

Rodrigo Duarte Silva

Aprendizagem de máquina aplicada a métodos de Classificação de Supernovas

Vitória - ES, Brasil

2018

Rodrigo Duarte Silva

Aprendizagem de máquina aplicada a métodos de Classificação de Supernovas

Dissertação de Mestrado apresentada ao curso de Pós-Graduação em Física da Universidade Federal do Espírito Santo como requisito essencial para obtenção do título de mestre em Física.

Universidade Federal do Espírito Santo – UFES

Departamento de Física

Orientador: Prof. Dr. Valerio Marra

Coorientador: Prof. Dr. Luciano Casarini

Vitória - ES, Brasil

2018

Dados Internacionais de Catalogação-na-publicação (CIP)
(Biblioteca Central da Universidade Federal do Espírito Santo, ES, Brasil)

S586a Silva, Rodrigo Duarte, 1985-
Aprendizagem de máquina aplicada a métodos de
classificação de supernovas / Rodrigo Duarte Silva. – 2018.
149 f. : il.

Orientador: Valerio Marra.
Dissertação (Mestrado em Física) – Universidade Federal do
Espírito Santo, Centro de Ciências Exatas.

1. Supernovas. 2. Cosmologia. 3. Aprendizado do
computador. 4. Simulação por computador. 5. Redes neurais
(Computação). I. Marra, Valério. II. Universidade Federal do
Espírito Santo. Centro de Ciências Exatas. III. Título.

CDU: 53

Elaborado por Perla Rodrigues Lôbo – CRB-6 ES-527/O



UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO

CENTRO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM FÍSICA

"Aprendizagem de máquina aplicada a métodos de classificação de supernovas"

Rodrigo Duarte Silva

Dissertação submetida ao Programa de Pós-Graduação em Física da Universidade Federal do Espírito Santo como requisito parcial para a obtenção do título de Mestre em Física.

Aprovada por:

Prof. Dr. Ribamar Rondon de Rezende dos Reis
(UFRJ-Participante remoto)

Prof. Dr. Valerio Marra
(Orientador /PPGFis)

Prof. Dr. Luciano Casarini
(Participante remoto PPGFis/UFES)

Prof. Dr. Julio César Fabris
(PPGFis/UFES)

Vitória-ES, 06 de setembro de 2018.

Agradecimentos

Na Física, descobri a Astrofísica e a Cosmologia, digo isso, porque o meu primeiro contato de imediato fascínio com essa ciência, ocorreu antes da graduação, através da Astronomia, durante certo tempo em que tive a oportunidade de ser monitor.

Durante a graduação entrei em maior contato com a Cosmologia, primeiro através dos professores Luciano Casarini e Galen Sotkov, e em seguida por meio do prof. Valerio Marra, durante o mestrado. Também não poderia deixar de citar os professores Oliver Piatela, Julio Fabris e Humberto Belich. Agradeço à todos os professores que contribuíram para a minha formação e funcionários do programa de pós-graduação em Física da Universidade Federal do Espírito Santo.

Agradecimentos especiais a Dr. Michelle Lochner do AIMS (African Institute for Mathematical Science), por ter disponibilizado o código de análise para modificações, e ao Dr. Ribamar Rondon de Rezende dos Reis, professor do Instituto de Física da Universidade Federal do Rio de Janeiro, por ajudar com as simulações no SNANA.

Agradeço à minha família e amigos, por todo o apoio e incentivo.

Aos membros da banca examinadora, ao PPGFis, ao PRPPG-UFES e à CAPES, muito obrigado!

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

"What we know is a drop; what we don't know is an ocean".

Issac Newton

Resumo

Futuras pesquisas observacionais com investimentos, telescópios e tecnologias nunca antes vistos, estão sendo propostas na tentativa de se desvendar os mistérios do Universo. Em nosso trabalho, fornecemos um panorama desse cenário, com especial atenção para a classificação de supernovas que será feita pelo LSST (Large Synoptic Survey Telescope) a partir de 2022.

Inicialmente introduzimos a física que envolve o evento de Supernovas e sua observação, com o objetivo de tratar o problema da classificação fotométrica de supernovas didaticamente. Fornecemos importantes referências no uso de diferentes aprendizagens de máquina e redes neurais para esse propósito. Incluimos resultados do uso de alguns dos métodos computacionais e a teoria por trás deles, destacando suas potencialidades e vulnerabilidades.

Os métodos de aprendizagem de máquina podem envolver supervisão ou não. Objetivamos descrever a aplicação destas poderosas ferramentas, na análise de dados observacionais e verificamos resultados inesperados.

Palavras-chave: Aprendizagem de máquina. Análise de dados. LSST. Supernova. Universo.

Abstract

Future observational research with investments, telescopes and technologies never before seen, are being proposed in an attempt to unravel the mysteries of the Universe. In our work, we provide an overview of this scenario, with special attention to the classification of supernovae that will be done by LSST (Large Synoptic Survey Telescope) from 2022.

Initially, we introduce the physics that involve the Supernova event and its observation, with the objective of treating the problem of photometric classification of supernovae. We provide important references in the use of different machine learning and neural networks for this purpose. We include results from the use of some of the computational methods and the theory behind them, highlighting their potentialities and vulnerabilities.

Machine learning methods may involve supervision or not. We aim to describe the application of these powerful tools in the analysis of observational data and verify unexpected results.

Keywords: Machine learning. Data analysis. LSST. Supernova. Universe.

Lista de ilustrações

Figura 1 – Tabela descritiva para os diferentes tipos e subtipos de supernovas (Wikipédia)	30
Figura 2 – Diferença entre a magnitude do pico observado em cada banda para as SN de (GUY, 2007), a previsão do modelo como função do comprimento de onda efetivo do filtro é utilizado (triângulos cinzas: SNLS SNe, quadrados cinzas: SNe próxima). Uma vez que as incertezas nas magnitudes B e V no máximo entram na normalização e avaliação de cores do modelo, as incertezas de correção K são definidas como zero para os comprimentos de onda da banda B e V.	55
Figura 3 – Diagrama de um neurônio artificial (<i>Linear Threshold Unit - LTU</i>) (GOOSSENS; RAHTZ; MITTELBAACH, 1997)	92
Figura 4 – Diagrama simplificado de uma rede neural MLP.	94
Figura 5 – Arquitetura tradicional de uma CNN.	97
Figura 6 – Priors para <i>SALT2</i> e modelos paramétricos. Os mesmos que foram utilizados em (LOCHNER et al., 2016).	104
Figura 7 – Visualizações t-SNE para os métodos de extração de características, com e sem a informação do redshift. <i>SALT2</i> e Modelo1.	106
Figura 8 – Visualizações t-SNE para os métodos de extração de características, com e sem a informação do redshift. Modelo2 e Wavelet.	107
Figura 9 – Curva ROC com e sem a informação do redshift para <i>SALT2</i> e Modelo1. Cada curva representa um algoritmo de aprendizagem de máquina diferente, com sua respectiva AUC informada na legenda.	108
Figura 10 – Curva ROC com e sem a informação do redshift para Modelo2 e Wavelet. Cada curva representa um algoritmo de aprendizagem de máquina diferente, com sua respectiva AUC informada na legenda.	109
Figura 11 – Exemplos de ajustes para curvas de luz obtidas através dos métodos de extração de característica, supernova tipo Ia.	110
Figura 12 – Exemplos de ajustes para curvas de luz obtidas através dos métodos de extração de característica, supernova tipo Ib.	111
Figura 13 – Exemplos de ajustes para curvas de luz obtidas através dos métodos de extração de característica, supernova tipo Ic.	112
Figura 14 – Exemplos de ajustes para curvas de luz obtidas através dos métodos de extração de característica, supernova tipo IIL.	113
Figura 15 – Exemplos de ajustes para curvas de luz obtidas através dos métodos de extração de característica, supernova tipo IIP.	114

Sumário

	Introdução	12
	INTRODUÇÃO GERAL	13
1	DADOS OBSERVACIONAIS E A CLASSIFICAÇÃO DE SUPER-NOVAS	18
1.1	Introdução à astrofísica de objetos compactos	18
1.1.0.1	Pressão de degenerescência do gás de Fermi e o limite de massa de Chandrasekhar	20
1.1.0.2	O <i>neutron drip</i> e estrelas de nêutrons	26
1.2	Aspectos históricos	28
1.3	Classificação espectroscópica e fotométrica	30
1.4	Dados de supernovas correlacionados	31
1.5	Não homogeneidade das Supernovas	32
1.6	Magnitudes	34
1.6.1	Sistema VEGA ou Jonhson	37
1.6.2	Sistema Gunn ou griz	38
1.6.3	Sistema AB	38
1.6.4	Sistema STmag	39
1.6.5	Sistema natural - LSST	39
1.7	Fotometria	40
1.7.1	Calibração da Fotometria LSST	41
2	AJUSTES E SIMULAÇÕES DE CURVAS DE LUZ	45
2.1	MLCS2k2	45
2.1.1	Aprendendo sobre curvas de luz (MLCS - Multicolor Light Curve Shape)	45
2.2	SALT2	51
2.2.1	Correção-k	54
2.3	Outros modelos de ajuste	55
2.3.1	Modelo1 - (NEWLING et al., 2010)	55
2.3.2	Modelo2 -(KARPENKA; FERROZ; HOBSON, 2012)	56
2.4	SNANA	57
2.5	SNcosmo	58
3	APRENDIZAGEM DE MÁQUINA	59
3.1	Validação Cruzada (<i>Cross-validation</i> - CV)	60
3.1.1	Validações cruzada exaustiva	60

3.1.1.1	Método <i>leave-p-out</i>	60
3.1.1.2	Método <i>leave-one-out</i>	60
3.1.2	Validações cruzadas não exaustiva	60
3.1.2.1	Método holdout	60
3.1.2.2	Método k-fold	61
3.2	Aprendizagem supervisionada	61
3.2.1	<i>k-Nearest Neighbors</i>	62
3.2.2	Métodos de regressão e modelos de ajustes	62
3.2.3	<i>Support Vector Machines(SVMs)</i>	62
3.2.4	<i>Decison Trees and Random Forest</i>	63
3.3	<i>Boosting</i>	65
3.3.1	AdaBoost:	65
3.3.2	Gradient Boosting:	66
3.4	Empacotamento (<i>Bagging</i>)	68
3.5	Aprendizagem não supervisionada	69
3.5.1	<i>KDE</i>	70
3.5.2	<i>k-Means</i>	71
3.5.3	Hierarchical Cluster Analysis(HCA)	71
3.5.4	Algoritmo EM e Extreme Deconvolution(XD)	72
3.6	Redução de dimensionalidade e extração de características	75
3.6.1	<i>Isomap</i>	76
3.6.2	<i>Principal Component Analysis (PCA)</i>	76
3.6.2.1	Transformada de Hotelling	77
3.6.3	<i>Independent Components Analysis (ICA)</i>	78
3.6.4	<i>Locally-Linear Embedding (LLE)</i>	80
3.6.5	t-distributed Stochastic Neighbor Embedding (t-SNE)	80
3.6.6	Ondaletas (<i>Wavelet</i>)	82
3.6.6.1	Transformada wavelet contínua (CWT)	84
3.6.6.2	Propriedades	85
3.6.6.3	<i>Power spectrum</i> e entropia wavelet	86
3.6.6.4	Wavelet de Haar	87
3.6.6.5	Wavelet de Shannon	87
3.6.6.6	Wavelet de Morlet	88
3.6.6.7	<i>Wavelet</i> chapéu mexicano	88
3.7	Avaliando a aprendizagem	88
3.7.1	Receiver Operating Characteristic (ROC)	90
4	<i>ARTIFICIAL NEURAL NETWORKS (ANN)</i>	91
4.1	<i>Perceptron, MLP e DNN</i>	94
4.2	<i>Convolutional Neural Network (CNN)</i>	95

4.2.0.1	<i>Convolutional layers (CL)</i>	95
4.2.0.2	<i>Pooling layers (PL)</i>	96
4.2.0.3	<i>Fully connected layers (FC)</i>	97
4.3	<i>Recurrent Neural Network (RNN)</i>	97
4.4	Processos de aprendizagem com ANNs	98
4.4.1	Regra Delta	98
4.4.2	Aprendizagem com memória	99
4.4.3	Regra de Hebb	100
4.4.4	Aprendizagem competitiva	100
4.4.5	Aprendizagem de Boltzmann	101
5	RESULTADOS	102
5.1	Modificações em snmachine e configuração do SNANA	102
5.2	Resultados gráficos	105
6	CONCLUSÃO	115
	REFERÊNCIAS	117
	APÊNDICE A – RESUMO DE ESTATÍSTICA	130
A.1	Probabilidades	130
A.2	Probalidades discretas ou contínuas	132
A.3	Teorema de Bayes	134
A.4	Montagem da Verossimilhança (Likelihood)	134
A.5	Evidência e <i>Odds ratio</i>	136

Introdução

No século XXI, iniciamos a era das grandes pesquisas observacionais com a utilização de aparelhagem moderna, motivadas em construir um "mapa cósmico" através da captura de informações para o maior número de objetos observáveis, possível. Nessa busca, pesquisadores inicialmente priorizavam o uso da espectroscopia, que exige um longo tempo de exposição do quadro de descanso (o que não otimiza a cadência de observação). Entretanto, pesquisas observacionais atuais e futuras estão buscando analisar dados apenas fotométricos. Como veremos o desenvolvimento de novas tecnologias digitais para observação e análise fotométrica, estão demonstrando superiores vantagens em comparação com a espectroscopia.

Aqui, fornecemos uma introdução teórica sobre Supernovas, astrofísica de Supernovas e classificação (espectroscópica×fotométrica) respectivamente, na Introdução Geral e no capítulo (1). Apresentamos os modelos e softwares mais utilizados em simulações e análise de dados para Supernovas no capítulo (2) e dedicamos o capítulo (3) a um panorama dos métodos de aprendizagem de máquina mais utilizados.

Os capítulos (3) e (4) tratam respectivamente de alguns dos diferentes métodos para a aprendizagem de máquina e topologias de redes neurais artificiais que são comumente utilizadas na análise de dados. No capítulo (5) fornecemos alguns dos resultados preliminares que foram obtidos através da simulação de dados fotométricos para 20.860 Supernovas, utilizando uma análise similar a realizada em (LOCHNER et al., 2016) e também, alguns exemplos de ajustes de curvas de luz, para diferentes tipos de supernovas, feitos utilizando quatro diferentes métodos de ajuste. Ao final, apresentamos as conclusões (capítulo 6).

Esperamos que nosso estudo desperte o interesse, em alunos de pós-graduação, para a utilização de métodos baseados em aprendizagens de máquina e redes neurais na realização de diferentes tarefas envolvendo o tratamento e a análise de dados.

INTRODUÇÃO GERAL

O estudo dos eventos Supernovas são fontes de muitas pesquisas, apresentando cada vez mais descobertas, o que vem propiciando um envolvimento cada vez maior dos físicos. Este fenômeno fascinante geralmente dá origem a um objeto compacto, como uma estrela de nêutrons ou pulsar e em casos extremos, buracos negros. Ele está relacionado com o importante processo de nucleossíntese, que, após a explosão, espalha elementos pesados por todo o Universo.

Supernovas despertam grande interesse nos físicos, porque em seu ambiente toda a física conhecida está envolvida. Além disso, no núcleo da estrela em colapso, estabelece-se um regime extremo, com densidade e temperatura muito grandes, só recentemente alcançadas em aceleradores de partículas. Esse é um ambiente propício para a descoberta do que podemos chamar de "nova física", que envolve novos fenômenos ainda não investigados por completo, um exemplo é que apesar de várias tentativas de se compreender os mecanismos físicos que conduzem ao colapso do núcleo em uma estrela e, por conseguinte, a explosão, esse cenário ainda não está completo (MEZZACAPPA; FULLER, 2005). A detecção de ondas eletromagnéticas, neutrinos e ondas gravitacionais emitidas durante a explosão de uma supernova, são as principais fontes de informações sobre este complexo fenômeno.

Emissões eletromagnéticas, relacionadas aos fenômenos Supernovas têm sido amplamente investigadas em diferentes regiões do espectro eletromagnético, como o visível e os intervalos de raios- X e raios γ . No entanto, as ondas eletromagnéticas são caracterizadas por interagirem com uma elevada seção de choque e, em um meio extremamente denso os fótons interiores a região não conseguem escapar, como consequência podemos observar apenas os fótons produzidos nas regiões mais externas da estrela. Portanto, o estudo desta emissão não pode explicar o que acontece nos primeiros segundos do colapso, que se caracteriza por rápidas mudanças na dinâmica do núcleo interno. Por outro lado, os neutrinos e as ondas gravitacionais, graças as suas baixas seções de choque, deixam a estrela durante os primeiros segundos do colapso, carregando informações valiosas sobre as regiões mais internas da estrela progenitora. Diversos projetos atualmente estão sendo desenvolvidos para detectar neutrinos e sinais de ondas gravitacionais emitidos por Supernovas (por exemplo, LIGO <<https://www.ligo.caltech.edu/>>, JUNO <<http://juno.ihep.cas.cn>>, dentre outros).

Uma Supernova tem sua origem com uma intensa explosão que determina o fim da vida de uma estrela. Após a queima de todo o combustível nuclear, o destino final da estrela depende basicamente de dois fatores: o primeiro é se a estrela faz ou não parte de um sistema binário ou múltiplo (cerca de 50% fazem parte deste tipo de sistema); o segundo fator é a massa da estrela. Se o sistema é binário ou múltiplo, além da dependência

fundamental da massa, há a dependência da separação entre as estrelas, isto é, um ponto da evolução em que as estrelas irão interagir e poderão “trocar” massas e, de certa maneira, alterar o comportamento deste sistema (CHOUDHURI, 2010).

Se a estrela é isolada, sua evolução só depende da massa inicial. Por exemplo, se uma estrela possuir uma massa menor que $0.8M_{\odot}$ (80% da massa solar), a idade do universo não é suficiente para que esta estrela (Anã Vermelha) evolua além da sequência principal no diagrama H-R (o diagrama Hertzsprung-Russell, retrata a evolução estelar através da relação existente entre a luminosidade de uma estrela e sua temperatura efetiva). Se a estrela possuir uma massa entre $0.8M_{\odot}$ e $4M_{\odot}$, após consumir o hidrogênio (combustível inicial da estrela), essa passará pela fase de gigante vermelha e depois de supergigante. Ela ejetará o que chamamos de nebulosa planetária, visível por aproximadamente 10.000 anos, e terminará sua vida como uma anã branca, com massa inferior a aproximadamente $1.4M_{\odot}$ e raio cerca de $10.000km$. Uma anã branca com massa superior ao limite *de massa de Chandrasekhar* dá origem a uma Supernova.

Se a estrela iniciar com massa entre $4M_{\odot}$ e $10M_{\odot}$, após a fase de supergigante ela ejetará a maior parte de seu material em uma explosão de supernova e terminará sua vida como uma estrela de nêutrons, com temperatura superficial acima de $10^6 K$, massa de cerca de $\approx 2M_{\odot}$, e raio da ordem de 10 km (CHOUDHURI, 2010). Se esta estrela possuir um campo magnético intenso, ela emitirá radiação direcionada em um cone em volta dos pólos magnéticos e será chamada de pulsar. Existem também estrelas que possuem massas entre $10M_{\odot}$ e $100M_{\odot}$ e após a fase de explosão o remanescente pode vir a ser um buraco negro, com uma massa da ordem de $6M_{\odot}$, e raio do horizonte menor que $10km$ (*raio de Schwarzschild* - $R_{(Sch)} = 2GM/c^2$). Se a estrela iniciar sua vida com massas superiores a $100M_{\odot}$, ela perderá a maior parte da sua massa ainda na sequência principal, devido a pressão de radiação (processo semelhante ao que ocorre com estrelas Wolf-Rayet) e depois evoluirá exatamente como uma estrela com massa menor que $100M_{\odot}$.

A emissão de radiação na frequência visível de uma explosão supernova inicia com uma grande luminosidade (muitas vezes mais luminosa que toda sua galáxia hospedeira) e que aumenta ainda mais, nas duas primeiras semanas, em média, até atingir o pico máximo. A partir deste ponto a luminosidade então diminui gradativamente, durante aproximadamente um ano (GIUNT; KIM, 2007). A explosão que dá origem à uma Supernova pode ser basicamente de dois tipos: colapso do núcleo ou termonuclear.

A explosão por colapso de núcleo faz parte do ciclo de vida normal de uma estrela que tem massa superior a aproximadamente $4M_{\odot}$, isso ocorre após sair da sequência principal no diagrama H-R (a maior parte das estrelas está alinhada ao longo de uma estreita faixa na diagonal do diagrama luminosidade×temperatura efetiva, que vai do extremo superior esquerdo (estrelas quentes e luminosas) até o extremo inferior direito (estrelas frias e pouco luminosas), a massa é o fator que determina a localização de uma estrela na

sequência principal). Ou seja, depois de “queimar” o hidrogênio no núcleo da estrela. Nesse ponto a geração de energia passa a ocorrer numa camada mais externa a este núcleo, onde temperatura e densidade são suficientes para manter as reações nucleares. Nenhuma energia, neste estágio, é gerada no núcleo, então a estrela se contrai rapidamente, e a luminosidade aumenta. As camadas mais externas se reajustam ao aumento de luminosidade expandindo-se e com o aumento da área superficial temos por consequência a redução da temperatura. Dessa maneira a luminosidade aumenta e a estrela torna-se mais vermelha, aproximando-se do ramo das gigantes vermelhas no diagrama H-R. O nosso Sol atingirá esta fase daqui aproximadamente 5 bilhões de anos e a reação triplo α iniciará combinando três partículas α (núcleos de hélio) em um núcleo de carbono. O Sol será, portanto, uma gigante vermelha, com um núcleo formado de hélio se transformando em carbono e hidrogênio transformando-se em hélio na fina camada mais externa da estrela. A massa do Sol não é suficiente para “queimar” o carbono, ou seja, alcançar temperaturas próximas de $10^9 K$.

Assim, a estrutura final do Sol será um núcleo de carbono extremamente denso e quente, onde não ocorrem mais fusões nucleares e o colapso gravitacional é impedido apenas pela pressão de degenerescência eletrônica. Ele esfriará vagarosamente durante centenas de bilhões de anos até deixar de ser visível. No diagrama HR o Sol descenderá para a região das anãs brancas (98% de todas as estrelas evoluirão até a fase de anã branca). Estrelas com massa acima de $10M_{\odot}$ evoluem muito rapidamente. Depois da fase de gigantes, passam para supergigantes, com temperaturas nucleares de alguns bilhões de Kelvin, permitindo assim a “queima” de outros elementos: O^{16} , Mg^{24} , Si^{28} , S^{32} , Cl^{35} , Ca^{40} , Sc^{45} , Ti^{48} ... Fe^{56} , em algumas centenas de milhões de anos. Este processo de fusão nuclear se encerra no Fe^{56} , uma vez que sua energia de ligação é muito alta para que o processo de fusão seja possível. Entretanto veremos que se a estrela for suficientemente grande, ela pode entrar na fase de Supernova, onde muitas reações nucleares importantes ocorrem. Isso porque durante a explosão os núcleos atômicos são acelerados a velocidades muito maiores do que as alcançadas no interior da estrela, nessas condições núcleos de Fe^{56} e outros podem se fundir em elementos mais pesados colidindo à velocidades relativísticas.

O colapso termonuclear ocorre somente em sistemas binários ou múltiplos. Nesse caso, depois de uma das estrelas se tornar uma Anã Branca, ela passa a crescer massa da outra estrela do sistema até ultrapassar uma certa massa crítica (em torno de $1.4M_{\odot}$) e sofrer a explosão termonuclear pelo mecanismo de acreção. Esses tipos de explosões são classificadas como Supernovas tipo Ia, conhecidas na cosmologia como "velas padrão" pois, a maioria dessas explosões ocorrem com aproximadamente a mesma magnitude absoluta (ou intrínseca) e por isso elas são perfeitas para o estudo das escalas de distâncias cosmológicas. Esse é um dos motivos pelo qual nos concentramos em métodos para a classificação de supernovas, mas os princípios subjacentes a esses métodos são mais gerais e podem ser aplicados em muitos outros problemas de classificação.

Historicamente essa classificação era realizada considerando características do espectro e propriedades da curva de luz. Essas características e propriedades dependem da composição do envoltório da estrela que deu origem a supernova, chamada progenitora (TURATTO; BENETTI; PASTORELLO, 2007). Assim, a classificação era feita pela análise da presença ou ausência no espectro dos elementos que compõem a estrela progenitora e pela análise do formato da curva de luz capturada ao longo do tempo em diferentes bandas do espectro eletromagnético (TURATTO, 2003).

Na era das grandes pesquisas astronômicas, a espectroscopia está dando lugar a classificação puramente fotométrica, devido as dificuldades inerentes do acompanhamento e da classificação espectroscópica do candidato. Felizmente pesquisas recentes tais HST, SNLS, SDSS dentre outras, têm de forma convincente demonstrado que a probabilidade de qualquer objeto ser uma SN Ia pode ser derivada através de sua curva de luz (KUNZ; BASSET; HLOZEK, 2007). Nesse trabalho apresentamos os mais novos e bem sucedidos métodos aplicados a classificação fotométrica de supernovas baseados no desenvolvimento de redes neurais e na aprendizagem de máquina que estão em fase de teste para aplicação no projeto do grande telescópio de levantamento sinóptico LSST (TYSON; STRAUSS, 2009) que ficará em operação pelo menos dez anos, tendo como previsão de início das atividades para Outubro de 2022. Está localizado próximo ao CTIO (Cerro Tololo Inter-American Observatory) no Cerro Pachón (Chile), a 100 km da cidade de apoio de La Serena, onde está localizada a base do LSST. Esse local foi estrategicamente escolhido devido suas ótimas condições para observação.

A pesquisa está sendo projetada para vasculhar os mistérios da matéria e da energia escura através da observação de Supernovas, estrelas variáveis e diversos outros transientes, identificar perigosos asteróides em distâncias remotas do Sistema Solar e estudar a estrutura e formação da Via Láctea. O telescópio LSST terá 8.4 metros de diâmetro e contará com um design especial de três espelhos, criando um campo de visão excepcionalmente amplo com capacidade de pesquisar quase todo o hemisfério celestial Sul, cerca de 20.000 deg^2 em apenas 3 noites (toda a esfera celeste tem aproximadamente 41.253 deg^2). Além disso, a câmera LSST, que contém mais de 3 Gigapixels em detectores de estado sólido CCD (Charge Couple Device) (TYSON; STRAUSS, 2009), deverá produzir dados com qualidade extremamente alta e velocidade de processamento nunca antes realizada. A parte delicada do projeto e que está em fase final de desenvolvimento, é o software, já que mais de 15 terabytes de dados devem ser processados e armazenados todas as noites na produção do maior conjunto de dados publicamente disponíveis no mundo, mais de 50 petabytes (PB, equivalente a 1024 terabytes) ao final dos 10 anos de pesquisa.

O software LSST é formado por uma pilha de códigos que será responsável por toda a infraestrutura de operação, processamento e armazenagem dos dados. Nosso estudo foi dedicado aos métodos computacionais para a classificação automática de supernovas

dentre outros transiente. Publicações recentes ([LOCHNER et al., 2016](#); [CHARNOCK; MOSS, 2017](#); [ANTARES-COLLABORATION et al., 2018](#); [DAI et al., 2018](#)) destacam os promissores resultados obtidos nesse sentido e são fundamentais para a escolha dos métodos estudados.

1 Dados observacionais e a classificação de Supernovas

Começamos com uma breve introdução à astrofísica estelar de anãs brancas (progenitoras) e estrelas de nêutrons (remanescentes de supernovas), acompanhada de uma revisão histórica. Em seguida tratamos de relatar alguns dos problemas inerentes a observação e classificação de supernovas. Ao final deste capítulo, fornecemos uma breve revisão de magnitudes e sistemas fotométricos com ênfase na pesquisa do LSST.

1.1 Introdução à astrofísica de objetos compactos

Para iniciarmos nosso estudo astrofísico, consideramos o problema na forma mais simples, comparando a estrela com uma distribuição esfericamente simétrica e uniforme de matéria. Se M_r é a massa no interior de uma esfera de raio r centrada na estrela, a massa no interior de uma esfera de raio $r + dr$ será $M_r + dM_r$, onde dM_r é a massa da casca esférica entre os raios r e $r + dr$. Se ρ é a densidade em r , então a massa da casca esférica será dada por:

$$dM_r = 4\pi\rho r^2 dr \longrightarrow \frac{dM_r}{dr} = 4\pi\rho r^2, \quad (1.1)$$

Se agora considerarmos uma pequena parte da casca entre r e $r + dr$ com um elemento infinitesimal de área transversal dA , as forças exercidas pela ação da pressão em cada lado de dA serão PdA e $-(P + dP)dA$, onde P e $P + dP$ são respectivamente as pressões em r e $r + dr$. Assim, a força efetiva devido a pressão $-dPdA$ deve ser contrabalanceada pelo campo gravitacional devido a massa M_r ($-GM_r/r^2$). Desde que a massa do elemento infinitesimal em consideração seja dada por $\rho dr dA$ podemos escrever a condição das forças para o equilíbrio:

$$-dPdA - \frac{GM_r}{r^2} \rho dr dA = 0, \quad (1.2)$$

ou equivalentemente,

$$\frac{dP}{dr} = -\frac{GM_r\rho}{r^2}, \quad (1.3)$$

que é a equação para o equilíbrio hidrostático. De modo geral essa relação implica que $\vec{\nabla}P = -\rho\vec{\nabla}\Phi$, onde Φ é o potencial gravitacional produzido pela distribuição de massa M_r e pode ser obtido a partir da *equação de Poisson* (FABRIS et al., 2010). Uma das consequências imediatas da lei do equilíbrio hidrostático é o chamado *teorema do virial*.

Para estrelas comuns como o nosso Sol, a atração gravitacional é compensada pelo excesso de pressão no interior quente da estrela. Assim esperamos que a energia térmica total seja da mesma ordem de grandeza que a energia gravitacional total da estrela. Isto pode ser estabelecido, multiplicando a equação (1.3) por $4\pi r^3$ em ambos os lados da igualdade e integrando do centro até o raio externo (R) da estrela (para obtermos a soma das contribuições de cada casca esférica com espessura infinitesimal para, respectivamente, energias térmica e gravitacional total da estrela):

$$\int_0^R \frac{dP}{dr} 4\pi r^3 dr = - \int_0^R \frac{GM_r \rho}{r^2} 4\pi r^3 dr \quad (1.4)$$

O lado esquerdo da equação anterior pode ser facilmente integrado por partes, fornecendo:

$$-3 \int_0^R P 4\pi r^2 dr = - \int_0^R \frac{GM_r \rho}{r} 4\pi r^2 dr. \quad (1.5)$$

O lado direito desta última relação é claramente a energia gravitacional total da estrela (E_G). Aproximando $(3/2)k_B T$ como a energia média térmica por partícula na região com temperatura T , ou seja, $(3/2)nk_B T$ é a energia térmica por unidade de volume. Podemos escrever a energia térmica total (E_T) da estrela:

$$E_T = \int_0^R \frac{3}{2} nk_B T \times 4\pi r^2 dr = \frac{3}{2} \int_0^R P 4\pi r^2 dr. \quad (1.6)$$

Através desse resultado podemos reescrever a equação 1.5 de uma forma mais elegante e famosa, conhecida como *teorema do virial*:

$$2E_T + E_G = 0 \quad (1.7)$$

Mas e quando o combustível da estrela se esgota e não há mais fonte de pressão térmica para equilibrar a gravidade? Claramente a estrela começa a se contrair e continua contraindo a menos que algum tipo de pressão seja capaz de equilibrar novamente a gravidade. Neste caso, temos que fazer uso de uma importante propriedade das partículas de Fermi. Em uma célula unitária no espaço de fase com volume h^3 em seis dimensões (posição e momento), não pode haver mais de duas partículas de Fermi (uma com spin "up" e outra com spin "down" que é o *princípio de exclusão de Pauli*). Os elétrons no interior da matéria estelar compõem um gás de Fermi, e quando a densidade dentro da estrela em contração se torna suficientemente alta, esse gás se torna degenerado, isto significa que o limite teórico de duas partículas por célula unitária do espaço de fase é quase ou totalmente atingido. Veremos que tal gás de Fermi degenerado exerce o que é conhecido como pressão de degenerescência. Acredita-se que estrelas anãs brancas representam configurações estelares nas quais as forças de atração da gravidade são equilibradas pela pressão de degenerescência do gás de elétrons.

Outra possível configuração final das estrelas é uma estrela de nêutrons. Como veremos, em densidades muito altas os elétrons são forçados a combinar-se com os núcleos para produzir matéria composta principalmente de nêutrons. Como os nêutrons são partículas Fermi, um gás de nêutrons também exerce pressão de degenerescência. Uma estrela de nêutrons é uma configuração estelar na qual a gravidade é balanceada pela pressão de degenerescência de nêutrons. Como não conhecemos com precisão a equação de estado da matéria nas extremamente altas densidades que prevalecem no interior das estrelas de nêutrons, a sua estrutura não é tão bem compreendida quanto a estrutura das anãs brancas. Estrelas de nêutrons também têm um limite superior de massa assim como anãs brancas. Mas esse limite de massa não é conhecido com exata precisão devido às incertezas em nosso conhecimento da equação de estado, mas cálculos sugerem que esse limite de massa não seja superior a $2M_{\odot}$ (CHOUDHURI, 2010).

Embora as estrelas de nêutrons tenham sido teoricamente postuladas nos anos 1930 logo após a descoberta do nêutron, elas permaneceram uma curiosidade teórica por mais de três décadas. Os pulsares foram descobertos em 1968 e foram rapidamente identificados como estrelas de nêutrons rotativas. A massa inicial de uma estrela não necessariamente precisa ser menor do que o limite de massa de anãs brancas ou estrelas de nêutrons para que a estrela termine em uma dessas configurações. Em (CHOUDHURI, 2010) é demonstrado que uma estrela pode perder uma parte considerável de sua massa durante as últimas fases da evolução, na forma de ventos solares intensos durante a fase gigante vermelha, ou através de mecanismos de ejeção mais drásticos, como o derramamento de camadas externas como uma nebulosa planetária ou uma explosão de supernova.

A partir de estudos estatísticos de vários tipos de estrelas, infere-se que estrelas menos massivas que cerca de $4M_{\odot}$ eventualmente tornam-se anãs brancas, enquanto estrelas com massas iniciais no intervalo entre $4M_{\odot}$ e $10M_{\odot}$ podem acabar como estrelas de nêutrons, tipicamente depois da explosão de supernova (SHAPIRO; TEUKOLSKY, 1983).

1.1.0.1 Pressão de degenerescência do gás de Fermi e o limite de massa de Chandrasekhar

A pressão de um gás é devido aos movimentos aleatórios das partículas que o constituem. Assim, se $4\pi f(p)p^2 dp$ é o número de partículas com momento entre p e $p + dp$ (assumindo uma função de distribuição isotrópica), onde v é a velocidade das partículas existentes com momento p , então a pressão P é dada pela seguinte expressão padrão da teoria cinemática:

$$P = \frac{1}{3} \int v p f(p) 4\pi p^2 dp \quad (1.8)$$

Para um gás comum, nós podemos substituir a distribuição de Maxwell na equação anterior, e a pressão será dada por $nk_B T$, onde n é o número de partículas por unidade de volume. Claramente essa pressão que advém dos movimentos térmicos das partículas deve ir a zero

em $T = 0K$ se assumirmos a validade da física clássica. No entanto, quando um gás de partículas Fermi é comprimido em alta densidade, muitas das partículas são forçadas a permanecerem com momentos não nulos mesmo em $T = 0K$, dando origem a pressão de degenerescência. Quando a matéria estelar é comprimida, elétrons tornam-se degenerados muito antes de prótons e outros nucleons. Isto ocorre porque a energia cinética é igualmente particionada entre os diferentes tipos de partículas e esperamos que partículas mais leves (como elétrons) tenham menores momentos. Assim, elétrons alcançam o limite teórico de ocupação do espaço de fase e tornam-se degenerados antes, apresentando uma densidade numérica muito superior a densidade numérica de outras partículas mais pesadas. Elétrons que ocupam um espaço real de volume V e têm momentos em d^3p espaço de momentos têm $2Vd^3p/h^3$ estados no espaço de fase disponíveis (devido aos dois estados de spin). Se d^3p corresponde a uma casca esférica entre os momentos p e $p + dp$, então claramente o número de estados por unidade de volume dentro da casca é $8\pi p^2 dp/h^3$ (os estados ocupados são dados pela estatística de Fermi-Dirac (PATHRIA, 1996)). Para simplificar a abordagem, vamos desconsiderar os efeitos devido a temperatura e assumir que todos os estados abaixo do momento de Fermi (p_F) estão ocupados. Nestas condições a densidade numérica de elétrons (n_e) será dada por:

$$n_e = \int_0^{p_F} \frac{8\pi}{h^3} p^2 dp = \frac{8\pi}{3h^3} p_F^3. \quad (1.9)$$

Se todos os estados entre p e $p + dp$ estão ocupados, $8\pi p^2 dp/h^3$ tem que ser igual a $4\pi f(p)p^2 dp$, implicando em $f(p) = 2/h^3$ se $p < p_F$ e 0 se $p > p_F$. Portanto,

$$P = \frac{8\pi}{3h^3} \int_0^{p_F} v p^3 dp. \quad (1.10)$$

Se agora substituirmos o momento da partícula pela expressão relativística $p = m\gamma v$, onde γ é o fator de lorentz (JACKSON, 1999), teremos:

$$v = \frac{p}{m\gamma} = \frac{pc^2}{E} = \frac{pc^2}{\sqrt{p^2 c^2 + m^2 c^4}}. \quad (1.11)$$

$$P = \frac{8\pi}{3h^3} \int_0^{p_F} \frac{p^4 c^2}{\sqrt{p^2 c^2 + m_e^2 c^4}} dp. \quad (1.12)$$

Nosso objetivo é determinar a equação de estado que conecta pressão e densidade. Prótons e outros núcleos pesados presentes no material estelar contribuem para a densidade, mas não para a pressão, porque eles não estão degenerados. Logo, vamos inicialmente encontrar a relação entre a densidade ρ e a densidade numérica de elétrons n_e . Se X é a fração total de massa de hidrogênio quando a densidade numérica de átomos de hidrogênio é $X\rho/m_H$, estes átomos contribuem com $X\rho/m_H$ elétrons por unidade de volume. Um átomo de hélio

tem massa atômica 4 e contribui com dois elétrons, logo ele contribui com 1/2 elétron por unidade de massa. Para átomos mais pesados que o hélio o número de elétrons também é a metade do número de núcleons. Em unidades de volume de matéria estelar esses átomos mais pesados contribuem com uma massa $(1 - X)\rho$ que corresponde a $(1 - X)\rho/m_H$ núcleons e $(1 - X)\rho/2m_H$ elétrons. Logo a densidade numérica de elétrons total pode ser dada por:

$$n_e = \frac{X\rho}{m_H} + \frac{(1 - X)\rho}{2m_H} = \frac{\rho}{2m_H}(1 + X). \quad (1.13)$$

Também podemos escrever na forma,

$$n_e = \frac{\rho}{\mu_e m_H}, \quad (1.14)$$

onde μ_e é a média molecular para os pesos dos elétrons dado por:

$$\mu_e = \frac{2}{1 + X}. \quad (1.15)$$

Combinando as equações (1.9) e (1.14) podemos escrever o momento de Fermi (p_F) como:

$$p_F = \left(\frac{3h^3\rho}{8\pi\mu_e m_H} \right)^{1/3}. \quad (1.16)$$

Avaliando a equação (1.12) com a expressão encontrada para p_F , nós obtemos a equação de estado que relaciona P e ρ .

Consideraremos agora os dois casos extremos, elétrons não relativísticos e elétrons totalmente relativísticos. Primeiro para o caso não relativístico, podemos aproximar:

$$\sqrt{p^2 c^2 + m_e^2 c^4} \approx m_e c^2, \quad (1.17)$$

substituindo em (1.12), obtemos:

$$P = \frac{8\pi}{15h^3 m_e} p_F^5. \quad (1.18)$$

substituindo p_F pela expressão encontrada em (1.16),

$$P = K_1 \rho^{5/3}, \quad (1.19)$$

onde K_1 é dado por:

$$K_1 = \frac{3^{2/3}}{20\pi^{2/3}} \frac{h^2}{m_e m_H^{5/3} \mu_e^{5/3}} = \frac{1.00 \times 10^7}{\mu_e^{5/3}}, \quad (1.20)$$

em unidades do S.I. A expressão não relativística descrita acima, é a equação de estado para elétrons degenerados, foi descrita pela primeira vez por (FOWLER, 1926), reconhecido como o primeiro a perceber que a gravidade no interior de anãs brancas deve ser equilibrada pela pressão eletrônica de degenerescência.

Na segunda situação extrema, quando os elétrons são totalmente relativísticos, temos:

$$\sqrt{p^2 c^2 + m_e^2 c^4} \approx pc, \quad (1.21)$$

e substituindo na equação (1.12),

$$P = \frac{2\pi c}{3h^3} p_F^4, \quad (1.22)$$

Podemos também escrever a expressão para a pressão de Fermi através de uma transformação politrópica (1.19) (transformação termodinâmica na qual a pressão e o volume de um gás (normalmente considerado ideal) apresentam uma relação do tipo PV^γ), obtendo:

$$P = K_2 \rho^{4/3}, \quad (1.23)$$

onde K_2 é dado por:

$$K_2 = \frac{3^{1/3}}{8\pi^{1/3}} \frac{hc}{m_H^{4/3} \mu_e^{4/3}} = \frac{1.24 \times 10^{10}}{\mu_e^{4/3}} \quad (1.24)$$

As equações de estado em (1.19) e (1.23) são da forma:

$$P = K \rho^{(1+\frac{1}{n})}, \quad (1.25)$$

com n assumindo os valores $3/2$ e 3 para os casos não relativístico e totalmente relativístico, respectivamente. A relação em (1.25) entre densidade e pressão é conhecida como politropo. Admitindo uma simetria esférica nós podemos ir além em nossa descrição e escrever a densidade no interior da estrela, na forma:

$$\rho(r) = \rho_c \theta^n(r), \quad (1.26)$$

onde $\rho_c = \rho(r=0)$ é a densidade central da estrela e θ é uma nova variável adimensional que claramente tem valor 1 no centro estelar. Substituindo (1.26) em (1.25), obtemos:

$$P = K \rho_c^{\frac{n+1}{n}} \theta^{n+1}. \quad (1.27)$$

Além disso, as equações de equilíbrio hidrostático (1.3) e continuidade de massa podem ser combinadas em uma equação diferencial de segunda ordem, substituindo (1.1) em (1.3), dividindo o resultado por ρ , multiplicando por r^2 e derivando ambos os lados da igualdade em relação a r , obtemos:

$$\frac{1}{r^2} \frac{d}{dr} \left(\frac{r^2}{\rho} \frac{dP}{dr} \right) = -4\pi G \rho. \quad (1.28)$$

Esta é a equação de Poisson. Utilizando a equação (1.26) e definindo a variável adimensional $r \equiv a\xi$, onde:

$$a = \left[\frac{(n+1)K\rho_c^{(1/n-1)}}{4\pi G} \right]^{1/2} \quad (1.29)$$

podemos reescrever a equação de Poisson (1.28) em uma forma mais elegante:

$$\frac{1}{\xi^2} \frac{d}{d\xi} \left(\xi^2 \frac{d\theta}{d\xi} \right) = -\theta^n, \quad (1.30)$$

que é mais conhecida como equação de Lane-Emden, em homenagem ao físico americano Jonathan Homer Lane (1819 – 1880), que derivou a equação de equilíbrio hidrostático em 1869 e as contribuições do físico suíço Robert Emden (1862 – 1940). Observamos que a relação polítrpos e a equação de Lane-Emden realizam um importante papel na descrição de estruturas estelares. Sabemos que tais abordagens fornecem apenas um modelo aproximado de estrelas normais, entretanto, no estudo da estrutura de anãs brancas, essa é a abordagem padrão.

Agora necessitamos resolver a equação de Lane-Emden com respeito a duas condições de contornos, são elas:

$$\theta(\xi = 0) = 1; \quad (1.31)$$

$$\left(\frac{d\theta}{d\xi} \right)_{\xi=0} = 0. \quad (1.32)$$

Resolver a equação (1.30) para $n < 5$ é relativamente fácil, pois θ vai para zero em um finito valor de ξ , que denotaremos por ξ_1 . Nós interpretamos esse resultado como a superfície da estrela, onde densidade e pressão dadas pelas equações (1.26) e (1.27) respectivamente, vão para zero. Devemos encontrar como ρ_c , R e M se relacionam, assim se ξ_1 é o valor de ξ quando θ vai para zero, o raio físico da estrela pode ser dado por:

$$R = a\xi_1. \quad (1.33)$$

Olhando para a expressão na equação (1.29), concluímos que:

$$R \propto \rho_c^{\frac{1-n}{2n}} \quad (1.34)$$

Considerando que todas as outras quantidades que aparecem na expressão para R sejam as mesmas para o grupo de estrelas em estudo. A massa de cada estrela poderá ser dada por:

$$M = \int_0^R 4\pi r^2 \rho dr = 4\pi a^3 \rho_c \int_0^{\xi_1} \xi^2 \theta^n d\xi. \quad (1.35)$$

Observando a dependência de a em ρ_c , encontramos

$$M \propto \left(\rho_c^{\frac{1-n}{2n}} \right)^3 \rho_c \equiv \rho_c^{\frac{3-n}{2n}}. \quad (1.36)$$

Logo, para $n = 3/2$ na equação (1.25) (isto é, para os casos não relativísticos) temos:

$$R \propto \rho_c^{-1/6}, M \propto \rho_c^{1/2}, \quad (1.37)$$

que podemos combinar para obter:

$$R \propto M^{-1/3}. \quad (1.38)$$

Esta é a importante relação massa-raio para anãs brancas quando a equação de estado não relativística equação (1.19) é satisfeita. Verificamos claramente que quanto mais aumenta a massa da anã branca menor é o seu tamanho.

Vamos agora considerar o caso relativístico, tomando a equação (1.25) com $n = 3$. Um resultado surpreendente é que a massa M se torna independente de ρ_c para $n = 3$ na equação (1.36). Em outras palavras, a massa da estrela obedece a equação de estado relativística (equação 1.23) e tem um valor fixo que pode ser obtido pela eq. (1.36). Multiplicando a equação (1.30) por ξ^2 e integrando para $\xi = 0$ até $\xi = \xi_1$, temos:

$$\int_0^{\xi_1} \xi^2 \theta^n d\xi = -\xi_1^2 \left(\frac{d\theta}{d\xi} \right)_{\xi=\xi_1}, \quad (1.39)$$

podemos agora, substituir a integral na (eq.1.36) por $|\xi_1^2 \theta'(\xi_1)|$, onde θ' denota a derivada com respeito a ξ . Podemos ainda substituir a expressão para a (fornecida em eq.1.29) e para K na eq.(1.24), em (1.36) e encontramos finalmente a relação para massa limite:

$$M_{Ch} = \frac{\sqrt{6}}{32\pi} \left(\frac{hc}{G} \right)^{3/2} \left(\frac{2}{\mu_e} \right)^2 \frac{\xi_1^2 |\theta'(\xi_1)|}{m_H^2}. \quad (1.40)$$

Resolvendo a equação de Lane-Emden numericamente para $n=3$, (CHOUDHURI, 2010) encontra $\xi_1^2 |\theta'(\xi_1)| = 2.018$. Substituindo este valor na equação anterior juntamente com os demais valores de constantes, obtemos:

$$M_{Ch} = 1.46 \left(\frac{2}{\mu_e} \right)^2 M_{\odot}. \quad (1.41)$$

A massa M_{Ch} corresponde ao limite relativístico da equação de estado em função da massa quando o raio da estrela vai para zero. Basicamente essa relação impede que existam anãs brancas com massas superiores a M_{Ch} , conhecida como *limite de massa de Chandrasekhar* (CHANDRASEKHAR, 1931).

Anãs brancas normalmente se formam em núcleos de estrelas progenitoras que já "queimaram" completamente o Hidrogênio na produção de Hélio (e outros elementos mais pesados em certas condições). Se a fração de massa de Hidrogênio $X \approx 0$, através da equação 1.15 temos que $\mu_e \approx 2$, o que estabelece o *limite de massa de Chandrasekhar* em torno de $1.4M_{\odot}$. Em (CHOUDHURI, 2010) é demonstrado que a equação de estado torna-se relativística quando a densidade no interior da estrela é da ordem de 10^9 kg/m^3 , que é a densidade típica no interior de uma anã branca. Se a massa típica das anãs brancas é da ordem de uma massa solar ($\approx 10^{30} \text{ kg}$), o raio de uma anã branca deve ter cerca de $10^7 \text{ m} \approx 10^4 \text{ km}$, que realmente é o tamanho típico dessas estrelas.

1.1.0.2 O *neutron drip* e estrelas de nêutrons

Assim como a pressão de degenerescência dos elétrons suporta uma anã branca contra a ação gravitacional, a pressão de degenerescência dos nêutrons suporta uma estrela de nêutron. Ao contrário dos prótons, nêutrons são eletricamente neutros, e portanto, muitos nêutrons podem ser agrupados na ausência de repulsão eletrostática. No entanto os nêutrons sofrem decaimento conforme a reação:



com meia-vida de aproximadamente 13 minutos (CHOUDHURI, 2010). A reação reversa também pode ser considerada possível no interior de estrelas de nêutrons progenitoras:



Por simplicidade estamos considerando que a matéria estelar comprimida seja formada apenas por prótons, nêutrons e elétrons (isto é, que as equações acima ocorrem apenas para a direita). Para um estudo mais preciso, neutrinos, pósitrons e outros núcleons devem ser incluídos ($n + e^+ \longleftrightarrow p + \bar{\nu}$).

Como vimos anteriormente com o aumento da densidade, os elétrons se tornam degenerados enquanto partículas mais pesadas permanecem não degeneradas. Vamos supor que queremos adicionar um elétron em uma região de alta densidade. Sabemos que todos os níveis estão preenchidos até o momento Fermi p_F , que se relaciona com a densidade numérica de elétrons através da eq.(1.9). Podemos descrever a energia Fermi (E_F) associada ao momento Fermi:

$$E_F = \sqrt{p_F^2 c^2 + m_e^2 c^4}. \quad (1.44)$$

Isto significa que uma energia igual ou superior a $E_F - m_e c^2$ deve ser adicionada a um elétron, para que seja possível colocar mais elétrons na região de alta densidade, uma vez que todos os estados de menor energia estão preenchidos. Imagine agora a situação em que esse excesso de energia requerida se torna igual ou maior que $(m_n - m_p - m_e)c^2$ a quantidade pela qual a massa de nêutrons excede a soma da massa de prótons e elétrons. Nesta situação, será energeticamente favorável para o elétron se combinar com um próton para produzir um nêutron. De acordo com a equação 1.43 a condição para essa situação crítica é

$$\sqrt{p_{F,c}^2 c^2 + m_e^2 c^4} - m_e c^2 = (m_n - m_p - m_e)c^2, \quad (1.45)$$

onde $p_{F,c}$ é o momento crítico de Fermi. Podemos também escrever:

$$m_e c^2 \left(1 + \frac{p_{F,c}^2}{m_e^2 c^2} \right)^{1/2} = Q c^2, \quad (1.46)$$

onde $Q = m_n - m_p$. Ou em termos do momento crítico de Fermi:

$$p_{F,c} = m_e c \left[\left(\frac{Q}{m_e} \right)^2 - 1 \right]^{1/2}. \quad (1.47)$$

Como o momento de Fermi aumenta com a densidade, esperamos que p_F seja menor que $p_{F,c}$. Nesta situação, os elétrons livres são favorecidos energeticamente e não esperamos que nenhum nêutron esteja presente, uma vez que eles decaem (eq.1.42). Assim, a densidade crítica, na qual $p_c = p_F = p_{F,c}$ pode ser obtida substituindo os valores das constantes fundamentais na equação anterior e multiplicando n_e por $m_p + m_e$ (já que apenas prótons e elétrons estão presentes abaixo da densidade crítica), o que fornece

$$p_c = 1.2 \times 10^{10} \text{ kg/m}^3. \quad (1.48)$$

Quando a densidade é aumentada, os elétrons começam a se combinar com prótons para dar origem a nêutrons. Esse fenômeno é chamado de *neutron drip* (neutronização). Em densidades bem acima da densidade crítica, a matéria consistiria principalmente de

nêutrons. Esses nêutrons não decaem de acordo com a equação (1.42), que é completamente suprimida, uma vez que não há estados livres para os elétrons produzidos ocuparem. Apresentamos aqui um cálculo simplificado de *neutron drip* sem considerar a possível formação de núcleos. Quando a existência de núcleos é levada em conta, o cálculo se torna muito mais complexo. Para maiores detalhes veja (SHAPIRO; TEUKOLSKY, 1983).

Os nêutrons, assim como os elétrons são partículas de Fermi e obedecem ao princípio de exclusão de Pauli, logo, nêutrons também podem causar uma pressão de degenerescência. Para elétrons derivamos a pressão de degenerescência através da estatística de Fermi-Dirac que assume não haver interação entre as partículas, o que é uma boa suposição para um gás de elétrons dentro de uma anã branca. Entretanto, quando os nêutrons são compactados em densidades próximas à densidade dentro de um núcleo atômico (como é o caso dos interiores das estrelas de nêutrons), os nêutrons vizinhos interagem entre si através de forças nucleares e não podemos tratá-los como não interagentes. Portanto, encontrar uma equação exata do estado para a matéria em densidades tão altas é muito complexo e o assunto ainda não está totalmente compreendido.

Assim como o limite de massa de Chandrasekhar para anãs brancas, as estrelas de nêutrons também têm um limite de massa. No entanto, esse limite de massa não é conhecido com muita precisão devido à incerteza em nosso conhecimento da equação de estado. Estudos anteriores e cálculos detalhados sugerem que uma estrela de nêutrons tipicamente tenha uma massa entre $2M_{\odot}$ e $3.2M_{\odot}$, com um raio aproximadamente da ordem de $10km$ e uma densidade interna média de $10^{18}kg/m^3$ (CHOUDHURI, 2010).

Outro fator importante é que em nossa aproximação, usamos simplificações da relatividade geral desprezando o fator $2GM/c^2r$ quando pequeno comparado com 1. Entretanto, para uma estrela de nêutrons com $1M_{\odot}$ e raio de $10km$, esse fator é aproximadamente 0.3 e os efeitos da relatividade geral não podem ser desprezados em um cálculo mais rigoroso. Isto é, as equações de estado devem ser modificadas para contemplar a relatividade geral, algo que está além dos propósitos de nosso trabalho.

1.2 Aspectos históricos

Ao longo da história da humanidade, foram registrados alguns milhares de fenômenos de supernovas, sendo que o registro mais antigo, remonta a China 185 d.C (SN185), hoje observamos apenas seu envoltório gasoso (RCW 86) remanescente registrado em (YAMAGUCHI; KOYAMA; UCHIDA, 2012). Em 1054 d.C, novamente os chineses, observaram e registraram a explosão da estrela que acabou por produzir a nebulosa do Caranguejo e o pulsar do Caranguejo (SN1054). Os chineses, apesar de terem registrado o evento, não sabiam o porque ocorriam nem que estavam relacionados a colapsos estelares. Alguns eventos de supernovas ocorridos em nossa galáxia foram observados a olho nu nos últimos

2000 anos: a SN1006 foi a explosão mais intensa observada, as outras duas foram vistas por Tycho Brahe em 1572 (SN1572) e por Joannes Kepler em 1604 (SN1604).

A maioria das supernovas registradas, foram observadas por telescópio no último século, cada vez mais em detalhes graças ao desenvolvimento da tecnologia observacional. Em 23 de fevereiro de 1987, em uma das estrelas na Grande Nuvem de Magalhães (aglomerado estelar satélite da via láctea), foi observada a supernova SN1987A, essa foi uma das mais bem estudadas nos últimos anos, ela teve grande relevância já que foi a primeira estudada com aparelhagem moderna, onde também foi possível detectar neutrinos, embora em pequenas quantidades. Ela foi responsável por vários estudos e publicações sobre a dinâmica de supernovas e também propriedades intrínsecas dos neutrinos.

Como mencionamos anteriormente, a classificação de supernovas vem sendo realizada considerando características espectrais. Essas características e propriedades dependem da composição do envoltório da estrela progenitora (TURATTO; BENETTI; PASTORELLO, 2007). Assim, a classificação é feita através da análise da presença ou ausência no espectro dos elementos que compõem a estrela progenitora e da análise do formato da curva de luz capturada ao longo do tempo (TURATTO, 2003). Mas devido ao longo tempo de exposição necessário para a captura do espectro, esse método está sendo fortemente restringido nas futuras missões que se propõem a investigar o Universo em busca de supernovas nas próximas décadas. O objetivo dessas missões se concentra em capturar o maior número de informações sobre as propriedades da *curva de luz*, através delas podemos dentre outras coisas, distinguir entre os diferentes tipos de supernovas (WHEELER; BENETTI, 2002).

Os esquemas de classificação desenvolvidos desde 1941 (MINKOWSKI, 1941), reconheceram duas grandes categorias inicialmente: Supernovas tipo I, caracterizadas pela ausência de hidrogênio em sua composição; e do tipo II, caracterizadas pela presença de hidrogênio em sua composição. Estudos de análise fotométrica também foram iniciados (BERTOLA, 1964), e foi observado que tipicamente a emissão óptica de SNs do tipo I e II começam com um crescimento na luminosidade durante a primeira e segunda semana devido à contração da superfície de luminosidade. Supernovas do tipo I possuem tipicamente um pico de luminosidade mais estreito, enquanto que do tipo II possuem picos mais amplos, da ordem de 100 dias. Depois do pico, a luminosidade decresce por aproximadamente um ano. A partir dessa classificação inicial foram derivados outros tipos de supernovas. Os esquemas de classificação atuais analisam também a presença de hélio e silício no espectro e o formato da curva de luz para classificar a supernova. Dois desses esquemas, apresentados em (GIUNT; KIM, 2007) e em (TURATTO; BENETTI; PASTORELLO, 2007), têm pelo menos sete diferentes tipos de supernovas: Ia, Ib, Ic, IIb, IIL, IIP e IIn.

Outro aspecto importante na classificação das supernovas é a estimativa da taxa de formação de supernovas que dependem fortemente do tipo de galáxia hospedeira onde esses eventos ocorrem.

Figura 1 – Tabela descritiva para os diferentes tipos e subtipos de supernovas (Wikipédia)

Type I No hydrogen	Type Ia Presents a singly ionized silicon (Si II) line at 615.0 nm (nanometers), near peak light		Thermal runaway
	Type Ib/c Weak or no silicon absorption feature	Type Ib Shows a non-ionized helium (He I) line at 587.6 nm Type Ic Weak or no helium	Core collapse
Type II Shows hydrogen	Type II-P/L/N Type II spectrum throughout	Type II-P/L No narrow lines	
		Type II-P Reaches a "plateau" in its light curve Type II-L Displays a "linear" decrease in its light curve (linear in magnitude versus time). ^[50]	
		Type IIn Some narrow lines	
	Type Iib Spectrum changes to become like Type Ib		

As linhas de emissão de supernovas do tipo Ia são devido principalmente ao decaimento do Ni^{56} , que é produzido em grandes quantidades durante o colapso das camadas mais externas do núcleo da anã branca. Contém também absorções de Si , que é produzido pela fusão do carbono e do oxigênio. O Ni^{56} decai em Co^{56} , que por sua vez decai em Fe^{56} , cujas linhas de emissão dominam o espectro após alguns meses. Supernovas do tipo Ia possuem um padrão de comportamento que foi identificado por algumas explosões de supernova em galáxias vizinhas. As curvas de luz (principalmente no espectro visível) caem praticamente sempre à mesma taxa e o total de energia emitido na explosão são aproximadamente iguais. Em 1993, (PHILLIPS, 1993) conseguiu empiricamente encontrar uma relação entre a fase de observação do pico de luminosidade da curva de luz (tempo do quadro de observação) e a luminosidade, a partir de um catálogo de supernovas do tipo Ia próximas com distâncias conhecidas. Esta relação nos permite usar as supernovas do tipo Ia como "velas padrão" para as medidas de distâncias cosmológicas.

1.3 Classificação espectroscópica e fotométrica

Como mencionamos anteriormente, futuras pesquisas em larga escala sobre supernovas enfrentarão o desafio de identificar os candidatos em grande parte sem o uso de espectroscopia. Ao longo dos últimos anos, uma série de métodos para a identificação de supernovas apenas através de informações fotométricas foram propostos. Alguns desses esquemas usam diagramas de cores ou cor-magnitude; outros simplesmente ajustam modelos aos dados de supernovas. Ambas as abordagens sofrem uma série de inconvenientes parcialmente solucionados na conhecida técnica Bayesiana para a classificação de supernovas. No entanto, as técnicas Bayesianas também são problemáticas na medida em que exigem que o candidato a supernova seja pertencente a um dos conjuntos conhecidos de tipos de supernovas. Isso apresenta uma série de problemas, o mais óbvio, são objetos que não

estão em conformidade com nenhum modelo atualmente conhecido nas grandes amostras de candidatos à supernova. No trabalho de (CONNOLLY; CONNOLLY, 2009) é proposto um novo esquema de classificação fotométrica que utiliza um fator Bayes baseado em cor para identificar supernovas por tipo. Este método não requer conhecimento do conjunto completo de possíveis objetos astronômicos que poderiam imitar um sinal de supernova, outro recente trabalho nessa área é (LECLERCQ, 2018) que utiliza o método da *likelihood-free* na inferência cosmológica.

1.4 Dados de supernovas correlacionados

A alta qualidade e quantidade de dados de Supernovas que serão coletados na próxima década (2×10^6 Supernovas previstas pelo LSST), permitirá a identificação de correlações até então não compreendidas devido à diversidade e complexidade de vários fatores contribuintes. Esquemáticamente as correlações podem ser devido a:

- Velocidades peculiares: quando as supernovas estiverem dentro de 50 Mpc aproximadamente umas das outras, a velocidade peculiar da galáxia hospedeira poderá ser correlacionada. Estas velocidades peculiares causam erros correlacionados de redshift, esses erros normalmente são convertidos em erros de módulo de distância adicionais (DAVIS, 2011).
- Cor-Redshift e Estiramento-Redshift: Se não houver observação do redshift espectroscópico da galáxia hospedeira, o redshift pode ser obtido fotometricamente a partir da curva de luz de uma SNIa, porém o efeito do redshift na curva de luz é degenerar as correlações de estiramento da distribuição espectral e cor afetando a estimativa do módulo da distância para a supernova (PALANQUE-DELABROUILLE, 2010).
- Erros de filtros: Erros na medida do ponto zero dos filtros de observação são uma fonte importante de incerteza nos levantamentos atuais que são iguais para todas as supernovas observadas com o mesmo telescópio (FEINDT; KOWALSKI; PAECH, 2012)
- Erros de correlações do modelo: evolução de erros correlacionados no redshift de supernovas devido a erros nas curvas de luz que formam o conjunto de treinamento. Um exemplo disso é a "anomalia U-band" que causa discrepâncias entre o SALT2 e ajustadores de curvas de luz MLCS2k2 que podem estar relacionados a um excesso de fluxo no UV de supernovas Ia em altos redshifts (FOLEY, 2012), (KESSLER, 2009).
- Condições de observação: mau tempo causará furos na cobertura da curva de luz de todas as supernovas visíveis em um determinado momento, quando as condições de visão afetarão as medições fotometricas de forma correlacionada. Isso induzirá correlações sutis entre objetos observado na mesma noite em condições semelhantes (TYSON; STRAUSS, 2009).

- Combinação de dados de vários telescópios: a matriz de covariância para combinar dados de múltiplos telescópios pode ser muito complexa, conforme discutido na análise SNLS de 3 anos em (CONLEY et al., 2010). Com exceção talvez do conjunto de dados finais do LSST, combinar dados de múltiplas pesquisas continua sendo uma estratégia padrão.
- Lentes gravitacionais: Supernovas que estão às mesmas distâncias do observador, isto é, em redshifts semelhantes, podem apresentar brilhos com intensidades similares ou sofrerem algum escurecimento devido à lentes, dependendo da distribuição de matéria ao longo da linha de visada. Futuras pesquisas terão mapas de distribuição de massa e serão capazes de prever e remover este efeito até certo ponto (AMENDOLA, 2010).
- Poeira: as supernovas vizinhas ao longo da linha de visada, sofrerão extinção semelhante a ocorrida na Via Láctea, mas devido meio intergaláctico o que também induzirá correlações (CORASANITI, 2006).
- Correlações entre galáxias hospedeiras: Existem evidências sólidas de que a dispersão no diagrama de Hubble correlaciona-se com as propriedades da galáxia hospedeira, particularmente com tipo de galáxia e metalicidade (KELLY, 2010).
- Correlações de alvos espectroscópicos: Desde que o acompanhamento fotométrico não seja aleatório, pode haver correlações ocultas. Por exemplo, o acompanhamento pode favorecer os candidatos bem separados do núcleo da galáxia hospedeira. Esse efeito, conhecido como viés de Malmquist (que também está relacionado com o fato de ser mais fácil detectar objetos intrinsicamente mais brilhantes em altos redshifts), pode causar correlações que dependem dos detalhes da pesquisa espectroscópica (SAKO, 2008).

1.5 Não homogeneidade das Supernovas

Desde o debate Curtis-Shapley de 1920 (SHAPLEY; CURTIS, 1921), a determinação das luminosidades de supernovas tem sido fundamental para a discussão de distâncias extragalácticas. Eles foram os primeiros a proporem uma distinção entre novas e supernovas que foi posteriormente explicitado em (BAADE; ZWICKY, 1934). Baade e Zwicky mostraram que, além da grande diferença em luminosidade absoluta, o comportamento fotométrico e espectroscópico das supernovas é distinto das novas. O trabalho de Baade (BAADE, 1938) mostrou que as supernovas eram mais uniformes do que novas, apresentando uma pequena dispersão no pico de magnitudes, tornando-as adequadas como indicadores de distância extragalácticos. A precisão das estimativas de distância da supernova aumentou à medida que a classe SN Ia foi melhor entendida e mais estreitamente definida. A baixa dispersão na amostra de "Baade" foi favorecida pela ausência fortuita de supernovas Tipo II observadas, que são significativamente menos luminosas na média. A partir do evento SN 1940B, as supernovas Tipo II foram classificadas pela presença de hidrogênio em seus espectros (MINKOWSKI, 1941). A crescente lista de supernovas espectroscopicamente definidas

de Tipo I apresentaram dispersões no pico entre 0.6 e 0.8 magnitudes ((MINKOWSKI, 1964; KOWAL, 1968; KIRSHNER et al., 1973; OKE; SEARLE, 1974). No entanto, esta amostra incluiu um número de peculiares supernovas Ia notáveis pela falta de silício, que agora são reconhecidas como provenientes de estrelas maciças que perdem o envelope antes do colapso de núcleo (WHEELER; HARKNESS, 1990). Depois de remover esses objetos deficientes em silício, agora classificados como Ib e Ic's (DOGETT; BRANCH, 1985; UOMOTO; KIRSHNER, 1985; WHEELER; LEVREAU, 1985; WHEELER et al., 1986), as SN Ia restantes formaram um conjunto mais homogêneo tornando-se os indicadores mais precisos para estimativas de distâncias cosmológicas. (LEIBUNDGUT; TAMMANN, 1989) desenvolveram um conjunto de modelos padrões para descrever o comportamento fotométrico de SN Ia e estimar o magnitude aparente do pico.

A hipótese de que as supernovas Ia são velas padrão atraiu apoio não só de dados observacionais, mas também de modelos teóricos que sugeriram uma relação entre o brilho intrínseco de qualquer supernova e o processo de ignição de uma estrela anã branca no limite de massa de Chandrasekhar (HOYLE; FOWLER, 1960; ARNETT, 1969; COLGATE; MCKEE, 1969). Nesses modelos, uma onda de choque supersônica viaja através da matéria degenerada que constitui a anã branca, queimando o material em ^{56}Ni a altíssimas temperaturas (WHEELER; MAZUREK, 1980; HOFLICH; MULLER; KHOKHLOV, 1993). A detonação é supersônica, devido ao fato de a região pré-choque não poder expandir para diminuir a pressão ou temperatura de queima. Além disso, a *pressão de fermi* do material degenerado na região de choque permanece insensível à temperatura por mais tempo do que a escala de tempo de queima. O resultado é uma incineração total e a produção de uma massa pura de níquel. Esse padrão de explosão uniforme de massa é o que leva a uma curva de luz homogênea e a uma luminosidade uniforme. No entanto, esses modelos completos de queima não reproduzem os elementos de massa intermediários que são vistos nos espectros de SN Ia (WHEELER; HARKNESS, 1990). Um modelo bem sucedido (NOMOTO; THIELEMAN; YOKOI, 1984) que correspondeu as restrições observacionais foi tão persuasivo que Arnett, Branch e Wheeler (1985) (ARNETT; BRANCH; WHEELER, 1985) e Branch (1992) (BRANCH, 1992) sugeriram a calibração da constante de Hubble baseando-se apenas em modelos teóricos de produção uniforme de níquel. No entanto, uma variedade de modelos (LIVNE, 1990; HOFLICH; MULLER; KHOKHLOV, 1993; P.HOFLICH; KHOKHLOV; WHEELER, 1995) combinam as características observadas do espectro e produzem uma gama de massas de níquel e uma gama de luminosidades previstas. Estes modelos empregam frentes de deflagração subsônicas, pulsações dentre outros. O sucesso desses modelos na reprodução dos espectros observados abre uma grande variedade de possibilidades teóricas. Ao contrário dos primeiros modelos monoenergéticos, esses modelos sugerem que uma grande variedade de luminosidades pode resultar da ignição de uma anã branca.

Observações recentes feitas com detalhes por detectores CCD (charge-coupled device)

evidenciam a não homogeneidade tanto na luminosidade quanto na forma da curva de luz das supernovas. Mesmo antes das recentes observações de não-homogeneidade, medidas fotométricas menos precisas feitas em (BARBON; CIATTI; ROSINO, 1973) sugeriram que existem duas classes fotométricas; a primeira com taxas de declínio "rápidas" após o máximo, que eram supernovas intrinsecamente mais brilhantes, e a segunda com taxas de declínio "lentas" que seriam mais fracas. Com a vantagem de melhores dados medidos, foi possível demonstrar evidências conclusivas para um declínio da luminosidade na curva de luz e o tipo de SNIa (PHILLIPS, 1993). Usando um conjunto de curvas de luz SNIa bem amostradas com fotometria precisa e distâncias relativas precisas, Phillips descobriu que a luminosidade absoluta em B, V e I estão correlacionadas com o declínio da banda B durante aproximadamente os quinze dias seguintes à luz máxima.

Esta relação empírica entre curvas de luz e luminosidade em SN Ia foi estimulada por uma abundância de modelos teóricos que podem explicar o comportamento observado (P.HOFLICH; KHOKHLOV; WHELLER, 1995; HOFLICH; KHOKHLOV, 1996; LIVNE; ARNETT, 1995). Tais modelos envolvem estudos em astrofísica estelar para compreensão das frentes de deflagração de queima, detonações fora do centro da estrela progenitora, queima de halo superficial de hélio, detonações tardia pulsada e progenitoras sub-Chandrasekhar (WOOLEY; WEAVER, 1994).

1.6 Magnitudes

As magnitudes comuns em todas observações são magnitudes aparentes e magnitudes absolutas. Além disso, para alguns casos específicos, (como medidas de magnitudes galácticas) pode se optar entre outras formas como, magnitudes isofotais, petrosians, métricas, magnitudes totais e de abertura.

O ato de se magnificar os objetos celestes remonta ao astrônomo, construtor, cartógrafo e matemático grego Hiparco (190 a.c - 120 a.c) que teria atribuído às estrelas mais brilhantes do céu a primeira magnitude $m = 1$, enquanto às estrelas mais tênues $m = 6$ (no limite da percepção visual humana). Cada magnitude era subjetivamente definida como o dobro do brilho da magnitude seguinte (escala logarítmica), já que não existiam instrumentos de medição para o propósito naquela época.

Foi somente em 1856 que o astrônomo inglês Norman Robert Pogson (1829-1891) formalizou o sistema como conhecemos hoje, definindo que uma estrela de primeira magnitude é 100 vezes mais brilhante do que uma estrela de sexta magnitude. Quanto mais tênue um objeto parece, maior é o valor numérico da sua magnitude aparente, sendo que uma diferença de cinco magnitudes corresponde a um fator de brilho de exatamente 100 vezes. Assim a magnitude aparente m foi definida para representar o brilho aparente de objetos celestes sem nenhum tipo de correção para os possíveis efeitos sofridos pela radiação ao longo da

linha de visada. Desta forma, em uma dada banda de passagem genérica x , a magnitude aparente é dada por:

$$m_x = -5 \log_{10} \left(\frac{F_x}{F_{x,0}} \right) \quad (1.49)$$

onde F_x é o fluxo luminoso observado usando o filtro x e $F_{x,0}$ é o fluxo de referência (que define o ponto zero de magnitude) para o filtro específico. O Fluxo luminoso (definido pela lei do inverso do quadrado da distância) é $F = L/4\pi d^2$ onde L é a luminosidade intrínseca e d a distância, (a luminosidade é a quantidade de energia que um corpo irradia por unidade de tempo, ou seja, potência irradiada, ela é tipicamente expressa em Watts). A unidade usual de fluxo é o $W.m^{-2}$ e a densidade de fluxo é medida em unidades de energia por área, sua unidade usual é o *Jansky* ($1Jy = 10^{-23} erg.cm^{-2}.s^{-1}.Hz^{-1} = 10^{-26} \frac{W}{m^2.Hz}$).

A magnitude aparente é mais comumente expressa em termos do logaritmo padrão (base 10):

$$m_x = -2,5 \log_{10} \left(\frac{F_x}{F_{x,0}} \right) \quad (1.50)$$

A razão de Pogson ($F_2/F_1 = 100^{\frac{\Delta m}{5}} = 10^{0,4\Delta m} \approx 2,512^{\Delta m}$) indica que cada aumento unitário de magnitude corresponde a um decréscimo de brilho por um fator de 2,512.

Já a magnitude absoluta é a medida do brilho ou luminosidade intrínseca de um objeto celeste. Definida como a hipotética magnitude aparente do objeto a uma distância padrão de exatamente 10 parsec (32,6 anos-luz) do observador em repouso (redshift zero), assumindo-se não haver nenhum tipo de efeito de extinção da luz. Assim podemos medir os objetos em uma base comum, o que permite a comparação da verdadeira emissão de objetos astronômicos, com correções de distorções provocadas por extinção. Como a maioria das magnitudes astronômicas, a magnitude absoluta pode ser especificada para diferentes intervalos de comprimentos de onda. Para estrelas, a mais comumente utilizada é a *magnitude visual absoluta*, que usa a banda visual do espectro (sistema Johnson-Morgan, *UBV*, *Ultraviolet-Blue-Visual system*) (JOHNSON; MORGAN, 1953). Apesar de não ser medida, outra magnitude também utilizada com frequência é a *magnitude bolométrica absoluta* (M_{bol}), que representa a luminosidade total expressa em unidades de magnitude, levando em conta a energia irradiada para todas as frequências. Devido ao fato de que simplesmente não conseguimos medir magnitudes bolométricas, mas apenas magnitudes para alguns filtros, necessitamos introduzir a chamada correção-K em nossa descrição (COELHO et al., 2014).

Se assumirmos a insignificância dos efeitos devido a extinção (como por exemplo, em uma aproximação para objetos dentro da Via Láctea) podemos escrever a relação entre

magnitude absoluta e magnitude aparente (ambas bolométricas, sem correção-K) para um objeto a uma distância d , na forma:

$$100^{\frac{m-M}{5}} = \frac{F_{10}}{F} = \left(\frac{d}{10pc} \right)^2 \quad (1.51)$$

onde F é o fluxo de radiação medido à distância d (em parsecs) e F_{10} é o fluxo de radiação medido à distância de 10 parsecs. Também podemos reescrever essa relação em termos de logaritmos:

$$M = m - 5 \log_{10} \left(\frac{d}{10pc} \right) \quad (1.52)$$

Para objetos fora de nossa galáxia, devemos usar a distância de luminosidade D_L no lugar de d . Além disso, para uma descrição mais realista da magnitude absoluta devemos considerar o filtro de referência e a respectiva correção-K. Assim, podemos reescrever:

$$M_y = m_x - 5 \log_{10} \left(\frac{D_L}{10pc} \right) + K_{yx}, \quad (1.53)$$

onde K é a correção-K ($K_{yx} = m_x - M_y - \mu$ (COELHO et al., 2014)) e D_L é a distância de luminosidade ($D_L = \sqrt{L/4\pi F}$ ou em termos do redshift e da distância transversal comóvel $D_L = (1+z)D_M$ (na métrica FLRW)) (PERLICK, 2000). Quanto mais distante observamos, mais estaremos sujeitos a anomalias de extinção da luminosidade devido a diversos fatores, desde efeitos provocados pela atmosfera terrestre, por poeiras e gases galáticos e intergaláticos, até efeitos de distorções provocados por lentes gravitacionais (PERLICK, 2010) e pela própria expansão do Universo (SCHMIDT et al., 1998).

Com equipamentos 100 % sensíveis a todos os comprimentos de onda, teoricamente poderíamos obter o fluxo em todo o espectro. A magnitude correspondente ao fluxo bolométrico F_{bol} em todos os comprimentos de onda é a magnitude bolométrica M_{bol} .

$$F = \frac{L}{4\pi d^2} \implies L = 4\pi d^2 \int_0^\infty F_\nu d\nu = 4\pi d^2 F_{bol} \quad (1.54)$$

onde F_ν é o fluxo luminoso na frequência ν . Sua integração de zero a infinito representa a soma das medidas dos fluxos para cada ν . Na prática, é impossível medir a magnitude bolométrica devido aos efeitos de extinção da radiação por espalhamento ou absorção, de forma, que normalmente essa magnitude é aproximada através da diferença entre a magnitude absoluta visual (M_V) e um fator de *correção bolométrica* (CB).

$$M_{bol} = M_V - CB \quad (1.55)$$

Como qualquer outra grandeza, magnitudes também precisam de referências. Na magnitude absoluta normalmente é considerada como referência a estrela α Lyrae (Vega) que possui magnitude aparente $m \approx 0$. Já na magnitude bolométrica a referência é o Sol, que possui correção bolométrica $CB = 0$. Desse modo, o Sol tem o mesmo valor 4,72 magnitude absoluta igual a magnitude bolométrica. Todas as outras estrelas possuem CB maior ou igual a zero, assim a magnitude bolométrica é sempre igual ou menor que a magnitude visual.

Quando medimos o brilho de uma estrela, o fluxo obtido depende da sensibilidade espectral do equipamento, ou seja, do conjunto de espelhos filtros e detectores. Se chamamos de $\Phi(\lambda)$ a eficiência espectral do equipamento normalizada, temos:

$$F_{obs} = \int_0^\infty \Phi(\lambda) F(\lambda) d\lambda \approx F(\lambda_0) \int_0^\infty \Phi(\lambda) d\lambda \quad (1.56)$$

onde $F(\lambda_0)$ é o fluxo no comprimento de onda efetivo do filtro. Um sistema de magnitudes é definido por $\Phi(\lambda)$ e por uma constante. Para encontrar a constante do sistema usamos estrelas padrões, ou seja, estrelas que têm magnitudes bem determinadas.

1.6.1 Sistema VEGA ou Jonhson

A referência de fluxo nesse sistema é definida com base na magnitude da estrela α Lyr (Vega) considerando sua magnitude visual $m_V \sim 0,03$ e em todas as outras bandas iguais a zero. Alternativamente, o ponto zero de magnitude pode ser definido como a média de uma série de magnitudes de estrelas da classe espectral A0, usando o conjunto de filtros Johnson-Morgan para corrigir a escala de fluxo. Algumas outras estrelas também são apropriadas, tal como a calibração realizada por Hayes e Lathan (1975) (HAYES; LATHAM, 1975), que produziu 3.500 Jy em 5556Å para α Lyr. Outros artigos discutem as bandas de passagem (BESSEL, 1991), (CLARET; DÍAS-COROVÉS; GIMÉNEZ, 2007), (LANDOLT, 2007).

O sistema Johnson original consistia apenas nos filtros UBV cuja calibração estava intimamente ligada aos detectores fotoelétricos disponíveis na época. O sistema foi desde então estendido para o vermelho com os filtros ópticos RI e infravermelhos próximos JHK . As definições desses filtros nem sempre são independentes dos detectores envolvidos e podem variar um pouco de observatório para observatório. Os filtros JHK foram utilizados na pesquisa *2MASS* (<https://www.ipac.caltech.edu/2mass/>) e são uma extensão importante do sistema Johnson para os comprimentos de onda do infravermelho próximo. A tecnologia requer diferentes detectores para esses comprimentos de onda, logo são necessárias diferentes estrelas de calibração. O *2MASS* usa um filtro "curto"K que é ligeiramente diferente do filtro K definido originalmente, mas que atualmente é comumente usado devido sua superior supressão de emissões térmicas terrestres.

1.6.2 Sistema Gunn ou griz

Originalmente definido em termos de detectores fotoelétricos (THUAN; GUNN, 1983) (WADE; HOESSEL; ELIAS, 1979), mas agora é usado principalmente com CCDs (SCHNEIDER; GUNN; G, 1983). O sistema *griz* é definido por algumas dúzias de estrelas padrões, além da estrela $BD + 17deg4708$, uma estrela *subdwarf* de classe espectral $F6$ e com $M_B - M_V = 0,43$, sendo definida como tendo todas as cores iguais a zero. A calibração absoluta deste sistema é simplesmente o fluxo monocromático da estrela (OKE; GUNN, 1983), escalado de $g = 9.50$ para $g = 0.0$, nos comprimentos de onda efetivos das bandas *griz*. Uma série de aspectos detalhados da fotometria de banda larga no contexto específico das galáxias em altos redshifts são revisados em (SCHNEIDER; GUNN; G, 1983).

1.6.3 Sistema AB

A magnitude monocromática absoluta (AB) é definida como o logaritmo de uma densidade de fluxo espectral com a escala usual de magnitudes astronômicas e densidade de fluxo igual a 3631 Jy para o ponto zero de magnitude (OKE; GUNN, 1983). Se a densidade de fluxo espectral for denotada por f_ν , a magnitude AB monocromática pode ser escrita (desconsiderando correções de extinção):

$$m_{AB} = -2,5 \log_{10} \left(\frac{f_\nu}{3631} \right) \quad (1.57)$$

ou em unidades do *cgs* ($ergs^{-1}cm^{-2}Hz^{-1}$),

$$m_{AB} = -2,5 \log_{10} f_\nu - 48.60 \quad (1.58)$$

Em alguns casos, as densidades de fluxo espectral são expressas por unidade de comprimento de onda, f_λ , em vez de frequência f_ν , utilizando a relação,

$$\nu f_\nu = \lambda f_\lambda \implies f_\nu = \frac{\lambda^2}{c} f_\lambda \quad (1.59)$$

As medidas reais são feitas em algum intervalo contínuo de comprimento de onda. A amplitude da banda de passagem em AB é definida para o ponto zero corresponder a uma densidade de fluxo espectral de 3631 Jy,

$$m_{AB} = -2,5 \log \left(\frac{\int f_\nu (h\nu)^{-1} A(\nu) d\nu}{3631 \int (h\nu)^{-1} A(\nu) d\nu} \right) \quad (1.60)$$

onde o termo $(h\nu)^{-1}$ assume que o detector é um dispositivo de contagem de fótons, como detectores CCD ou fotomultiplicadores e $A(\nu)$ é a seção transversal de captura definida em (TONRY et al., 2012).

1.6.4 Sistema STmag

O sistema *STmag* é definido de maneira similar ao *AB*, mas para o fluxo constante por unidade de intervalo de comprimento. Este sistema de magnitude é definido de tal forma que um objeto com fluxo constante por intervalo de comprimento de onda unitário tem cor zero. É o padrão adotado pelos pacotes de fotometria do Hubble Space Telescope (DEUSTUA et al., 2017), para mais detalhes e exemplos de aplicações na linguagem python consulte também SynPhot User's Guide (www.stsci.edu/institute/software_hardware/stsdas/synphot/SynphotManual.pdf).

Os sistemas STmag e AB definem que a densidade do fluxo de uma determinada fonte, deve corresponder à densidade de fluxo de uma fonte de forma espectral predefinida. A conversão é feita de forma que a magnitude visual corresponda aproximadamente a do sistema Jonhson. No sistema STmag, o fluxo é expresso em função de λ ao invés de ν e a densidade de fluxo referência é f_λ . Um objeto observado em um respectivo filtro, terá ponto zero de magnitude em $m = 21.10$, logo:

$$m_{STmag} = -2,5 \log_{10} f_\lambda - 21.10 \quad (1.61)$$

Veja Bohlin 2014 (BOHLIN, 2014), para mais detalhes sobre a definição de fluxo e densidade de fluxo.

1.6.5 Sistema natural - LSST

Uma das limitações da fotometria multibanda é que as medidas de fluxo F_b^{meas} não podem ser relacionadas com a precisão de $F_\nu(\lambda)$ a menos que a eficiência espectral para cada banda de observação ($\phi_b(\lambda)$) seja conhecida. Outra limitação adicional é que F_b^{meas} pode variar mesmo quando $F_\nu(\lambda)$ é constante porque geralmente ϕ_b é variável. Tradicionalmente, esta variação de fluxo é calibrada usando termos de extinção atmosférica e cor, que funcionam bem para fontes com distribuições de energia espectral relativamente suaves. Em outras palavras, estritamente esse efeito não pode ser calibrado a menos que à forma da distribuição espectral de energia (SED) da fonte seja,

$$f_\nu(\lambda) = F_\nu(\lambda)/F_0 \quad (1.62)$$

onde F_0 é uma normalização, conhecida previamente. Se $f_\nu(\lambda)$ for conhecido para uma função normalizada de resposta padrão, ϕ_b^{std} (obtido pela média de um conjunto apropriado de ϕ_b^{meas} dos dados acumulados pela pesquisa em datas periódicas), as medidas expressas sobre o sistema fotométrico podem ser padronizadas através da relação:

$$m_b^{std} - m_b^{meas} \equiv \Delta m^{std} = 2.5 \log \left(\frac{\int_0^\infty f_\nu(\lambda) \phi_b^{meas}(\lambda) d\lambda}{\int_0^\infty f_\nu(\lambda) \phi_b^{std}(\lambda) d\lambda} \right) \quad (1.63)$$

Embora essa transformação seja em princípio exata, m_b^{std} herda os erros de medição em m_b^{meas} , bem como um erro adicional devido à diferença entre o verdadeiro $\phi_b(\lambda)$ e o medido ϕ_b^{meas} , que vai ser o usado na prática. Incertezas no conhecimento de $f_\nu(\lambda)$ contribuirão com um termo de erro adicional para m_b^{std} . Dependendo do estudo se poderá escolher corrigir m_b^{meas} usando o pré-calculado Δm^{std} para distribuições espectrais de energia típicas (vários tipos de galáxias, estrelas, objetos do sistema solar, média de distribuições espectrais de energia para quasares, etc), ou calcular o valor de Δm^{std} para uma escolha particular de $f_\nu(\lambda)$ (TYSON; STRAUSS, 2009).

1.7 Fotometria

A fotometria é uma técnica de observação astronômica que relaciona a medida do fluxo ou intensidade da radiação eletromagnética em diferentes bandas do espectro. Observações fotométricas podem ser processadas por fotometria relativa ou absoluta. A *fotometria relativa* é a medida do brilho aparente de múltiplos objetos em relação uns aos outros e a *fotometria absoluta* é a medida do brilho aparente de um objeto em um sistema fotométrico padrão.

Um sistema fotométrico assim como um sistema de magnitude, é composto por um conjunto de filtros com sensibilidade conhecidas para certas bandas de passagem com comprimentos bem definidos. A sensibilidade é um efeito combinado do sistema óptico, dos detectores e dos filtros utilizados. O primeiro sistema fotométrico padronizado conhecido é o sistema Johnson-Morgan ou *UBV* (1953), atualmente existem mais de 200 sistemas fotométricos (em SNCosmo encontramos um resumo dos sistemas mais conhecidos <<http://sncosmo.readthedocs.io/en/v1.5.x/bandpass-list.html>>).

Cada letra fotométrica representa uma determinada seção do espectro eletromagnético que abrangem basicamente desde o ultravioleta passando pelo visível, até o infravermelho. Por exemplo, os espectros de cores laranja, amarelo e verde estão na banda visual (*V*), enquanto violetas e roxas estão na banda azul (*B*). A seguir, apresentamos uma lista:

Banda	$\lambda_{eff}(nm)$	$\Delta\lambda$ (nm)	Descrição	variantes
<i>U</i>	365	66	Ultravioleta	<i>u, u', u*</i>
<i>B</i>	445	94	Azul	<i>b</i>
<i>V</i>	551	88	Visível	<i>v, v'</i>
<i>G</i>	464	128	Verde visível	<i>g, g'</i>
<i>R</i>	658	138	Vermelho	<i>r, r', R', ...</i>
<i>I</i>	806	149	Infravermelho	<i>i, i', ...</i>
<i>Z</i>	900		Infravermelho próximo	<i>z, z'</i>
<i>Y</i>	1020	120	Infravermelho próximo	<i>y</i>
<i>J</i>	1220	213	Infravermelho próximo	<i>J', J_s</i>
<i>H</i>	1630	307	Infravermelho próximo	
<i>K</i>	2190	390	Infravermelho próximo	<i>K', K_s, K^8, nbK, ...</i>
<i>L</i>	3450	472	Infravermelho próximo	<i>L', nbL'</i>
<i>M</i>	4750	460	Infravermelho médio	<i>M', nbM</i>
<i>N</i>	10500	2500	Infravermelho médio	
<i>Q</i>	21000	5800	Infravermelho médio	<i>Q'</i>

Aqui λ_{eff} é o comprimento de onda médio efetivo para o filtro padrão e $\Delta\lambda$ é a largura da banda de passagem em torno de λ_{eff} . As variantes de cada letra fotométrica representam valores um pouco diferentes de λ_{eff} e $\Delta\lambda$, elas variam conforme o sistema fotométrico adotado.

1.7.1 Calibração da Fotometria LSST

O Telescópio LSST (TYSON; STRAUSS, 2009) tem detectores de banda larga (CCD) que serão capazes de detectar a integral da densidade de fluxo luminoso para fontes específicas, como se as observássemos do topo da atmosfera terrestre, essa densidade de fluxo ($F_\nu(\lambda)$) é dada em Jansky (1 Jansky = $10^{-26} \text{ W Hz}^{-1} \text{ m}^{-2}$). A densidade de fluxo corrigida pela função de resposta normalizada ϕ_b (que inclui os efeitos da atmosfera e todos os elementos ópticos do telescópio) fornece,

$$F_b = \int_0^\infty F_\nu(\lambda) \phi_b(\lambda) d\lambda, \quad (1.64)$$

onde o índice b corresponde a um dos filtros de banda ($b = \text{ugrizy}$, para o LSST). A unidade escolhida para o fluxo F_b é W.m^{-2} e por definição, $\int_0^\infty \phi_b(\lambda) d\lambda = 1$. A magnitude correspondente é dada por:

$$m_b = -2.5 \log_{10} \left(\frac{F_b}{F_{AB}} \right) \quad (1.65)$$

O fluxo normalizado $F_{AB} = \int_0^\infty 3631 \phi_b(\nu) d\nu$ segue o padrão de (OKE; GUNN, 1983).

A função de resposta normalizada corresponde à relação:

$$\phi_b(\lambda) \equiv \frac{\lambda^{-1}T_b(\lambda)}{\int_0^\infty \lambda^{-1}T_b(\lambda)d\lambda} \quad (1.66)$$

O termo λ^{-1} reflete o fato de que os sensores CCD utilizados na câmera são dispositivos contadores de fótons ao invés de calorímetros (TYSON; STRAUSS, 2009). Na relação anterior $T_b(\lambda)$ é a função de resposta do sistema,

$$T_b(\lambda) = T_b^{instr}(\lambda) \times T^{atm}(\lambda), \quad (1.67)$$

onde T^{atm} é a transmitância óptica da atmosfera para o conjunto óptico do telescópio e T^{instr} é a função de resposta do sistema instrumental da "pupila" do telescópio (TYSON; STRAUSS, 2009). Observe que as normalizações gerais de T_b^{instr} e T^{atm} são canceladas em $\phi_b(\lambda)$. Dessa forma, os resultados básicos da fotometria LSST serão reportados em um *sistema fotométrico natural*, isso significa que para cada medida fotométrica F_b^{means} uma correspondente medida da função de resposta normalizada $\phi_b^{means}(\lambda)$ também será fornecida juntamente com as estimativas de erros para ambos. A pesquisa tem previsão de coletar $\approx 10^{12}$ pares ($F_b^{means}, \phi_b^{means}(\lambda)$), um par para cada fonte detectada.

Uma característica inevitável da fotometria terrestre é que a função de resposta normalizada $\phi_b(\lambda)$, varia com o tempo, a posição no céu e o detector, devido a variações nas formas dos perfis espectrais de T^{atm} e T_b^{instr} . Tradicionalmente, esses efeitos são calibrados usando um conjunto de estrelas padrões. Entretanto, dados observacionais existentes (por exemplo, do SDSS (KESSLER, 2009)) demonstraram que este método não fornece a precisão fotométrica necessária em condições gerais de observação. Por isso, o sistema LSST medirá $T^{atm}(\lambda)$ e $T_b^{instr}(\lambda)$ (produzindo medidas de S^{atm} e S_b^{instr}) nas escalas ondulatórias, temporais e angulares relevantes. Os valores de S_b^{instr} incluem a reflexividade dos espelhos, transmissão da óptica refrativa e filtros, eficiência quântica, desempenho e linearidade da leitura eletrônica dos sensores na câmera. No LSST um conjunto de diodos precisos serão aplicados para normalizar o fluxo de fótons integrado durante as exposições. Esses fotodiodos, juntamente com sua leitura eletrônica, serão calibrados no US National Institute of Standards (NIST - <<https://www.nist.gov/>>) para $\approx 0,1\%$ de precisão relativa em comprimentos de onda de $450nm$ a $950nm$. A resposta desses diodos varia suavemente em toda essa faixa de comprimentos de onda fornecendo uma referência "bem comportada" (STUBBS et al., 2004). O ajuste do comprimento de onda da fonte de luz pode ser tão bom quanto um nanômetro, e permitirá um monitoramento preciso da forma das bandas de passagem do sistema instrumental durante o curso da pesquisa (STUBBS; TONRY, 2006). Já o valor de S^{atm} será obtido através do produto entre dois fatores: um determinado a partir de medidas espectroscópicas repetidas de um pequeno conjunto de estrelas de referência por um telescópio auxiliar dedicado (TYSON;

[STRAUSS, 2009](#)), e o outro determinado a partir das imagens obtidas pelo LSST como parte de uma auto-calibração fotométrica global da pesquisa, baseada nos fluxos medidos de um grande conjunto de estrelas de referência que cobrem a área de pesquisa e são observadas durante muitas épocas.

O design LSST inclui um telescópio de calibração auxiliar de 1,2m. A idéia é fazer medições do total de variação espacial e temporal da extinção atmosférica ao longo de cada noite independentemente das operações do telescópio principal de pesquisa. Isso será feito tomando repetidamente espectros de um pequeno conjunto de estrelas referências, enquanto o telescópio atravessar foco sobre o céu todas as noites. Essas estrelas estão espalhadas de forma a cobrirem totalmente a área pesquisada pelo LSST. A calibração usará os bem sucedidos modelos atmosféricos aplicados em ([STUBBS et al., 2007](#)) e códigos prontamente disponíveis (MODTRAN - MODerate resolution atmospheric TRANsmission - modtran.spectral.com) que calculam com precisão as assinaturas espectrais de todos os componentes atmosféricos significativos. Isso permitirá a interpolação da mistura atmosférica presente ao longo de qualquer linha de visada a qualquer momento partindo dos dados atmosféricos coletados. As estrelas de referência serão observadas muitas vezes durante a pesquisa LSST, então a distribuição espectral de energia (SED) de cada estrela poderá ser encontrada a partir dos dados. A resposta instrumental espectrocópica também poderá ser obtida dos dados, que incluem estrelas com uma variedade de SEDs e uma ampla gama de massas de ar ao longo da linha de visada.

A calibração fotométrica para o LSST será realizada em dois níveis com diferentes cadências e com diferentes resoluções de imagem. A primeira será realizada com base no melhor conjunto de dados disponíveis previamente por outras pesquisas observacionais, fornecendo uma melhor precisão. Essa calibração será usada para garantir qualidade na geração de alertas para transientes ([ANTARES-COLLABORATION et al., 2018](#)). Uma análise mais completa será utilizada para recalibração baseada nos dados acumulados pela pesquisa em datas periódicas. Ambos níveis de calibração serão separados em três partes (calibração relativa, calibração absoluta de cor e calibração absoluta de fluxo) que abordam diferentes requisitos específicos:

- Calibração relativa: A precisão da calibração fotométrica relativa do LSST será realizada pela análise de repetidas observações da ordem de 10^8 estrelas selecionadas por brilho durante as operações científicas da pesquisa. O procedimento de calibração em seguida, realiza um pré-processamento dos dados em um subsistema auxiliar de calibração, para determinar a banda de passagem óptica apropriada para cada imagem. Após o pré-processamento de cada imagem acumulada na pesquisa, será executado um procedimento de auto-calibração global que procura minimizar a dispersão dos erros em todas as observações de todas as estrelas de referência. Este processo é baseado em técnicas utilizadas em pesquisas fotométricas anteriores ([GLAZEBROOK et al., 1994](#)), ([MACDONALD et al., 2004](#)). A

específica implementação usada pelo LSST vai ser baseada no procedimento "Ubercal" desenvolvido para o SDSS (PADMANABHAN et al., 2008). As magnitudes medidas das estrelas de referência serão transformadas para o sistema padrão LSST usando as estimativas acumuladas das cores de cada estrela e as correspondentes observações medidas nas bandas de passagem. O procedimento de calibração então realizará a minimização do erro relativo $\delta_b(p, j)$ na fotometria para cada trecho p (a câmera focal plana do LSST será definida para uma calibração por trechos de épocas de observações para controle de erros sistemáticos (PADMANABHAN et al., 2008)) em cada imagem i acumulada na pesquisa j , minimizando o χ^2 :

$$\chi^2 = \sum_{(i,j)} \frac{(m_b^{std,meas}(i,j) - (m_b^{std,true}(i) + \delta_b(p,j)))^2}{(\delta m_b^{std,meas}(i,j))^2}, \quad (1.68)$$

onde as magnitudes estão no sistema padrão, e a soma é sobre todas as estrelas, em todas as imagens j . Os erros $\delta_b(p, j)$ serão usados para corrigir a fotometria para todas as outras fontes no trecho de época p e na imagem j .

- Calibração absoluta de cor e fluxo: Existem seis valores de Δ_b para toda a pesquisa LSST, que definem o "ponto zero" das bandas de passagem padrão para os seis filtros (u, g, r, i, z, y). Estes seis valores podem ser expressos em termos de uma única banda fiducial, por exemplo a banda u,

$$\Delta_b = \Delta_u + \Delta_{bu} \quad (1.69)$$

A estratégia de medir a banda de passagem observacional para cada fonte foi projetada com o intuito de reduzir erros em Δ_{bu} , para cinco das seis bandas, atendendo as especificações e requisitos da pesquisa (TYSON; STRAUSS, 2009). Esse procedimento será validado através do fluxo medido de uma ou mais fontes celestes, muito provavelmente anãs brancas quentes cujas atmosferas simples, são razoavelmente bem compreendidas. A determinação de Δ_u será realizada através de um fluxo externo padrão que ainda está em aberto.

2 Ajustes e simulações de curvas de luz

Neste capítulo, falamos sobre aspectos gerais das simulações fotométricas para supernovas. Destacamos alguns dos métodos e modelos de ajustes mais utilizados, e em seguida, introduzimos alguns dos softwares que se propõem a essa tarefa.

2.1 MLCS2k2

É um método para ajuste de curvas de luz multicolor com aplicação de correções-K. Inicialmente, apenas chamado de LCS (*Light Curve Shape*), foi desenvolvido com o propósito de utilizar a forma de curvas de luz independente da distância para melhorar a estimativa de supernovas tipo Ia. Dada pelo menos uma curva de luz em um filtro, observada dentro de dez dias após o máximo, o método MLCS pode distinguir entre os efeitos de distância, luminosidade intrínseca e poeira para todos os eventos de supernovas (RIESS; PRESS; KIRSHNER, 1996).

2.1.1 Aprendendo sobre curvas de luz (MLCS - Multicolor Light Curve Shape)

As curvas de luz para supernovas são séries temporais desproporcionais e ruidosas, que podem ser obtidas em diferentes bandas do espectro. Embora algumas SN Ia sejam encontradas em galáxias elípticas, a maioria é encontrada em espirais e irregularidades onde podem estar sujeitas a importantes extinções de luminosidade (FILIPPENKO, 1989). Atualmente, o número de curvas de luz SNs bem amostradas disponíveis em sistemas fotométricos moderno está sendo ampliado, mas ainda é limitado, somente 21.909 supernovas têm curvas de luz publicamente disponíveis para serem baixado, e apenas 6.421 possuem espectros ("The Open Supernova Catalog" - <<https://sne.space>> - (GUILLOCHON et al., 1993)).

Para uma primeira aproximação, consideramos que as curvas de luz em diferentes bandas do espectro observado de supernovas Ia são homogêneas, assim o modelo é formado pelo padrão curvas de luz e cor com a adição de ruídos (n), compensações individuais que resultam do módulo da distância (μ) e um excesso de cor (E_{cor}) devido ao "avermelhamento" por extinção:

$$\mathbf{m}_V = \mathbf{M}_V + \mu + \mathbf{n} \quad (2.1)$$

$$\mathbf{m}_B - \mathbf{m}_V = (\mathbf{B} - \mathbf{V})_0 + E_{B-V} + \mathbf{n} \quad (2.2)$$

$$\mathbf{m}_V - \mathbf{m}_R = (\mathbf{V} - \mathbf{R})_0 + E_{V-R} + \mathbf{n} \quad (2.3)$$

$$\mathbf{m}_V - \mathbf{m}_I = (\mathbf{V} - \mathbf{I})_0 + E_{V-I} + \mathbf{n} \quad (2.4)$$

Na convenção usual $\mathbf{M}$ é a magnitude absoluta e $\mathbf{m}$ é a magnitude aparente. O negrito denota que são vetores cujas entradas são medidas ou determinadas em função do tempo, assim como os $\mathbf{n}$ ruídos. Melhorias na qualidade do conjunto de dados disponíveis e o sucesso de Phillips ([PHILLIPS, 1993](#)) motivaram o uso da abordagem de primeira ordem para explicar as variações observadas nas curvas de luz e cor com a luminosidade intrínseca $\mathbf{M}$.

A abordagem mais econômica é adotar um único parâmetro e correlacioná-lo com as variações das curvas observadas. Naturalmente o parâmetro escolhido deve ser a quantidade pela qual a luminosidade intrínseca difere de uma SN Ia com brilho e forma da curva de luz padrão, já que no final, é essa diferença que esperamos medir. Chamamos isso de "correção de luminosidade" (Δ), onde $\Delta \equiv M_V - M_{V_{stand}}$, $M_{V_{stand}}$ é a luminosidade da SN Ia escolhida para descrever o evento *SN Ia "padrão"*. O cálculo de χ^2 é empregado para estabelecer a relação empírica entre as formas das curvas fotométricas e a luminosidade de um subconjunto de SNIa próximas, aparentemente conhecidas com precisão de distâncias e estimativas de extinções. Esta diferença é medida por convenção na magnitude máxima em B ("ponto zero" da escala de tempos). As funções do tempo de primeira ordem em Δ e as correções das curvas observadas com as curvas do modelo são as "correções do modelo" (R_V ou R_{cor}). A suposição de que a luminosidade se correlaciona linearmente com a forma da curva de luz e cor não fornece uma descrição completamente precisa, mas considerando as limitações dos nossos conjuntos de dados reais, é uma maneira razoável de aproximação. O modelo melhorado pode ser assim descrito,

$$\mathbf{m}_V = \mathbf{M}_V + \mu + \mathbf{R}_V \Delta + \mathbf{n} \quad (2.5)$$

onde $\mathbf{R}_V(t)\Delta$ fornece o desvio para a magnitude do evento padrão $\mathbf{M}_V(t)$ em cada tempo. De forma semelhante podemos escrever:

$$\mathbf{m}_B - \mathbf{m}_V = (\mathbf{B} - \mathbf{V})_0 + E_{B-V} + \mathbf{R}_{B-V} \Delta + \mathbf{n} \quad (2.6)$$

$$\mathbf{m}_V - \mathbf{m}_R = (\mathbf{V} - \mathbf{R})_0 + E_{V-R} + \mathbf{R}_{V-R} \Delta + \mathbf{n} \quad (2.7)$$

$$\mathbf{m}_V - \mathbf{m}_I = (\mathbf{V} - \mathbf{I})_0 + E_{V-I} + \mathbf{R}_{V-I} \Delta + \mathbf{n} \quad (2.8)$$

$$(2.9)$$

onde $(\mathbf{B}-\mathbf{V})_0$, $(\mathbf{V}-\mathbf{R})_0$ e $(\mathbf{V}-\mathbf{I})_0$ são os padrões das curvas luz para os filtros (interessante comparar essas equações que permitem SN Ia inhomogêneas, com as equações 2.1, 2.2, 2.3 e 2.4 que impõem a homogeneidade).

Podemos escrever as relações anteriores, na forma matricial:

$$\begin{bmatrix} m_V(t_1) \\ m_V(t_2) \\ \dots \\ m_V(t_N) \\ \\ m_B(t_1) - m_V(t_1) \\ m_B(t_2) - m_V(t_2) \\ \dots \\ m_B(t_N) - m_V(t_N) \\ \\ m_V(t_1) - m_R(t_1) \\ m_V(t_2) - m_R(t_2) \\ \dots \\ m_V(t_N) - m_R(t_N) \\ \\ m_V(t_1) - m_I(t_1) \\ m_V(t_2) - m_I(t_2) \\ \dots \\ m_V(t_N) - m_I(t_N) \end{bmatrix} = \begin{bmatrix} M_V(t_1) \\ M_V(t_2) \\ \dots \\ M_V(t_N) \\ \\ (B-V)_0(t_1) \\ (B-V)_0(t_2) \\ \dots \\ (B-V)_0(t_N) \\ \\ (V-R)_0(t_1) \\ (V-R)_0(t_2) \\ \dots \\ (V-R)_0(t_N) \\ \\ (V-I)_0(t_1) \\ (V-I)_0(t_2) \\ \dots \\ (V-I)_0(t_N) \end{bmatrix} + \begin{bmatrix} 1 & 0 & R_V(t_1) \\ 1 & 0 & R_V(t_2) \\ \dots & \dots & \dots \\ 1 & 0 & R_V(t_N) \\ \\ 0 & \frac{1}{3.1} & R_{B-V}(t_1) \\ 0 & \frac{1}{3.1} & R_{B-V}(t_2) \\ \dots & \dots & \dots \\ 0 & \frac{1}{3.1} & R_{B-V}(t_N) \\ \\ 0 & \frac{1}{3.9} & R_{V-R}(t_1) \\ 0 & \frac{1}{3.9} & R_{V-R}(t_2) \\ \dots & \dots & \dots \\ 0 & \frac{1}{3.9} & R_{V-R}(t_N) \\ \\ 0 & \frac{1}{1.9} & R_{V-R}(t_1) \\ 0 & \frac{1}{3.9} & R_{V-R}(t_2) \\ \dots & \dots & \dots \\ 0 & \frac{1}{3.9} & R_{V-R}(t_N) \end{bmatrix} * \begin{bmatrix} \mu \\ A_V \\ \Delta \end{bmatrix} + n(t)$$

Nessa descrição os termos de medidas simultâneas E_{cor} foram combinados em A_v considerando a parametrização $\frac{A_V}{E(B-V)} = 3.1$, $\frac{A_V}{E(V-R)} = 3.9$, $\frac{A_V}{E(V-I)} = 1.9$, fornecida em (SAVAGE; MATHIS, 1979). Por simplicidade, estamos assumindo que os índices de avermelhamento por "poeira" em nossa galáxia são equivalentes aos efeitos de extinção em galáxias distantes.

Podemos então escrever o ajuste entre os dados e o modelo:

$$\chi^2 = (\mathbf{n})^T \mathbf{C}^{-1} (\mathbf{n}) \quad (2.10)$$

onde n é o vetor de ruídos descrito anteriormente, e $\mathbf{C} = \mathbf{S} + \mathbf{N}$ é a matriz de correlação construída pela soma da matriz de correlação dos sinais ($\mathbf{S}$) com a matriz de correlação dos ruídos ($\mathbf{N}$). A diagonal da matriz $\mathbf{S}$ é formada pelas medições da dispersão, para cada membro do conjunto de treinamento, em torno da média dos conjuntos R_V e R_{cor} obtidos

analiticamente através da minimização de χ^2 independente de $\mathbf{C}$:

$$R_V(t) = \frac{\langle [m_V(t) - M_V(t) - \mu]\Delta \rangle}{\langle \Delta^2 \rangle} \quad (2.11)$$

$$R_{cor}(t) = \frac{\langle [m_{cor}(t) - M_{cor}(t) - E_{cor}]\Delta \rangle}{\langle \Delta^2 \rangle} \quad (2.12)$$

os colchetes indicam uma média ponderada pela incerteza nas estimativas dadas de Δ , sobre o conjunto de treinamento. Já os elementos fora da diagonal requerem mais informações, eles podem representar covariâncias entre observações em dias diferentes com mesmo filtro, em filtros diferentes no mesmo dia, e em filtros diferentes em dias diferentes (além das indesejáveis correlações de medida, que normalmente não são completamente fornecidas pelos observadores e devem ser filtradas).

Uma outra abordagem simples para Δ , seria darmos um peso, aos elementos da diagonal da matrix de correlação S , suficiente para compensar a nossa incapacidade em estimar os termos fora da diagonal. Procuramos pesar os dados em cada banda de passagem pela sua capacidade de prever o parâmetro em comum, Δ (algumas bandas fotométricas podem ser melhor descritas por nosso modelo linear, do que outras na estimativa do parâmetro Δ e os pesos para cada banda devem estar de acordo). Para determinar o peso em cada "cor" (banda fotométricas), permitimos que os pesos, na matriz de correlação, variem e maximizamos a função de log-likelihood para a determinação de Δ ,

$$\mathcal{L} \equiv -\frac{1}{2}(\chi_\Delta^2 - \sum_{i=1}^n \ln \frac{1}{\sigma_{\Delta_i^2}}) \quad (2.13)$$

Maximizar $\mathcal{L}$ é a forma desejada para determinar os parâmetros da matriz de correlação, onde a abordagem convencional de minimizar χ^2 necessariamente conduziria os pesos e χ^2 para zero. Desta forma, ao maximizar $\mathcal{L}$ otimizamos a capacidade dos nossos dados de treinamento em prever o parâmetro Δ , enquanto simultaneamente exigimos que o erro estimado seja razoável. Com a melhora da qualidade e aumento na quantidade dos dados observacionais para SNIa, será possível futuramente estimar toda a matriz de correlação apenas pela maximização de $\mathcal{L}$.

A conveniente escolha de um objeto particular como SNIa "padrão" com curvas de luz e cor que estabelecem a forma das curvas padrão do modelo, M_V e M_{cor} , por definição tem uma variação de pico de luminosidade $\Delta \equiv 0$. A luminosidade absoluta M de uma SNIa é normalmente calibrada com observações de Cefeidas, como por exemplo em (SANDAGE et al., 1979).

Adicionando modelos de correções suficientes ao modelo padrão tem-se uma família empírica de curvas de luz para diferentes filtros, as quais demonstram relações interessantes entre

a forma da curva de luz e a luminosidade antes do máximo e mais do que quinze dias após o máximo. Em (HAMUY et al., 1996) observamos que SNIa intrinsecamente fracas aumentam rapidamente a luminosidade em V em comparação com SNIa intrinsecamente mais luminosas. A curva de luz na banda B tem um comportamento semelhante ao da banda V , SNIa fracas tem subida e descida mais rápida em B e V do que SNIa mais luminosas. Na banda R SNIa mais luminosas tem um "ombro" em ~ 25 dias após máximo de B . Para SNIa fracas, este "ombro" é menos pronunciado e desaparece completamente para os objetos mais sub-luminosos. Na banda I , SNIa mais brilhantes tem dois máximos. O primeiro ocorre muito cedo, ~ 5 dias antes do máximo na banda B . O segundo é mais amplo e ocorre ~ 30 dias após o máximo em B . Para a maioria das SNIa subluminosas, os dois máximos se fundem em um máximo que é largo e ocorre ~ 5 dias após a máxima luminosidade em B (HAMUY et al., 1996).

As reconstituições de curvas de luz fotométricas, proporcionam um poderoso método para medir distâncias com correções de extinção para todas as fases de observação. SNIa podem parecer fracas porque elas estão distantes, obscurecida por poeiras, gases, etc, ou porque realmente são intrinsecamente mais fracas. Podemos distinguir entre estas possibilidades, utilizando as formas das curvas de luz e cor (que determinam a intrínseca luminosidade e cor) e medir os deslocamentos observados (que determinam a medida da distância com a correção de extinção). Dessa forma, dada uma curva de luz fotométrica, buscamos o melhor conjunto de curvas para um valor fixo de Δ que minimiza o χ^2 entre o modelo e os dados, fornecendo as melhores estimativas de μ e A_V .

Para a forma matricial descrita anteriormente, podemos utilizar a seguinte notação: $\mathbf{x}$ é a coluna com as medidas das magnitudes aparentes, $\mathbf{r}$ a coluna com as medidas padrão do modelo, $\mathbf{L}$ é a matriz com três colunas de correções e deslocamentos do modelo e $\mathbf{k}$ é a coluna com os três parâmetros livres (μ , A_V e Δ). Assim podemos reescrever a equação 2.10 para χ^2 ,

$$\chi^2 = (\mathbf{x} - \mathbf{r} - \mathbf{Lk})^T \mathbf{C}^{-1} (\mathbf{x} - \mathbf{r} - \mathbf{Lk}) \quad (2.14)$$

A minimização de χ^2 com respeito a coluna de parâmetros livres ($\mathbf{k}$) fornece,

$$q_{best} = [\mu \quad A_V \quad \Delta]_{best}^T = (\mathbf{L}^T \mathbf{C}^{-1} \mathbf{L})^{-1} \mathbf{L}^T \mathbf{C}^{-1} [\mathbf{x} - \mathbf{r}] \quad (2.15)$$

onde o erro padrão para os parâmetros em $\mathbf{k}$ são dados por $(\mathbf{L}^T \mathbf{C}^{-1} \mathbf{L})^{-1}$. Essa equação mede simultaneamente a distância e a extinção usando todas as observações de curvas de luz que estiverem disponíveis. Quanto ao erro na medida do módulo da distância (μ) extraídos dos dados da banda de passagem V , ele pode ser obtido pelo produto do primeiro termo da matriz de covariância do modelo $((\mathbf{L}^T \mathbf{C}^{-1} \mathbf{L})^{-1})$ com a minimização de χ_V^2 para

a forma da curva de luz V :

$$\sigma_\mu^2 = (L^T C^{-1} L)_{(1,1)}^{-1} \chi_V^2(V) \quad (2.16)$$

Da mesma forma, o erro de extinção derivado para $B - V, V - R$ e $V - I$ pode, por exemplo, ser descrito como o produto entre o elemento $(2, 2)$ da matriz de covariância $((L^T C^{-1} L)^{-1})$ e a minimização de χ^2 para a forma da "curva de cor",

$$\sigma_{A_V}^2 = (L^T C^{-1} L)_{(2,2)}^{-1} \chi_V^2(B - V, V - R, V - I) \quad (2.17)$$

A distância com a devida correção de extinção é dada por $\mu - A_V$, logo a variância é igual a soma das duas equações anteriores menos o dobro da estimativa de covariância para μ e A_V :

$$\sigma_{\mu - A_V}^2 = \sigma_\mu^2 + \sigma_{A_V}^2 - 2(L^T C^{-1} L)_{(1,2)}^{-1} \sqrt{\chi_V^2 \chi_V^2(B - V, V - R, V - I)} \quad (2.18)$$

Se medirmos A_V devido ao efeito de extinção por "poeira", podemos afirmar a priori, que a magnitude aparente deve ser corrigida por um A_V positivo (a luminosidade deve cair). Assim podemos incorporar esse prior em nossa análise (apesar que em princípio, se a SN for intrinsecamente mais fraca, o melhor ajuste para A_v pode ser negativo). Supondo que queremos estimar um A_V com média $\hat{A}_V$ e variância σ_{A_V} temos informações sobre a distribuição $p(A_V)$ e utilizando o teorema de Bayes podemos escrever,

$$p(A_V | \hat{A}_V, \sigma_{A_V}) = \frac{p(\hat{A}_V | A_V, \sigma_{A_V}) p(A_V)}{p(\hat{A}_V)} = \frac{e^{-\frac{(A_V - \hat{A}_V)^2}{2\sigma_{A_V}^2}} p(A_V)}{\int_{-\infty}^{\infty} p(A_V) e^{-\frac{(A_V - \hat{A}_V)^2}{2\sigma_{A_V}^2}} dA_V} \quad (2.19)$$

essa simplificada abordagem Bayesiana, fornece a verdadeira distribuição de probabilidade para A_V que é utilizada para se obter as melhores estimativas de A_V e do erro.

Essas são as idéias básicas do algoritmo para medir distâncias com correções de extinção através de curvas de luz multicolor mencionadas em (RIESS; PRESS; KIRSHNER, 1996). Porém existem algumas complicações na prática para se medir as distâncias para supernovas através da forma da curva de luz, dentre elas podemos citar a correção-K (OKE; SANDAGE, 1968), efeitos de extinção e redshift para a forma da curva de luz de SNIa distantes (LEIBUNDGUT et al., 1996). O algoritmo atualizado *MLCS2k2* inclui a correção-K e segue as leis para a extinção e procedimentos descritos inicialmente em (CARDELLI; CLAYTON; MATHIS, 1989), com eficiência comprovada em (NUGENT; KIM; PERLMUTTER, 2002). Para mais detalhes sobre a atualização do modelo (*MLCS2k2*), sugerimos (JHA; RIESS; KIRSHNER, 2007).

2.2 SALT₂

Esse método de ajuste fotométrico, utiliza um conjunto de curvas de luz e dados espectroscópicos para treinar o modelo, aumentando a resolução no espaço de comprimentos de onda, sendo totalmente independente do trabalho de (NUGENT; KIM; PERLMUTTER, 2002). Ele modela o sinal da supernova no espaço espectroscópico (em termos dos comprimentos de onda) garantindo que as correções-K serão tratadas de forma consistente, uma vez que existe um único modelo que aborda curvas de luz e espectros. Também permite com isso, uma coerente propagação de erros no encaixe das curvas de luz às estimativas de distâncias. O modelo pode variar em função da fase e comprimento de onda, com um pequeno número de parâmetros intrínsecos a priori desconhecidos, e uma lei de variação de cores que é ajustada durante o processo de treinamento.

O fluxo luminoso normalizado de cada supernova é uma função livre do modelo. Portanto, não precisamos saber a distância para realizar o treinamento. Isso nos permite usar as supernovas próximas que não estão no fluxo Hubble e também as supernovas em altos redshifts, sem qualquer prior cosmológico. Supernovas em altos redshifts podem ser utilizadas para modelar a emissão de UV de cada quadro de descanso (rest frame), que é fundamental para melhorar as estimativas de desvios nos dados de supernovas encontradas em redshifts maiores que 0,8 (GUY et al., 2005), (GUY, 2007). O objetivo principal deste método é fornecer a melhor sequência espectral média e o principal componente responsável pela variabilidade de SNIa, de modo que o modelo pode explicar possíveis variações nos espectros de supernovas Ia para qualquer fase dada.

Pretendemos então modelar a evolução média da distribuição espectral de energia (*SED*) de SNIa e sua variação com poucos componentes dominantes, incluindo uma independência do tempo na variação de cor, seja intrínseca ou devido à extinção por poeira na galáxia hospedeira (ou ambos). A forma funcional para o fluxo é dada por:

$$F(S, N, p, \lambda) = x_0 \times [M_0(p, \lambda) + x_1 M_1(p, \lambda) + \dots] \times \exp[cCL(\lambda)] \quad (2.20)$$

onde p é o tempo do quadro de observação desde a data da máxima luminosidade na banda b (p é a fase) e λ é o comprimento de onda no quadro de descanso observado (rest frame). M_0 é a sequência espectral média enquanto $M_k(p, \lambda)$ para $k > 0$ são componentes adicionais que descrevem a variabilidade nas SNIa. $CL(\lambda)$ representa a média da lei de correção de cor e c é a profundidade óptica que é expressa usando um deslocamento de cor em relação a média na data de luminosidade máxima na banda B, $c = (B - V)_{MAX} - \langle B - V \rangle$ (em (GUY et al., 2008) $\langle B - V \rangle = -0,057$ foi a cor de referência escolhida para SNIa, com base nas magnitudes de bandas fotométricas da estrela Vega 1.6.1). Esta parametrização modela a parte da variação de cor que é independente da fase, enquanto a variação de cor restante, com a fase, é explicada por componentes lineares. x_0 é a normalização

da distribuição espectral de energia e x_k para $k > 0$, são os parâmetros intrínsecos da supernova (como por exemplo, um fator de estiramento devido ao redshift). Em suma, M_k e CL são propriedades do modelo global, x_k e c são parâmetros de uma específica supernova dada e, portanto, diferem para cada supernova. Exceto pelo termo exponencial de cor, a equação anterior é equivalente a uma decomposição em componentes principais (veja também, 3.6.2). No entanto, na análise de componentes principais normalmente exigimos um conjunto homogêneo e denso de observações para cada supernova (o que está em construção atualmente), no entanto, o SALT2 foi desenvolvido para ser capaz de tratar problemas com dados faltantes. Normalmente temos para cada supernova um conjunto limitado de pontos de curvas de luz observados em diferentes filtros e, para algumas poucas supernovas, um ou vários espectros em diferentes fases. No entanto, se usarmos um conjunto de dados de supernovas grande o suficiente, poderemos modelar o espaço de fase corretamente (comprimentos de onda no intervalo de tempo do quadro de descanso) e vários componente principais poderão ser extraídos. Para relacionar o modelo é definida uma função de base dependente da fase e do comprimento de onda $[f_i(p, \lambda)]$. A escolha da real função de base, não é importante nas regiões do espaço de fase que são densamente cobertas por dados, desde que forneça uma resolução suficiente para seguir a variabilidade observada na distribuição espectral de energia, em função da fase e do comprimento de onda. A escolha da base modificará o modelo em algumas regiões do espaço de fase, como por exemplo, para espectros capturados muito antes do máximo ($p < -15$ dias), neste caso, essas regiões são identificadas após o treinamento usando uma técnica chamada de "canivete" como veremos mais a frente.

Nessa estrutura o modelo é uma combinação linear das funções de base e pode ser descrito por um vetor $\mathbf{M}$. Cada medida em uma dada fase e comprimento de onda $m(p_m, \lambda_m)$ é então comparada ao modelo com um vetor $\mathbf{H}_m$ (com valores $H_{m,i} = f_i(p_m, \lambda_m)$), de modo que o valor esperado para o modelo em (p_m, λ_m) seja o produto escalar $\mathbf{H}_m^T \mathbf{M}$.

O processo de convergência consiste em minimizar χ^2 permitindo a comparação do conjunto de dados completo com o modelo da equação dada anteriormente. Para cada SN, os parâmetros são a normalização e coordenadas ao longo dos componentes principais (x_k), uma cor e parâmetros de recalibração para espectros, se houver. Os componentes (M_k) e os parâmetros da lei de cores $CL(\lambda)$ também devem ser estimados. Este procedimento requer uma primeira adivinhação para os componentes do modelo (M_k), para uma primeira estimativa de normalização, recalibração do espectro e cor. Em (GUY, 2007) foi utilizada a sequência SED do modelo SALT para uma SN com estiramento = 1 para M_0 e a diferença da sequência SED de um SN com estiramento $x_0 = 1.1$ e um valor prévio para M_1 (ou seja, uma versão linearizada do modelo SALT), encontrando mais de 3000 parâmetros adequados, com obviamente não-linearidades, de modo que foi utilizado o procedimento Gauss-Newton e consistiu em:

1. Aproximar localmente χ^2 por uma função quadrática dos parâmetros.
2. Resolver um sistema linear grande para obter um incremento nos parâmetros (δP_i).
3. Incrementar os parâmetros e iterar até o decréscimo de χ^2 em relação à iteração anterior se torne insignificante.

Primeiro, a média do modelo é estimada juntamente com a lei de cores, os coeficientes de calibração para os espectros e os parâmetros do SN $((x_k), c)$. Quando o sistema converge, adiciona-se outro componente, e todos os parâmetros são ajustados novamente (componentes, leis de cores, parâmetros de SNs). O algoritmo de convergência é insensível ao conjunto de componentes de entrada.

Pode haver alguma degenerescência em parte do espaço de fase para o conjunto de dados, como por exemplo, se uma região de comprimento de onda $\times$ (fase) é coberta por fotometria e não por espectroscopia, não temos dados suficientes para restringir as combinações de parâmetros que modelam características espectrais, enquanto ainda podemos modelar uma medida fotométrica, já que o sinal está integrado em uma ou mais bandas espectrais. Adicionar um termo de regularização em χ^2 ajuda a resolver esse problema. Em (GUY, 2007) foi escolhido minimizar as segundas derivadas em relação à fase e ao comprimento de onda (essa abordagem é efetiva somente quando não existem dados suficientes). Nesse caso o termo de regularização pode ser definido por:

$$\chi_{REGUL}^2 = n \times \sum \mathbf{M}_k^T \mathbf{D}^T \mathbf{D} \mathbf{M}_k \quad (2.21)$$

onde $\mathbf{M}_k$ é o vetor que descreve o componente k , $\mathbf{D}$ é a matriz de derivadas segundas e n uma normalização que controla o peso desta regularização em relação aos dados. Para esse propósito, até então foram utilizados conjuntos de dados simulados que ajudam a definir a resolução do modelo. Cada SN da amostra de treinamento é ajustada usando o modelo *SALT2*, logo as curvas de luz falsas e os espectros são calculados substituindo cada medida verdadeira de SN pelo melhor valor de ajuste do modelo. Em (GUY, 2007) o procedimento de treinamento apresentou resultados ligeiramente tendenciosos devido ao termo de regularização em χ^2 na região do ultravioleta. O peso do termo de regularização (normalização n) foi escolhido para que o viés nas correções-K seja menor que $0,005mag$ para todos os comprimentos de onda, o que é significativamente inferior às incertezas estatísticas.

A escolha da resolução do modelo é imposta pelo conjunto de dados que utilizamos. Em (GUY, 2007) foram utilizados 10×120 parâmetros para M_0 (10 ao longo do eixo de tempo e 120 parâmetros de comprimento de onda), em uma faixa de fase de $[-20, +50]$ dias e uma faixa espectral de $[2000, 9200]\text{\AA}$. Isso dá uma resolução espectral da ordem de $60\text{\AA}$ que é suficiente para a modelagem de SN com amplas linhas espectrais devido à velocidade de ejeção da matéria durante a explosão. Para M_1 , em (GUY, 2007) foi

escolhido usar uma resolução mais baixa (10×60 parâmetros). O eixo do tempo foi remapeado para que a resolução seja duas vezes melhor nas observações entre -20 e $+50$ dias (aproximadamente 4, 5 e 9 dias, respectivamente). Utilizando esses parâmetros no treinamento foi possível identificar que a resolução mais baixa, apresenta uma dispersão em cores de apenas $0,01\text{mag}$, que tem uma contribuição desprezível nas estimativas de distâncias em comparação com a dispersão intrínseca de luminosidades de SNIa. Em (GUY et al., 2008) foi escolhido parametrizar as curvas de luz utilizando uma única cor (as possibilidades de implementações com os métodos de ajustes de curvas de luz é em grande parte arbitrária).

2.2.1 Correção-k

A qualidade das correções-k pode ser obtida diretamente pela comparação da magnitude do pico observado de uma curva de luz em um determinado filtro com sua previsão pelo modelo usando um ajuste das curvas de luz. A figura 2 apresenta as diferenças entre as magnitudes observadas e previstas como uma função do comprimento de onda efetivo do quadro de descanso e da resposta do instrumento utilizado. Uma abordagem mais elaborada consiste na modelagem de erros de correção-k como uma função paramétrica de λ em que os valores para os comprimentos de onda correspondentes às bandas B e V omitidos do quadro de descanso também possam ser estimados (erros nas magnitudes máximas de B e V são inserido na normalização e avaliação das bandas espectrais). Para cada SN com curvas de luz suficientes, esses parâmetros adicionais de correção-K podem ser estimados e seu desvio padrão usado para derivar um modelo de erros de correção K. Esse modelo é representado pela linha sólida da figura a seguir e é dado pela seguinte fórmula:

$$\sigma_K(\lambda) = 0.022 \left(\frac{\lambda - \lambda_b}{\lambda_u - \lambda_B} \right)^3 = 0.018 \left(\frac{\lambda - \lambda_V}{\lambda_R - \lambda_V} \right) \quad (2.22)$$

uma vez que a estimativa de σ_K se baseia no ajuste da normalização das curvas de luz, ele mede a dispersão de cores média no intervalo de fase definido pelo conjunto de dados. Pela equação anterior e o gráfico a seguir, podemos afirmar que os erros de correção-K são grandes na faixa do UV. Além disso, esses erros devem ser adicionados aos erros estatísticos nas normalizações, mas não são responsáveis pela dispersão total observada, como pode ser visto em (GUY, 2007).

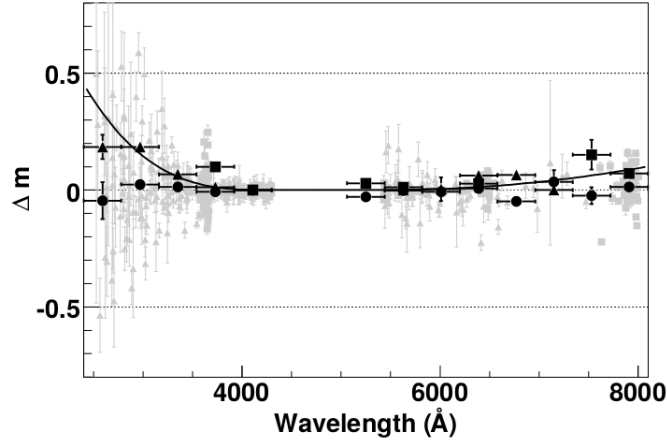


Figura 2 – Diferença entre a magnitude do pico observado em cada banda para as SN de (GUY, 2007), a previsão do modelo como função do comprimento de onda efetivo do filtro é utilizado (triângulos cinzas: SNLS SNe, quadrados cinzas: SNe próxima). Uma vez que as incertezas nas magnitudes B e V no máximo entram na normalização e avaliação de cores do modelo, as incertezas de correção K são definidas como zero para os comprimentos de onda da banda B e V.

2.3 Outros modelos de ajuste

2.3.1 Modelo1 - (NEWLING et al., 2010)

Esse modelo foi originalmente desenvolvido para o *SNPhotCC* (KESSLER et al., 2010b) e consiste basicamente em padronizar os dados brutos das curvas de luz em cada banda de cor, ajustando por mínimos quadrados ponderados, uma função parametrizada às curvas de luz e "cor". Os parâmetro do modelo são $(A, \phi, \psi, k, \sigma)$ e o fluxo em cada banda é modelado por:

$$F(t) = A \left(\frac{t - \phi}{\sigma} \right)^k \exp \left(-\frac{t - \phi}{\sigma} \right) k^{-k} e^k + \Psi(t) \quad (2.23)$$

Os parâmetros de melhor ajuste em cada banda fotométrica, seguem as interpretações: $A + \phi$ é o pico (fluxo máximo), ψ é o tempo inicial da explosão, k determina o aumento relativo e os tempos de decaimento, σ é um termo de estiramento temporal e $\Psi(t)$ é a função de "cauda", de forma que $F(t) \rightarrow \psi$ quando $t \rightarrow \infty$. A forma exata de Ψ é:

$$\Psi(t) = \begin{cases} 0 & -\infty < t < \phi \\ \text{cubic spline} & \phi < t < \tau \\ \psi & \tau < t < \infty \end{cases} \quad (2.24)$$

onde τ é a duração do pico de fluxo e a *cubic spline* é determinada a ter derivadas zeros em $t = \phi$ e $t = \tau$. Assim cada banda fotométrica tem 5 parâmetro livres $(A, \phi, \psi, k, \sigma)$ o que implica, por exemplo no caso da pesquisa LSST (*ugrizY*), em ajustes de pelo

menos 30 parâmetros para cada supernova. Esses ajustes, geralmente são feitos utilizando algum package *MultiNest* (FEROZ; HOBSON, 2008; FEROZ; HOBSON; BRIDGES, 2009) que é construído com um conjunto de *nested sampling* (SIVIA; SKILLING, 2006; SKILLING, 2004a). Em (NEWLING et al., 2010) são aplicados métodos de classificação por densidade de estimativa de kernel e *boosting* a este modelo, analisando a influência da representatividade e não representatividade dos dados de treinamento.

2.3.2 Modelo2 -(KARPENKA; FEROZ; HOBSON, 2012)

Inicialmente tratado em (KARPENKA; FEROZ; HOBSON, 2012) seu ajuste, assim como o do *modelo1*, segue uma parametrizada forma funcional que é baseada em (BAZIN et al., 2009) e (KESSLER et al., 2010a), mas utilizando uma melhor precisão na parametrização:

$$f(t) = A[1 + B(t - t_1)^2] \frac{e^{-(t-t_0)/T_{fall}}}{1 + e^{-(t-t_0)/T_{rise}}} \quad (2.25)$$

onde para cada SN, $t = 0$ corresponde ao tempo da primeira observação da curva de luz na banda r . Os parâmetros $A_0, t_0, T_{rise}, T_{fall}, B$ e t_1 são ajustados separadamente para cada banda de passagem. Embora esta forma não tenha motivação física particular, é suficientemente geral para se ajustar à forma de praticamente todos os tipos de curvas de luz para supernovas, incluindo aquelas que apresentam picos duplos (diferentemente da função de ajuste usada em (NEWLING et al., 2010) que não lida bem com esses tipos de curvas de luz). Ao contrário da função em (BAZIN et al., 2009) o parâmetro aditivo de compensação é omitido, assumindo a simplificação de que o fluxo é zero fora do evento SN. Mas este parâmetro pode ser incluído, por exemplo, se estivermos tratando dados reais brutos. Além disso, a parametrização do *modelo2* é feita para um pré-fator quadrático em $f(t)$. Para cada SN, em (KARPENKA; FEROZ; HOBSON, 2012), foram realizados ajustes filtro por filtro separadamente, assumindo a seguinte função de verossimilhança (likelihood) gaussiana:

$$\mathcal{L}(\Theta) = \exp \left[-\frac{1}{2} \chi^2(\Theta) \right] \quad (2.26)$$

onde o parâmetro Θ é um vetor ($\Theta = \{A, B, t_1, t_0, T_{rise}, T_{fall}\}$) e:

$$\chi^2(\Theta) = \sum_{k=1}^n \frac{[F_k - f(t_k; \Theta)]^2}{\sigma_k^2} \quad (2.27)$$

em que n é a quantidade de observação do fluxo para cada SN e filtro em consideração e F_k é o verdadeiro fluxo contido no conjunto de dados. Assim como no *modelo1*, o ajuste dos parâmetros é realizado através de algum package *MultiNest* (FEROZ; HOBSON, 2008; FEROZ; HOBSON; BRIDGES, 2009). Em (KARPENKA; FEROZ; HOBSON, 2012) a

classificação é feita por uma rede neural artificial multicamadas *perceptron*, treinada e validada sobre os parâmetros do *modelo2* extraídos de curvas de luz para supernovas do *SNPhotCC* (KESSLER et al., 2010b).

2.4 SNANA

O SNANA foi inicialmente desenvolvido para o SDSS-II Supernova Survey (KESSLER et al., 2009), com o objetivo principal de usar SNs Ia como indicadores de distância para a determinação de parâmetros cosmológicos, mas também pode ser aplicado no estudo de eficiências em análises de taxas de SNs, estimativas de contaminações de SNs nIa e otimizações de futuras pesquisas. Esse software é capaz de realizar simulações de dados observacionais através do ajustes da curvas de luz para todos os tipos conhecidos de supernovas. As simulações não utilizam pixels ou imagens, apesar de fazerem uso de informações geradas por imagens, tais como a escala e o erro de fluxo. As curvas de luz são geradas através da utilização de diferentes modelos que incluem *MLCS2k2* (JHA; RIESS; KIRSHNER, 2007), *SALT2*, (GUY et al., 2008), *SNooPy* (BURNS et al.,), *stretch* (GOLDHABER et al.,), *two-stretch* (HAYDEN; GARNAVICH; SURVEY,) e *core-collapse* (KESSLER et al., 2009).

A simulação com SNANA foi projetada para ser rápida gerando de $\sim 10^2 - 10^3$ curvas de luz por segundo, fornecendo acurácia e descrição realista para curvas de luz de supernovas. O SNANA permite a configuração das variações de ruídos atmosféricos, tratamentos de efeitos de extinção, correção-K e cadência de observação para pesquisas de telescópios arbitrários, sendo por isso, utilizado em várias colaborações, o que vem tornando-o mais robusto e fácil de usar, já que foi projetado para passar por melhorias continuadas da comunidade científica a medida que aprendemos mais sobre o universo observável. A confiabilidade das simulações é baseada na acurácia das "condições observacionais" contidas em um arquivo "*SIMLIB*" que descreve os ruídos atmosféricos, o ponto zero de magnitude de cada filtro e a cadência de observação para a pesquisa em interesse, para mais informações, veja também 5.1.

O SNANA executa uma pilha de códigos binários subjacentes baseados em uma mistura das linguagens *fortran* e *C*. Enquanto a simulação é um programa autônomo sem interface para usuários, os programas de análise permitem a interação do usuário de várias maneiras utilizando as opções privadas descritas em (KESSLER, 2018). Essa arquitetura permite que os usuários se concentrem em escrever novos recursos de análises, sem preocupações com a sobrecarga de leitura dos arquivos de entrada. Em geral, existem três maneiras de utilizar as opções privadas: escrever um pacote de análise ou ajuste inteiro (por exemplo *snlc_fit* ou *psnid*, usar códigos já existentes modificando algoritmos subjacentes ou adicionar códigos privados nas rotinas de interface para usuário (USRINI, USRANA, USREND)

como por exemplo, para calcular um novo parâmetro ou escrever informações em um formato específico ([KESSLER, 2018](#)).

Nós seguimos o padrão SNANA para simulação da pesquisa LSST, o mesmo que foi adotado em ([DAI et al., 2018](#)), para gerar cerca de 144.246 supernovas entre Ia e nIa. Para mais detalhes sobre configurações do código SNANA, veja [5.1](#).

2.5 SNcosmo

O SNcosmo é uma biblioteca em python para análise de cosmologia de supernovas em alto nível. O objetivo é tornar essa análise tão flexível e clara quanto possível. Ele é construído utilizando os pacotes NumPy, SciPy e AstroPy.

O SNcosmo também conta com modelos de supernovas incorporados, como SALT2, MLCS2k2, Hsiao, Nugent, PSNID, SNANA e Whalen, bem como uma variedade de bandas de passagem e sistemas de magnitude incorporados, o que lhe confere a capacidade de sintetizar espectros de supernovas e fotometria através de modelos. Além disso, rápidas implementações de várias leis de extinção podem ser utilizadas para construir um modelo que inclua "poeira". Novos modelos, bandas e sistemas de magnitude podem ser definidos, usando uma interface orientada a objetos.

Em nosso estudo, ele foi empregado apenas na simulação dos resultados obtidos com o SNANA e na leitura pelo código utilizado em ([LOCHNER et al., 2016](#)). Para maiores detalhes da utilização do SNcosmo e exemplos, veja ([BARBARY; CONTRIBUTORS, 2018](#)).

3 Aprendizagem de máquina

Aprendizagem de máquina (em inglês, *Machine Learning*) é o campo de estudos que fornece a computadores capacidade de aprenderem com os dados. Além disso, a aplicação de ML em extensos conjuntos de dados (*big data*) pode auxiliar na descoberta de padrões que não são evidentes, ajudando humanos a aprenderem com os dados, o que também é chamado de *data mining*.

Uma típica tarefa de *machine learning* é a predição de valores numéricos alvo (*target*) para problemas envolvendo regressão ou classificação. Isso normalmente é feito, aplicando uma função custo (ou função de ajuste) durante as interações realizadas na etapa de treinamento. Na maioria das vezes, essa função é utilizada para medir a qualidade do modelo e atualizá-lo buscando a minimização do custo. Tanto o treinamento, quanto o teste, em métodos com ou sem supervisão, são fundamentalmente realizados através de repetitivas amostragens parciais do conjunto de dados total, isso é o que chamamos de validação cruzada (*cross-validation*).

Os sistemas baseados em aprendizagem de máquina podem ser classificados de acordo com o tipo de supervisionamento fornecido durante a etapa de treinamento. Existem vários tipos diferentes de sistemas baseados em aprendizagem de máquina, que são, até então, classificados em categorias seguindo os seguintes critérios:

- São ou não, treinados com supervisão humana;
- Podem ou não aprender incrementalmente em tempo real (*on-line* versus *batch-learning*) (Géron, 2017);
- Simplesmente realizam comparações de novos pontos de dados com pontos de dados conhecidos ou, em vez disso, detectam padrões nos dados de treinamento e constroem um modelo preditivo, como fazem os cientistas (baseados em medidas de similaridade entre dados (instâncias) × modelo de aprendizagem).

Esses critérios não são exclusivos e podem ser combinados na busca pelo melhor resultado. Em (Géron, 2017) são citadas quatro grandes categorias: *supervised learning*, *unsupervised learning*, *semisupervised learning*, e *reinforcement learning*.

Neste capítulo, introduzimos as teorias subjacente as técnicas para aprendizagem de máquina. Destacamos a importância dos métodos de validação cruzada para o treinamento e teste (validação) juntamente com as principais diferenças entre métodos supervisionados e não supervisionados. Citamos alguns dos métodos mais utilizados na redução de dimensionalidade de conjuntos de dados e extração de características, e encerramos com uma seção dedicada a avaliação das técnicas para aprendizagem de máquina.

3.1 Validação Cruzada (*Cross-validation* - CV)

Validação cruzada é uma técnica de verificação da qualidade de modelos, que avalia os resultados de uma análise estatística ou *aprendizagem de máquina* (para uma introdução a estatística, veja apêndice A). É principalmente aplicada, em cenários onde o objetivo é a previsão e se deseja estimar a precisão com que um modelo preditivo funcionará na prática. Em um problema de previsão, o modelo geralmente recebe um conjunto de dados conhecidos no qual o treinamento é executado (conjunto de dados de treinamento) e um conjunto de dados desconhecidos, contra o qual, o modelo é testado (chamado conjunto de dados de validação ou teste). O objetivo da CV é definir um conjunto de dados para "testar" o modelo na fase de treinamento (isto é, com o conjunto de validação).

Aqui, fornecemos uma breve revisão dos métodos mais comuns em CV.

3.1.1 Validações cruzada exaustiva

3.1.1.1 Método *leave-p-out*

O método *leave-p-out* (*LpO*) envolve o uso de p observações como o conjunto de validação e as observações restantes como o conjunto de treinamento. Isso se repete de todas as maneiras possíveis de reduzir a amostra original em um conjunto de validação com p observações em um conjunto de treinamento.

A validação cruzada *LpO* requer treinamento e validação do modelo C_p^n vezes, onde n é o número de observações total e C_p^n é o coeficiente binomial.

3.1.1.2 Método *leave-one-out*

A validação cruzada *leave-one-out* (*LOO*) é um caso particular de validação cruzada *leave-p-out* com $p = 1$. O processo é semelhante ao jackknife (MILLER, 1974), no entanto, com validação cruzada calculamos uma estatística nas amostras excluídas, enquanto que, com o jackknife, calculamos uma estatística somente a partir das amostragens mantidas.

A validação cruzada *LOO* não tem o mesmo problema de tempo de computação excessivo que a validação cruzada de *LpO* em geral, porque exige C_1^n treinamentos.

3.1.2 Validações cruzadas não exaustiva

3.1.2.1 Método holdout

O método holdout consiste em dividir o conjunto total de dados em dois subconjuntos mutuamente exclusivos, um para treinamento (estimação dos parâmetros) e outro para teste (validação). O conjunto de dados pode ser separado em quantidades iguais ou não.

Uma proporção muito comum é considerar $1/3$ dos dados para teste e os $2/3$ restantes para treinamento.

Após o particionamento, a estimação do modelo é realizada e, posteriormente, os dados de teste são aplicados e o erro de predição calculado. Esta abordagem é indicada quando temos disponíveis uma grande quantidade de dados para análise.

3.1.2.2 Método k-fold

Na validação cruzada k-fold, a amostra original é dividida aleatoriamente em k subamostras de tamanhos iguais. Das k , uma única subamostra é retida como os dados de validação para testar o modelo, e as restantes subamostras ($k - 1$) são usadas como dados de treinamento. O processo de validação cruzada é então repetido k vezes, com cada uma das k subamostras usadas apenas uma vez como dados de validação. Os k resultados das dobras podem então ser calculados para produzir uma única estimativa. A vantagem deste método em relação à subamostragem aleatória repetida, é que todas as observações são usadas para treinamento e validação, e cada observação é usada para validação exatamente uma vez. Validação cruzada de 10 vezes é comumente usada, mas em geral k permanece um parâmetro não fixo.

Por exemplo, definir $k = 2$ resulta em validação cruzada de duas vezes. Na validação cruzada de duas vezes, aleatoriamente misturamos o conjunto de dados em dois conjuntos d_0 e d_1 , para que ambos os conjuntos tenham o mesmo tamanho (isso geralmente é implementado ao embaralhar o array de dados e depois dividi-lo em dois). Em seguida, treinamos em d_0 e validamos em d_1 e depois fazemos o treinamento em d_1 e validamos em d_0 .

Quando $k = n$ (o número de observações), a validação cruzada k-fold é exatamente a validação cruzada leave-one-out.

3.2 Aprendizagem supervisionada

A aprendizagem supervisionada é aplicada na resolução de problemas onde conhecemos, pelo menos, parte dos resultados, isto significa, que podemos treinar a máquina várias vezes, até que ela ajuste seus parâmetros internos (para cada método adotado) e apresente valores satisfatórios (ou seja, dentro de uma tolerância pré-definida) quando comparados com os resultados esperados. Iniciamos com uma breve revisão dos métodos supervisionados mais comuns e ao final tratamos dos métodos mais "robustos".

3.2.1 *k*-Nearest Neighbors

É uma técnica de estimativa de densidade simples e frequentemente utilizada, baseada na distribuição dos vizinhos mais próximos. Nesse método originalmente proposto para um contexto astronômico em (DRESSLER, 1980), a densidade local implícita em uma posição arbitrária x é definida por:

$$\hat{f}_k(x) = \frac{k}{V_D(d_k)} \quad (3.1)$$

onde k é o número de vizinhos mais próximos d_k é a dimensão dos dados e $V_D(d_k)$ é o volume da hipersfera d_k – *dimensional*. O erro em $\hat{f}_k(x)$ é dado por $\sigma_f = k^{1/2}/V_D(d_k)$. Na prática k deve ser pelo menos 5 porque o estimador tem grande variância para k menor (IVEZIC et al., 2014).

Este método pode ser aplicado a redução de dimensionalidade, porém, sofre efeitos da "maldição de dimensionalidade", discutida em 3.6.

3.2.2 Métodos de regressão e modelos de ajustes

A regressão é um dos procedimentos gerais de ajuste e seleção de modelos. Ela pode ser definida como a relação que descreve o valor esperado de y dado x , onde y é uma variável dependente e x é um conjunto de variáveis independentes. O objetivo de se obter um modelo de "melhor ajuste" é diferente do interesse de se obter os valores dos parâmetros do modelo (por exemplo as propriedades da matéria escura ou de um planeta recém descoberto). Lembramos aqui, que um modelo de "melhor ajuste" pode gerar overfitting, generalizando, isto significa que ele se ajusta muito bem ao conjunto de dados observados, porém, se mostra ineficaz na previsão de novos resultados. O uso da palavra regressão para essa relação remonta a *Francis Galton* (1822 – 1911) (IVEZIC et al., 2014), que foi o primeiro a aplicar métodos estatísticos nos estudos da genealogia e psicometria.

Para um estudo detalhado dos métodos de regressão (como regressão linear, regressão polinomial, regressão logística, regressão não linear, regressão de kernel, regressão robusta) e exemplos de aplicações na linguagem *python* sugerimos (IVEZIC et al., 2014).

3.2.3 *Support Vector Machines*(SVMs)

Um vetor de suporte de máquina (SVM) é um poderoso e versátil modelo de aprendizagem de máquina, capaz de realizar lineares e não lineares classificações e regressões, detectando *outliers*. Ele é o mais popular modelo em aprendizagem de máquina e qualquer um, interessado em *Machine Learning* deve conhecê-lo. Os SVMs são particularmente bem adequados para a classificação de conjuntos de dados complexos pequenos ou médios (SHMILOVICI, 2009).

O SVM padrão recebe um conjunto de dados de entrada e prediz, para cada entrada dada, qual de duas possíveis classes a entrada faz parte, o que faz do SVM um classificador binário não probabilístico. Apesar de ser um classificador binário, vários SVMs podem ser combinados na solução de problemas com múltiplas classificações.

Dado um conjunto de treinamento, onde cada ponto é bem classificado como pertencendo a uma das duas classes, o algoritmo de treinamento do SVM constrói um modelo permitindo classificar novos dados reais ou simulados. Um modelo SVM é uma representação dos dados como pontos no espaço, mapeados de maneira que os pontos de cada classe sejam divididos por um espaço claro e tão amplo quanto possível. Assim novos dados reais e simulados podem então serem mapeados no mesmo espaço e classificados como pertencentes a uma ou outra classe, baseando-se em qual posição do espaço eles estão localizados. Em outras palavras, na classificação o SVM busca encontrar uma linha de separação (ou curva nos casos não lineares, o que é chamado também de truque de kernel (BOSER; GUYON; VAPNIK, 1992)), mais comumente chamada de hiperplano (ou hipersuperfície) entre dados de duas classes. Essa "linha" busca maximizar a distância entre os pontos mais próximos (chamada de margem) em relação a cada uma das classes. Para maiores detalhes sobre o algoritmo sugerimos (CORTES; VAPNIK, 1995).

3.2.4 *Decison Trees and Random Forest*

Assim como o SVM as técnicas de árvores de decisão (*Decision Trees*) são versáteis algoritmos de aprendizagem de máquina para a realização de ambas tarefas de regressão e classificação para complexos conjuntos de dados. Elas são construídas de cima para baixo (abordagem conhecida como divisão recursiva) e são capazes de realizar classificações múltiplas, partindo de métodos condicionais binários. As árvores de decisão são os blocos fundamentais do método de floresta aleatória (*Random Forest*) muito utilizado na classificação múltipla de conjuntos de dados complexos. Realmente os métodos baseados em árvores de decisão, são considerados uns dos melhores e mais populares métodos de aprendizagem supervisionada, sendo capazes de mapear relações não lineares entre os dados de forma eficiente com grande precisão, sendo aplicado em diversas tarefas de aprendizagem.

Basicamente uma árvore de decisão pode ser representada por um conjunto de regras e condições previamente conhecidas ou não. Na construção de uma árvore de decisão a estratégia é "dividir para conquistar", de forma que um problema complexo é decomposto em sub-problemas mais simples que recursivamente seguem a mesma estratégia. Assim as divisões estratégicas, afetam fortemente os resultados e vários algoritmos podem ser utilizados para a tomada de decisão de dividir um nó (ou folha) em dois ou mais sub-nós, baseando-se na divisão do espaço definido pelas características dos dados de treinamento, em sub-espacos distintos e não sobrepostos, associando no final, cada sub-espaco a uma classe. Esse processo de divisão é continuado até que um critério de parada definido pelo

usuário, seja alcançado.

Por padrão é utilizada a medida do *coeficiente de Gini* (calculado como uma razão entre as áreas no diagrama da *curva de Lorenz* (GASTWIRTH, 1971)), mas também podemos utilizar o conceito de *ganho de informação* derivado da *entropia de informação* que mede a homogeneidade dos dados de entrada com relação a sua classificação. Por exemplo, dado um conjunto de entrada D que pode ter c classes distintas, a entropia $H(D)$ será dada por:

$$H(D) = - \sum_{i=1}^c P_i \log(P_i) \quad (3.2)$$

onde P_i é a proporção dos dados que pertence a classe i . O ganho de informação para um parâmetro (y) do conjunto de dados D , fornece a redução da entropia esperada quando utilizamos y para particionar o conjunto de dados. Por exemplo, podemos considerar que o ganho de informação para um determinado atributo de um conjunto de dados D fornece uma medida da redução de entropia esperada quando utilizamos esse atributo para particionar os dados. Seja $P(D)$ o conjunto dos valores que D pode assumir, onde x é um elemento do subconjunto D_x formado pelos dados em que $D = x$, a entropia que se obtém ao particionar D em função desse atributo é dada por:

$$S(\text{atributo}) = \sum_{x=\text{atributo}} \frac{|D_x|}{|D|} H(D_x), \quad (3.3)$$

e o ganho de informação é dado por:

$$\text{ganho} = H(D) - S(\text{atributo}) \quad (3.4)$$

onde $H(D)$ é uma medida de não homogeneidade do conjunto D .

Esses critérios podem ser calculados, para cada nó (folha) da árvore de decisão, fornecendo estimativas de erros e realizando "podas". Assim a construção de uma árvore de decisão é guiada pelo objetivo de diminuir a entropia, ou seja, a aleatoriedade na classificação dos dados utilizando o mínimo de nós possíveis para a tarefa.

O método *Random Forest* basicamente é um conjunto de árvores de decisão onde os dados de entrada são selecionados de forma aleatória utilizando a técnica *bootstrapping* (BREIMAN, 1996). Em geral as características são divididas em subconjuntos aleatórios e o algoritmo constrói árvores menores a partir de cada subconjunto, depois do treinamento as sub-árvores são combinadas. Essa abordagem é menos suscetível a *outlier* e consequentemente ao *overfitting*.

Outra diferença entre os métodos *Decision Tree* e *Random Forest*, é que ao invés de utilizar o *índice de Gini* ou o *ganho de informação* para gerar regras e nós (folhas), o *Random*

Forest faz isso de modo aleatório, baseando em hiperparâmetros como o número máximo de árvores construídas, o número máximo de características que serão analisadas e a quantidade mínima de nós que deve haver em uma dada árvore.

Algumas técnicas podem otimizar o uso de árvores de decisão e consequentemente a aplicação do *Random Forest*, como a validação cruzada, o *boosting* e o *bootstrap*.

3.3 Boosting

O termo *boosting* (impulsionar) faz referência a qualquer método que consiste em aprender de forma iterativa os classificadores fracos de uma distribuição e combiná-los em um classificador forte ao final. Essa combinação normalmente é feita, adicionando os classificadores fracos ponderados de alguma forma, de acordo com sua precisão. Existem muito algoritmos de "impulsão" sendo os mais populares o *AdaBoost* (*Adaptive Boosting*) e *Gradient Boosting* (FREUND; SCHAPIRE, 1996b; FREUND; SCHAPIRE, 1996a).

3.3.1 AdaBoost:

O algoritmo de impulsionamento adaptativo primeiro treina um preditor (classificador) base (como por exemplo, uma árvore de decisão) e usa para fazer previsões sobre o conjunto de treinamento. O peso relativo de cada característica do treinamento classificada incorretamente é então aumentado e um segundo classificador é treinado usando os pesos atualizados, novamente faz-se previsões sobre o conjunto de treinamento, os pesos são atualizados e assim por diante. Nesse algoritmo cada característica inicializa com um peso $w = 1/n$, onde n é o tamanho do conjunto de treinamento. Desta forma, a taxa de erro do peso para o j -ésimo preditor (r_j) é dado por:

$$r_j = \frac{\sum_{i=1}^n w^{(i)} \mathbb{1}_{\hat{y}_j^{(i)} \neq y^{(i)}}}{\sum_{i=1}^n w^{(i)}} \quad (3.5)$$

onde $\hat{y}_j^{(i)}$ é a predição do j -ésimo preditor para a i -ésima instância. O incremento ou decremento ao peso α_j (peso dos preditores) pode então ser calculado:

$$\alpha_j = \eta \log \left(\frac{1 - r_j}{r_j} \right) \quad (3.6)$$

onde η é um hiperparâmetro relacionado a taxa de aprendizagem (O algoritmo original AdaBoost, faz uso de $\eta = 0,5$). Em seguida os pesos são atualizados para cada predição $\hat{y}_j^{(i)}$ através da relação:

Para $i = 1, 2, \dots, n$,

$$w^{(i)} \leftarrow \begin{cases} w^{(i)} & ; \text{para } \hat{y}_j^{(i)} = y^{(i)} \\ w^{(i)} \exp(\alpha_j) & ; \text{para } \hat{y}_j^{(i)} \neq y^{(i)} \end{cases} \quad (3.7)$$

Todos os pesos são normalizados (divididos por $\sum_{i=1}^n w^{(i)}$). Finalmente um novo preditor é treinado usando os pesos atualizados, e todo o processo é repetido (os pesos para cada predição são calculados, os pesos de cada instância i são atualizados, então outro preditor é treinado e assim por diante). O algoritmo pára quando um número desejado de preditores é atingido ou quando um preditor perfeito é encontrado.

O AdaBoost calcula as predições feitas por todos os preditores e as atualiza usando os pesos α_j . A classe prevista será aquela que receber a maioria das predições ponderadas.

$$\hat{y}(x) = \arg \max_k \sum_{\substack{j=1 \\ \hat{y}_j(x)=k}}^N \alpha_j, \quad (3.8)$$

onde N é o número de preditores (por exemplo, o número de árvores de decisão).

3.3.2 Gradient Boosting:

Assim como o *AdaBoost*, o *Gradient Boosting* funciona adicionando preditores sequencialmente a um conjunto, cada um corrigindo seu antecessor. No entanto, em vez de ajustar os pesos das instância em cada iteração, como o AdaBoost, esse método tenta ajustar o novo preditor aos erros residuais feitos pelo preditor anterior. Assim ele constrói o modelo por etapas, sendo otimizado através da minimização de uma *função de perda* particular.

Considere o problema em que temos um conjunto de treinamento $\{y_i, x_i\}_1^N$ onde conhecemos os valores dos dados de entrada ($\mathbf{x} = \{x_1, \dots, x_n\}$) e saída (y). O objetivo do *gradient boosting* é encontrar uma função $F^*(\mathbf{x})$ que mapeia $\mathbf{x}$ em y , de tal forma que o valor esperado de alguma função de perda específica $L(y, F(\mathbf{x}))$ seja calculada sobre a distribuição conjunta de todos os valores $(y, \mathbf{x})$:

$$F^*(\mathbf{x}) = \arg \min_{F(x)} E_{\mathbf{x}}[E_y(L(y, F(\mathbf{x}))) | \mathbf{x}] = \arg \min_{F(x)} E_{y, \mathbf{x}} L(y, F(\mathbf{x})) \quad (3.9)$$

A função de perda $L(y, F)$ frequentemente é empregada incluindo a estimativa do desvio mínimo quadrado $(y - F)^2$ (L_2) ou o desvio mínimo absoluto $|y - F|$ (L_1) para os casos de regressão (BUHLMANN; HOTHORN, 2007), enquanto para problemas de classificação, por exemplo binária, a função de perda é implementada com o log-likelihood de uma distribuição binomial negativa, $\log(1 + e^{-2yF})$ (FRIEDMAN; HASTIER; TIBSHIRANI, 2000), em outros casos também costuma-se usar o critério de perda exponencial e^{-yF} (FREUND;

SCHAPIRE, 1996b; SCHAPIRE; SINGER, 1998). O *gradiente boosting* aproxima $F^*(x)$ pelo somatório de uma expansão na forma:

$$F(x) = \sum_{m=1}^M \beta_m h(\mathbf{x}; a_m), \quad (3.10)$$

onde a função $h(\mathbf{x}, \mathbf{a})$ é conhecida como "*base learner*" (ou "weak learner") sendo usualmente escolhida como uma simples função de x com parâmetros $\mathbf{a} = \{a_1, a_2, \dots, a_M\}$. A expansão dos coeficientes $\{\beta_m\}_1^M$ e dos parâmetros $\{a_m\}_1^M$ são conjuntamente ajustados pelos dados de treinamento em etapas. Começando com uma hipótese inicial para $F_0(x)$, e depois para $m = 1, 2, \dots, M$. Tais expansões estão no centro de muitos métodos de aproximação de funções, como SVM, *decision tree*, redes neurais, wavelets, entre outros (FRIEDMAN, 1999).

A escolha da parametrização do modelo altera o problema de otimização da função, para uma otimização de parâmetros:

$$(\beta_m, a_m) = \arg \max_{\beta, a} \sum_{i=1}^N L(y_i, F_{m-1}(x_i) + \beta h(x_i; a)) \quad (3.11)$$

onde,

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \beta_m h(\mathbf{x}; a_m) \quad (3.12)$$

Em geral, a solução de 3.11 é difícil de ser obtida. Dada qualquer aproximação para $F_{m-1}(x)$, a função $\beta_m h(\mathbf{x}; a_m)$ pode ser vista como o melhor passo a ser tomado, em direção a estimativa baseada nos dados de $F^*(\mathbf{x})$ (3.9), sob a restrição de que a "direção" do passo em cada etapa ($h(\mathbf{x}; a_m)$), seja um dos membros do conjunto de funções $h(\mathbf{x}; \mathbf{a})$. Sob essa restrição, podemos então considerar que este é o passo de descida mais íngreme. Com isso, podemos definir o irrestrito gradiente negativo para nosso exemplo:

$$-g_m(x_i) = - \left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \right]_{F(\mathbf{x})=F_{m-1}(x)}, \quad (3.13)$$

que fornece a direção do passo rumo à descida mais íngreme $-\mathbf{g}_m = -\{g_m(x_i)\}_1^N$ em um espaço N -dimensional em $F_{m-1}(\mathbf{x})$. No entanto, esse gradiente é definido apenas nos pontos $\{x_i\}_1^N$ e não pode ser generalizado para outros valores de $\mathbf{x}$. Uma possibilidade nesse caso, é escolher o membro de $h(\mathbf{x}; \mathbf{a})$ que produz $\mathbf{h}_m = \{h(x_i; a_m)\}_1^N$ mais paralelo a $\mathbf{g}_m$. Assim, esse é o $h(\mathbf{x}; \mathbf{a})$ mais altamente correlacionado com $-g_m(\mathbf{x})$ sobre os dados da

distribuição, que pode ser obtido pela solução de

$$a_m = \arg \min_{\beta, a} \sum_{i=1}^N [-g_m(x_i) - \beta h(x_i; a)]^2 \quad (3.14)$$

Substituímos $-g_m(\mathbf{x})$ pela restrição $h(\mathbf{x}; a_m)$ na direção de descida mais íngreme. Podemos então, executar a *linha de pesquisa*:

$$\rho_m = \arg \min_{\rho} \sum_{i=1}^N L(y_i, F_{m-1}(x_i) + \rho h(x_i; a_m)) \quad (3.15)$$

e a atualização da aproximação,

$$F_m(\mathbf{x}) := F_{m-1}(\mathbf{x}) + \rho_m h(\mathbf{x}; a_m) \quad (3.16)$$

Basicamente, em vez de obtermos a solução pela restrição de 3.11, a restrição é aplicada à solução irrestrita, ajustando $h(\mathbf{x}; \mathbf{a})$ às "pseudo-respostas" $\{\tilde{y}_i\}_1^N$,

$$\tilde{y}_i = E_y \left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \middle| x_i \right]_{F(\mathbf{x})=F_{m-1}(\mathbf{x})}. \quad (3.17)$$

Isso permite substituir o difícil problema de minimização em 3.11 pela minimização da função de mínimos quadrados 3.14, seguida por uma otimização de parâmetro com base em 3.15. Desta forma, qualquer $h(\mathbf{x}; \mathbf{a})$ para o qual exista um algoritmo de mínimos quadrados viável, esta abordagem pode ser adotada.

Algoritmo *Gradient Boosting*:

1. $F_0(\mathbf{x}) = \arg \min_{\rho} L(y_i, \rho)$

2. Para $m = 1$ até M :

$$\tilde{y}_i = - \left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \right]_{F(\mathbf{x})=F_{m-1}(\mathbf{x})}, \text{ para } i = 1, \dots, N$$

$$a_m = \arg \min_{\beta, a} \sum_{i=1}^N [-g_m(x_i) - \beta h(x_i; a)]^2$$

Assim o algoritmo *gradient boosting* aproxima a solução da equação 3.11 por uma função de perda, $L(y, F(x))$ (*loss function*), diferenciável arbitrária, segundo um procedimento de duas etapas (FRIEDMAN, 1999).

3.4 Empacotamento (*Bagging*)

Uma das formas de se obter um conjunto de classificadores, é o uso de muitos algoritmos de treinamento diferentes. Outra forma, é utilizar um mesmo algoritmo de treinamento para todos os preditores e treinar com subconjuntos diferentes do conjunto de treinamento

original, aleatoriamente. Quando a amostragem é realizada com substituição do subconjunto de treinamento, o método é chamado por *bagging* (abreviação de *bootstrap aggregating*) (BREIMAN, 1996). Quando a amostragem é realizada sem substituição é chamada de *pasting* (BREIMAN, 1999).

Os preditores podem ser treinados em paralelo, através de diferentes núcleos CPU, ou até mesmo, servidores diferentes. Está é uma das razões para que *bagging* e *pasting* sejam métodos populares. Ambos, permitem que conjuntos de treinamento sejam amostrados varias vezes em diferentes preditores, mas somente o *bagging* permite que diferentes conjuntos de treinamento sejam amostrados várias vezes, para um mesmo preditor. Uma vez que todos os preditores são treinados, o conjunto pode fazer uma previsão para um novo conjunto de dados agregando as previsões de todos os preditores. Cada preditor, individualmente tem uma tendência (bias) mais alta do que se fosse treinado com o conjunto de treinamento original, mais a agregação reduz o bias e a variância. O resultado líquido do *bagging*, é que o conjunto tem um bias semelhante, mas uma variação menor do que um único preditor treinado com o conjunto de treinamento original (GÉRON, 2017).

Em geral, o *bagging* é padronizado para realizar a amostragem aleatória de, em média, 63% do conjunto de treinamento para cada preditor. Os 37% restante do conjunto de treinamento que não são amostrados são chamados de *out-of-bag (oob)*. Como um preditor nunca vê o conjunto *oob* durante a fase de treinamento, ele pode ser avaliado nesse conjunto sem a necessidade de utilização dos métodos de validação cruzada.

O *bootstrap* introduz um pouco mais de diversidade aos subconjuntos nos quais cada preditor é treinado, portanto, o *bagging* acaba com um bias ligeiramente maior do que o *pasting*, mas isso também significa que os preditores se tornam menos correlacionados o que reduz a variância do conjunto. Em geral, o *bagging* muitas vezes resulta em melhores modelo, o que explica porque é geralmente preferido, entretanto, os métodos de validação cruzada podem ser usados para avaliar *bagging* e *pasting* selecionando aquele que funciona melhor.

3.5 Aprendizagem não supervisionada

A aprendizagem de máquina não supervisionada permite a abordagem de problemas com pouca, ou nenhuma idéia de como os nossos resultados devem aparentar. Essas técnicas são poderosas ferramentas para derivar estruturas intrínsecas as propriedades dos dados que de outra forma, provavelmente não seríamos capazes de detectar. Dessa maneira, esse métodos formam agrupamentos, extraem características principais e realizam reduções de dimensionalidade, apenas comparando as informações contida nos dados de análise.

3.5.1 KDE

As técnicas para Estimativa de Densidade Kernel (KDE) são aplicadas na análise de dados que apresentam diferentes contribuições para a distribuição de probabilidade. Elas podem ser aplicadas tanto a problemas de regressão quanto classificação. Em geral nesses casos, uma solução é permitir que cada ponto tenha sua própria contribuição intrínseca com sobreposição de cada contribuição. Em essência, cada ponto é substituído por um "grão" (*kernel*) (que pode ser representado por um círculo, uma caixa, uma distribuição gaussiana ou algo mais complexo) com altura e largura previamente definida. O uso de um *kernel* muito estreito leva a uma distribuição "barulhenta", enquanto o uso de um *kernel* muito amplo, leva a uma suavização excessiva e consequentemente à perda de informação.

Uma das maiores complexidades na estimativa por kernel está na seleção da largura de banda. Por exemplo, para o caso mais simples de uma caixa, dado um conjunto de N medições $\{x_i\}$ o estimador de densidade de kernel (isto é, o estimador da PDF *função de densidade de probabilidade* subjacente) em uma posição arbitrária x é definido como:

$$\hat{f}_N(x) = \frac{1}{Nh^D} \sum_{i=1}^N \underbrace{K\left(\frac{d(x, x_i)}{h}\right)}_u \quad (3.18)$$

onde $K(u)$ é a função de kernel e h é conhecida como largura de banda. A densidade local é estimada como a média ponderada de todos os pontos com pesos $K(u)$ e sistematicamente diminuem com a distância $d(x, x_i)$. A função de kernel pode ser qualquer função positiva ($K(u) \geq 0$), normalizada ($\int K(u)du = 1$) e com variância ($\sigma_x^2 = \int u^2 K(u)du$) maior que zero.

Outros kernels comumente utilizados são o top-hat, o kernel gaussiano $K(u) = \left(\frac{1}{(2\pi)^{D/2}}e^{-u^2/2}\right)$, onde D é a dimensão do espaço dos parâmetros e $u = \frac{d(x, x_i)}{h}$, e o kernel exponencial $K(u) = \frac{1}{D!V_D(1)}e^{-u}$, onde $V_D(r)$ é o volume de uma hipersfera D-dimensional de raio r . Idealmente selecionaríamos um kernel que tivesse uma largura de banda menor possível, porém se h se torna muito pequeno, aumentamos a variância da estimativa de densidade e se h for muito grande, então suavizamos a distribuição, mas as custas de uma maior tendência nas estimativa de densidade. Existem muitos métodos na busca da largura de banda h ideal para um kernel. Dentre elas, métodos envolvendo validação cruzada e regularizações ($L1$ e $L2$), kernel Epanechnikov, a regra de Silverman's, etc.

Em geral o KDE estima a densidade de um conjunto de dados fixando um kernel para cada ponto dos dados. Uma outra alternativa é permitir que cada ponto tenha a sua contribuição intrínseca para a densidade, ajustando a largura de banda dos kernels localmente. Isto é chamado de modelo de misturas e pode ser entendido de duas formas: como um modelo de estimativa de densidade ou como algoritmo de agrupamento (classificação), onde a localização e h de cada componente reflete alguma propriedade subjacente dos dados.

3.5.2 *k*-Means

É um dos métodos mais simples para particionamento de dados em classes, como por exemplo, em um problema de classificação. Basicamente o método procura uma partição dos pontos em k subconjuntos disjuntos C_k contendo N_k pontos, de modo que a função para a soma de mínimos quadrados seja minimizada:

$$\min \sum_{k=1}^k \sum_{i \in C_k} \|x_i - \mu_k\|^2 \quad (3.19)$$

$$\mu_k = \frac{1}{N_k} \sum_{i \in C_k} x_i \quad (3.20)$$

onde μ_k é a média dos pontos na classe C_k e $C(x_i) = C_k$ denota que a classe do ponto x_i é C_k .

O procedimento k-Means inicia-se com a escolha do centróide, μ_k , para cada classe k . Em seguida cada ponto é atribuído a classe mais próxima ($C(x_i) = \operatorname{argmin}_k \|x_i - \mu_k\|$) e o centróide de cada classe é recalculado. Isso é repetido até que não existam mais pontos a serem atribuídos às classes. Na prática o método k-Means é realizado várias vezes em uma análise, iniciado com diferentes valores para os centróides de C_k e o resultado com menor error na soma dos quadrados é utilizado.

3.5.3 Hierarchical Cluster Analysis(HCA)

Esse é um método processual, isto é, um método onde não temos formalmente relação com alguma função de densidade subjacente. Ele é capaz de encontrar todos os agrupamentos em todas as escalas, não havendo a necessidade de se especificar inicialmente a quantidade k de grupos existentes.

O método consiste em dividir os dados em N classes, uma para cada ponto nos dados, em seguida, os pares de classes mais próximas são mescladas. As opções para definir a distância entre duas classes C_k e $C_{k'}$, incluem:

$$d_{\min}(C_k, C_{k'}) = \min_{(x \in C_k, x' \in C_{k'})} \|x - x'\| \quad (3.21)$$

$$d_{\max}(C_k, C_{k'}) = \max_{(x \in C_k, x' \in C_{k'})} \|x - x'\| \quad (3.22)$$

$$d_{\text{med}}(C_k, C_{k'}) = \frac{1}{N_k N_{k'}} \sum_{x \in C_k} \sum_{x' \in C_{k'}} \|x - x'\| \quad (3.23)$$

$$d_{\text{cen}}(C_k, C_{k'}) = \|\mu_k - \mu_{k'}\| \quad (3.24)$$

onde x e x' são os pontos nas classes C_k e $C_{k'}$, respectivamente, N_k e $N_{k'}$ são os números de pontos em cada classe e μ_k e $\mu_{k'}$, os respectivos centróides.

O uso de d_{min} (conhecido como árvore de extensão mínima - *minimum spanning tree*) resulta em classes com extensas cadeias de pontos, enquanto $d_{máx}$ tende a produzir um agrupamento hierárquico com classes compactas. Os outros dois exemplos de distâncias se comportam entre esses extremos.

3.5.4 Algoritmo EM e Extreme Deconvolution(XD)

O EM é um processo iterativo capaz de encontrar o valor que maximiza a função de verossimilhança (likelihood) de uma mistura de gaussianas. Ele possui basicamente duas etapas iterativas: **E-step** - calcula a probabilidade de cada ponto pertencer a uma das distribuições gaussianas fornecendo a estimativa para a likelihood, **M-step** - nesta etapa a maximização da likelihood é calculada.

Por exemplo, considerando o ponto de dados x de um conjunto de N medidas, x_i , como amostra de uma das M distribuições gaussianas diferentes com média μ_j e covariância σ_j , temos a *PDF*:

$$P(x) = \sum_j \alpha_j \mathcal{N}(x|\mu_j, \sigma_j) \quad (3.25)$$

onde α_j respeita a restrição das PDF's através da normalização $\sum_{j=1}^M \alpha_j = 1$ a função de densidade da distribuição normal é dada por:

$$\mathcal{N}(x|\mu_j, \sigma_j) = \frac{1}{\sqrt{(2\pi)^D \det(\sigma_j)}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right) \quad (3.26)$$

O logaritmo da função de verossimilhança (likelihood) para todo o conjunto de dados, é:

$$\ln L = \sum_{i=1}^N \ln \left[\sum_{j=1}^M \alpha_j \mathcal{N}(\mu_j, \sigma_j) \right] \quad (3.27)$$

O EM não é limitado apenas a misturas gaussianas ($\mathcal{N}(\mu_j, \sigma_j)$), podendo também ser usada uma *PDF* mais geral para cada componente da mistura $P_j(x_i|\theta)$, onde θ é o vetor de parâmetros da distribuição. Se a classe de cada valor do conjunto de dados não for conhecida, podemos determinar a probabilidade que cada ponto x_i tem de ser gerado por determinada classe j , utilizando a *regra de Bayes*:

$$P(j|x_i) = \frac{\alpha_j P_j(x_i|\theta)}{\sum_{j=1}^M \alpha_j P_j(x_i|\theta)} \quad (3.28)$$

Nesse caso, o $\ln L$ será dado por:

$$\ln L = \sum_{i=1}^N \ln \left[\sum_{j=1}^M \alpha_j P_j(x_i|\theta) \right] \quad (3.29)$$

Tomando a derivada parcial de $\ln L$ com relação ao parâmetro θ_j ,

$$\frac{\partial \ln L}{\partial \theta_j} = \sum_{i=1}^N \frac{\alpha_j}{\sum_{j=1}^M \alpha_j P_j(x_i|\theta)} \left[\frac{\partial P_j(x_i|\theta)}{\partial \theta_j} \right] \quad (3.30)$$

multiplicando e dividindo por $P_j(x_i|\theta)$, podemos reescrever:

$$\frac{\partial \ln L}{\partial \theta_j} = \sum_{i=1}^N \left[\frac{\alpha_j P_j(x_i|\theta)}{\sum_{j=1}^M \alpha_j P_j(x_i|\theta)} \right] \left[\frac{1}{P_j(x_i|\theta)} \frac{\partial P_j(x_i|\theta)}{\partial \theta_j} \right] \quad (3.31)$$

Nesse ultimo resultado, vemos que o primeiro termo corresponde a probabilidade de classe, $P(j|x_i)$ que por simplicidade chamaremos por $w_{ij} = P(j|x_i)$ e o segundo termo é a derivada parcial de $\ln[P_j(x_i|\theta)]$. Quando $P_j(x_i|\theta)$ é gaussiana temos uma particularmente simples restrição dos parâmetros do modelo, tomando o logaritmo da PDF antes da derivada:

$$\frac{\partial \ln L}{\partial \theta_j} = - \sum_{i=1}^N w_{ij} \frac{\partial}{\partial \theta_j} \left[\ln \sigma_j + \frac{(x_i - \mu)^2}{2\sigma_j^2} \right] \quad (3.32)$$

onde θ_j agora corresponde a μ_j e σ_j . Nessa configuração a derivada de $\ln L$ com respeito a μ_j e σ_j igual a zero, fornece os estimadores:

$$\mu_j = \frac{\sum_{i=1}^N w_{ij} x_i}{\sum_{i=1}^N w_{ij}}, \quad (3.33)$$

$$\sigma_j = \frac{\sum_{i=1}^N w_{ij} (x_i - \mu_j)^2}{\sum_{i=1}^N w_{ij}}, \quad (3.34)$$

e a partir da restrição de normalização,

$$\alpha_j = \frac{1}{N} \sum_{i=1}^N w_{ij}. \quad (3.35)$$

As expressões vista até então formam a base do processo iterativo do algoritmo EM para o caso de misturas gaussianas. O processo inicia com um palpite para w_{ij} (E-step) e os valores de α_j , μ_j e σ_j são estimados para mais próximo do máximo local (M-step). Em seguida w_{ij} é atualizado (E-step) e novamente os parâmetros (α, μ, σ) são estimados para maximizar $\ln L$. O processo se repete até que um máximo local provável seja encontrado.

O EM pode falhar devido certas dificuldades numéricas, especialmente quando os dados disponíveis são escassos, apresentam outliers ou são repetidos.

Na prática a maximização de $\ln L$ é feita com o uso do algoritmo Levenberg-Marquardt, que combina os algoritmo de gradiente descendente e Gauss-Newton (MORÉ, 1978). Outra possibilidade é o uso dos métodos Monte-Carlo via cadeia de Markov(MCMC) (GILKS; RICHARDSON; SPIEGELHALTER, 1995).

A generalização do algoritmo EM é conhecida como deconvolução extrema (XD), nesse caso, os dados apresentam erros de medição. Assim, pressupomos que as observações ruidosas x_i e os valores de interesse v_i estão relacionados seguindo a descrição:

$$\mathbf{x}_i = \mathbf{R}_i \mathbf{v}_i + \mathcal{E}_i, \quad (3.36)$$

onde $\mathbf{R}_i$ é a matriz de projeção que pode ser ou não invertível e assumimos que os ruídos $\mathcal{E}_i$ sejam extraídos de uma distribuição gaussiana com $\mu_i = 0$ e variância $\mathbf{S}_i$. Dadas as matrizes $\mathbf{R}_i$ e $\mathbf{S}_i$, o objetivo da XD é encontrar os parâmetros das gaussianas subjacentes μ_j , Σ_j e os pesos α_i , maximizando a likelihood para os dados observados. Desta forma, a generalização do algoritmo EM (XD) é realizada conforme os passos a seguir:

Expectativa (E-step):

$$Q_{ij} \leftarrow \frac{\alpha_j \mathcal{N}(\mathbf{w}_i | \mathbf{R}_i | \mu_j, \mathbf{T}_{ij})}{\sum_j \alpha_k \mathcal{N}(\mathbf{w}_i | \mathbf{R}_i | \mu_k, \mathbf{T}_{ij})}, \quad (3.37)$$

$$\mathbf{b}_{ij} \leftarrow \mu_j + \Sigma_j \mathbf{R}_i^T \mathbf{T}_{ij}^{-1} (\mathbf{w}_i - \mathbf{R}_i \mu_j), \quad (3.38)$$

$$\mathbf{B}_{ij} \leftarrow \Sigma_j - \Sigma_j \mathbf{R}_i^T \mathbf{T}_{ij}^{-1} \mathbf{R}_i \Sigma_j \quad (3.39)$$

onde, $\mathbf{T}_{ij} = \mathbf{R}_i \Sigma_j \mathbf{R}_i^T + \mathbf{S}_i$

Maximização (M-step):

$$\alpha_i \leftarrow \frac{1}{N \sum_i Q_{ij}} \quad (3.40)$$

$$\mu_j = \frac{1}{Q_j} \sum_i Q_{ij} \mathbf{b}_{ij}, \quad (3.41)$$

$$\Sigma_j = \frac{1}{Q_j} \sum_i Q_{ij} [(\mu_j - \mathbf{b}_{ij})(\mu_j - \mathbf{b}_{ij}^T) + \mathbf{B}_{ij}] \quad (3.42)$$

onde, $Q_j = \sum_i Q_{ij}$.

A extensão XD é adequada para tratar distribuições com dados faltantes, sendo muito utilizada no tratamento de dados observacionais principalmente em astrofísica e cosmologia.

3.6 Redução de dimensionalidade e extração de características

Nesta seção, pretendemos responder a questão de quantos dados são necessários para estimar a probabilidade de se encontrar pontos em um espaço de alta dimensão. Para exemplificar esse problema, digamos que temos N pontos em uma distribuição uniforme d -dimensional, descrita por um hipercubo (centrado na origem) com comprimento de aresta 2 (isto é, cada coordenada está entre $[-1, 1]$). Considerando a métrica Euclidiana, qual a probabilidade de encontrarmos pontos dentro de uma unidade de distância da origem? Se os pontos estiverem uniformemente distribuídos ao longo do volume, então uma boa estimativa é a razão entre o volume de uma hiperesfera e um hipercubo, ambos de raio unitário centrados na origem. Desta forma, para duas dimensões teríamos:

$$f_2 = \frac{\pi r^2}{(2r)^2} = \frac{\pi}{4} \approx 78,5\% \quad (3.43)$$

para três dimensões:

$$f_3 = \frac{(4/3)\pi r^3}{(2r)^3} = \frac{\pi}{6} \approx 52,3\% \quad (3.44)$$

Generalizar para dimensões mais altas exigem algumas fórmulas menos "familiares" para volumes de hiperesferas. Pode-se demonstrar que o volume de uma hiperesfera d -dimensional com raio r é dado por:

$$V_d(r) = \frac{2r^d \pi^{d/2}}{d\Gamma(d/2)}, \quad (3.45)$$

onde $\Gamma()$ é a função gamma de euler ([IVEZIC et al., 2014](#)). Verificamos que para $d = 2$ e $d = 3$ utilizamos respectivamente, as fórmulas familiares para a área de um círculo e volume de uma esfera. No caso d -dimensional teremos:

$$f_d = \frac{V_d(r)}{(2r)^d} = \frac{\pi^{d/2}}{2^{d-1} d \Gamma(d/2)}; \quad (3.46)$$

e podemos facilmente mostrar que,

$$\lim_{d \rightarrow \infty} f_d = 0. \quad (3.47)$$

Ou seja, no limite de infinitas dimensões, não existirá um único ponto dentro de uma distância r da origem. Em outras palavras, a fração de pontos dentro de um raio de

pesquisa r tende a zero à medida que é aumentada a dimensionalidade e, portanto, o número de pontos em um conjunto de dados necessário para a amostra uniforme desse hipervolume cresce exponencialmente. Isso é conhecido como a *maldição da dimensionalidade*. Entretanto, como veremos a seguir, existem algoritmos de projeções e extração de características que capturam as principais correlações físicas e estatísticas entre os dados medidos, reduzindo a dimensionalidade dos dados subjacentes sem perda de informação. Isso é o que veremos nas subseções seguinte.

Os métodos para redução de dimensionalidade e extração de características descritos a seguir, também fazem parte das aprendizagens não supervisionadas.

3.6.1 *Isomap*

Uma variante da *Classical Scaling* que pode lidar com dados que apresentem relações não-lineares, foi definida em (TENENBAUM, 1998; TENENBAUM; SILVA; LANGFORD, 2000), conhecida como Isometric Feature Mapping (ISOMAP). Essa técnica utiliza distâncias geodésicas ao invés de Euclidianas. Normalmente, o ISOMAP utiliza o algoritmo de *Dijkstra*(1959) (GUO et al.,) que aplica a sequência de Fibonacci para calcular a menor distância entre dois vértices através de grafos criados por distâncias geodésicas, apresentando complexidade computacional $\mathcal{O}(N^2)(k + \log N)$. Outros métodos como o algoritmo Floyd-Warshall (BURFIELD, 2013) realizam as mesmas operações com $\mathcal{O}(N^3)$, uma abordagem de "força bruta" para este tipo de problema, exigiria um tempo total de computação $\mathcal{O}(N^2 k^N)$.

3.6.2 *Principal Component Analysis (PCA)*

A análise de componentes principais (*PCA*), assim como o método de análise de fatores (*Factor Analysis*) (NG, 2000), busca identificar o subespaço em que os dados aproximadamente se encontram relacionando suas características principais. No entanto o PCA faz isso de forma mais direta exigindo apenas um cálculo de vetor próprio e não necessitando do *algoritmo EM*.

A *PCA* também é conhecida como *transformada discreta de Karhunen-Loève (KLT)* ou ainda *transformada de Hotelling* (em homenagem a Kari Karhunen, Michel Loève e Harold Hotelling) e também, decomposição ortogonal própria (*POD*). Basicamente, ela transforma linearmente variáveis discretas em coeficientes descorrelacionados, sendo referenciada como transformação linear ótima e tendo inúmeras aplicações no reconhecimento de padrões.

Os componentes principais representam as características mais relevantes dos dados linearmente não correlacionadas. Eles são obtidos através do cálculo dos autovalores e autovetores da matriz de covariância, que compõem a *matriz da transformada de Hotelling*.

3.6.2.1 Transformada de Hotelling

Sendo a matriz de covariância real e simétrica, é sempre possível encontrar um conjunto de autovetores ortonormais. As linhas da *matriz da transformada de Hotelling* são formadas a partir dos autovetores da matriz de covariância arranjados em ordem decrescente, com a primeira linha sendo o autovetor correspondente ao maior autovalor e sucessivamente até o menor autovalor.

Considere o exemplo em que queremos mapear o espaço $\mathcal{X}$ em um subespaço principal $\mathcal{Y}$. Para isso, podemos descrever uma transformação a partir da matriz de Hotelling ($\mathcal{H}$) que mapea os valores x em uma distribuição de valores y com $\mu_y = 0$ e cuja matriz de covariância ($\mathbf{C}_y$) pode ser obtida através de $\mathcal{H}$ e $\mathbf{C}_x$:

$$\mathbf{C}_y = \mathcal{H}^T \mathbf{C}_x \mathcal{H} \quad (3.48)$$

e assim, temos que:

$$\mathbf{C}_x = \mathcal{H} \mathbf{C}_y \mathcal{H}^T \quad (3.49)$$

Nessa matriz diagonal os elementos são os autovalores de $\mathbf{C}_x$, como os elementos fora da diagonal principal de $\mathbf{C}_y$ são zeros, os elementos dos vetores y são descorrelacionados.

Desta forma, podemos escrever a transformada de Hotelling para o nosso exemplo, pela relação:

$$y = \mathcal{H}(x - \mu_x) \quad (3.50)$$

Podemos também, recuperar qualquer vetor x a partir do y correspondente:

$$x = \mathcal{H}^T y + \mu_x \quad (3.51)$$

Os autovetores da matriz de covariância de fato formam uma nova base que segue a variância nos dados, com o maior autovalor correspondendo a característica que apresenta maior variância.

Outra técnica muito utilizada em PCA é a decomposição em valores singulares (SVD) que representa a fatoração de uma matriz real ou complexa (em nosso caso a matriz de covariância ou de correlação) e tem ampla aplicação em processamento de sinais na estatística. A decomposição em valores singulares é bastante geral, já que pode ser aplicada a qualquer matriz $m \times n$, ao passo que a decomposição em autovalores pode ser aplicada

apenas para alguns tipos de matrizes quadradas. Para o nosso caso a SVD escalonada pode se escrita,

$$\mathbf{U}\mathbf{\Sigma}\mathbf{V}^* = \frac{1}{\sqrt{N-1}}\mathbf{X} \quad (3.52)$$

onde a matriz $\mathbf{X}$ contém as subtrações $(x - \mu)$, $\mathbf{U}$ é a matriz dos singulares vetores esquerdos (unitária e pode ser real ou complexa), $\mathbf{\Sigma}$ é uma matriz quadrada com números reais positivos na diagonal e $\mathbf{V}^*$ é a matriz transposta conjugada, também unitária e pode ser real ou complexa. Usando essa expressão para a matriz de covariância, podemos escrever:

$$\mathbf{C}_x = \left[\frac{1}{\sqrt{N-1}}\mathbf{X} \right]^T \left[\frac{1}{\sqrt{N-1}}\mathbf{X} \right] \quad (3.53)$$

$$\mathbf{C}_x = \mathbf{V}\mathbf{\Sigma}\mathbf{U}^T\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \quad (3.54)$$

$$\mathbf{C}_x = \mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^T \quad (3.55)$$

Comparando esse último resultado com $\mathbf{C}_x = \mathcal{H}\mathbf{C}_y\mathcal{H}^T$, vemos que os vetores singulares direitos $\mathbf{V}$ correspondem a matriz de Hotelling e a matriz diagonal de autovalores $\mathbf{C}_y$ é equivalente ao quadrado dos valores singulares:

$$\mathbf{\Sigma}^2 = \mathbf{C}_y \quad (3.56)$$

O formalismo SVD também pode ser usado para observar a relação entre a matriz de covariância $\mathbf{C}_x$ e a matriz de correlação $\mathbf{M}_x$,

$$\mathbf{M}_x = \frac{1}{N-1}\mathbf{X}\mathbf{X}^T = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\mathbf{V}\mathbf{\Sigma}\mathbf{U}^T = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^T \quad (3.57)$$

Assim, temos três formas equivalentes de realizar a análise de componentes principais: a decomposição em autovalores de $\mathbf{C}_x$, a decomposição em autovalores de $\mathbf{M}_x$ ou a SVD de $\mathbf{X}$. Geralmente os resultados de PCA são discutidos em termos pontuações (scores) de componentes, onde os valores das variáveis transformadas correspondem a uma pontuação particular, e ao peso pelo qual cada variável normalizada original deve ser multiplicada para se obter a pontuação do componente.

3.6.3 Independent Components Analysis (ICA)

A análise de componentes independentes (ICA) é uma técnica muito comum na biomedicina estatística e em processamento de sinais, para a resolução da questão, muitas vezes referida como "problema do coquetel da festa". Nesse problema temos vários microfones situados em

todo o quarto onde temos $\mathbf{n}$ pessoas conversando. Cada microfone captura uma combinação linear das $[\mathbf{n}]$ vozes. O objetivo da ICA é usar o conceito de independência estatística (NG, 2000) para isolar os sinais sonoros individuais. No contexto astronômico podemos considerar, por exemplo, a aplicação da ICA para uma série de espectros de galáxias. Neste contexto, cada espectro de galáxia pode ser comparado a um microfone que capta uma combinação linear de sinais individuais de estrelas e regiões com abundância de H_I . Desta forma, cada espectro de galáxia pode ser descrito por:

$$\begin{aligned}x_1 &= a_{11}S_1(k) + a_{12}S_2(k) + a_{13}S_3(k) + \dots \\x_2 &= a_{21}S_1(k) + a_{22}S_2(k) + a_{23}S_3(k) + \dots \\x_3 &= a_{31}S_1(k) + a_{32}S_2(k) + a_{33}S_3(k) + \dots\end{aligned}\tag{3.58}$$

onde $S_i(k)$ são os espectros estelares individuais e a_{ij} são as amplitudes apropriadas a mistura. Na forma matricial podemos reescrever:

$$\mathbf{X} = \mathbf{AS}\tag{3.59}$$

onde $\mathbf{X}$ e $\mathbf{S}$ são respectivamente as matrizes do conjunto de espectros de entrada e do conjunto de espectros estelares individuais. Extrair estes sinais de espectros separados é equivalente a estimar uma matriz de pesos $\mathbf{W}$ apropriada de tal forma que

$$\mathbf{S} = \mathbf{WX}\tag{3.60}$$

O princípio fundamental da ICA é que os sinais de entrada observados $\mathbf{S}_i(k)$ devem ser *estatisticamente independentes*. Duas variáveis aleatórias são ditas independentes estatisticamente se a distribuição de probabilidade do conjunto $f(x, y)$ pode ser totalmente descrita por uma combinação de suas probabilidades marginais,

$$f(x^p, y^q) = f(x^p)f(y^q),\tag{3.61}$$

onde, p e q representam momentos arbitrários de maior ordem da distribuição. Para o caso da PCA, $p = q = 1$, o que suaviza as condições mais sutis na correlação dos dados.

Relacionar a ICA com a redução de dimensionalidade e extração de características é importante, uma vez que ICAs são interessantes para separar informações através dos dados de entrada. A diferença é que em uma busca por projeção identificamos as informações uma de cada vez enquanto na ICA a pesquisa dos dados individuais pode ser realizada ao mesmo tempo. A busca por projeções pode ser entendida como um subconjunto da análise de componentes independentes.

3.6.4 Locally-Linear Embedding (LLE)

A incorporação localmente linear (*LLE*) (Roweis e Saul, 2000, 2001) é uma técnica para redução de dimensionalidade baseada na suposição de que localmente os dados são lineares, isto significa dizer que os dados são invariantes por rotação, translação e escala, localmente. Para mais detalhes e acesso ao código veja cs.nyu.edu/~roweis/lle/.

Consideramos novamente o exemplo anterior, onde queremos representar o espaço $\mathcal{X}$ por um subespaço $\mathcal{Y}$ d -dimensional. Nesse caso, o algoritmo LLE primeiro encontra os vizinhos mais próximos de cada objeto em $\mathcal{X}$, em seguida expressa cada objeto x_i como uma combinação linear dos outros objetos, isto é, $x_i = \sum_j w_{ij}x_j$, onde $w_{ij} = 0$ se x_j não é um vizinho próximo de x_i . Enfim, ele encontra as coordenadas de cada ponto no espaço d -dimensional usando os pesos encontrados anteriormente. A matriz de pesos $\mathcal{W}$, cujas entradas são w_{ij} , é encontrada através da minimização do erro:

$$error(\mathcal{W}) = \sum_i \left(x_i - \sum_j w_{ij}x_j \right)^2 \quad (3.62)$$

Com a matriz de pesos $\mathcal{W}$ e o número de dimensões d escolhido, o algoritmo realiza um "mapeamento preservando a vizinhança" dos dados em um espaço de menor dimensão representado pela matriz de dados $\mathcal{Y}$, onde a i -ésima linha é o y_i que corresponde a x_i . Então essa projeção pode ser obtida encontrando a matriz $\mathcal{Y}$ que minimiza a relação:

$$error(\mathcal{Y}) = \sum_i \left(y_i - \sum_j w_{ij}y_j \right)^2 \quad (3.63)$$

Essa minimização é normalmente realizada por métodos que utilizam autovetores da matriz $error(\mathcal{Y})_{(n \times n)}$. Desta forma, LLE é uma técnica $\mathcal{O}(n^2)$, porém alguns métodos para determinação de autovetores podem ser empregados para reduzir essa complexidade.

3.6.5 t-distributed Stochastic Neighbor Embedding (t-SNE)

É uma técnica de redução de dimensionalidade não linear que é particularmente adequada para incorporar dados de alta dimensão em um espaço de duas ou três dimensões. Especificamente, ele modela cada objeto de alta dimensão por um ponto (bi-tridimensional), de tal forma, que objetos semelhantes são modelados como pontos próximos e objetos diferentes são modelados por pontos distantes.

O algoritmo *t-SNE* compreende duas etapas principais. Em primeiro lugar, t-SNE constrói uma distribuição de probabilidade sobre pares de objetos em alta dimensão fazendo com que objetos semelhantes tenham uma alta probabilidade de serem coletados, enquanto objetos diferentes têm uma probabilidade extremamente pequena de serem colhidos. Em

segundo lugar, o *t-SNE* define uma distribuição de probabilidade semelhante aos pontos no mapa de baixa dimensão e minimiza a divergência de *Kullback-Leibler* entre cada duas distribuições em relação as suas localizações no mapa.

A divergência de *Kullback-Leibler* é uma medida de como uma distribuição de probabilidade difere de outra distribuição de probabilidade esperada. Considere as distribuições P e Q , desta forma podemos descrever a divergência através da relação:

$$KL(P||Q) = \sum_{i \neq j} P_{ij} \log \frac{P_{ij}}{Q_{ij}} \quad (3.64)$$

A minimização da divergência de Kullback-Leibler é realizada utilizando o algoritmo de gradiente descendente com um elemento estocástico para evitar mínimos locais. O resultado dessa otimização é um mapa que reflete bem as semelhanças entre as entradas de alta dimensão.

Para um exemplo teórico de aplicação, considere um conjunto de n dados em alta dimensão $(x_1 + x_2 + \dots + x_n)$, o t-SNE primeiro calcula as probabilidades P_{ij} que são proporcionais à semelhança entre x_i e x_j , através das seguintes relações:

$$P_{ij} = \frac{P_{j|i} + P_{i|j}}{2n}, \quad (3.65)$$

$$P_{j|i} = \frac{e^{(-\|x_i - x_j\|^2 / 2\sigma_i^2)}}{\sum_{k \neq i} e^{(-\|x_i - x_k\|^2 / 2\sigma_i^2)}}. \quad (3.66)$$

Que são respectivamente as probabilidades condicionais e conjuntas entre i e j . Segundo *Geofrey Hinton* e *Laurens Van der Maten*, "A semelhança do ponto x_j com o ponto x_i é a probabilidade condicional $P(x_j|x_i)$, o ponto x_i só escolherá x_j como vizinho em proporção à sua densidade de probabilidade sob uma distribuição gaussiana centrada em x_i ".

O método *t-SNE* diferencia-se do método *SNE*, pela escolha da distribuição de probabilidade definida entre pontos, usando a distribuição *t-student* em vez de gaussiana, e pela escolha exata da função de custo para minimização. A distribuição *t-student* é semelhante a distribuição normal, porém com caudas mais longas, ou seja, ela é capaz de englobar mais valores extremos. Por exemplo, considere que o conjunto dados anteriormente mencionado tenha valor esperado igual a μ , assim podemos escrever as seguintes relações:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}; \sigma^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2; \quad (3.67)$$

$$t = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}. \quad (3.68)$$

A distribuição *t-student* com ν graus de liberdade pode ser também definida como a distribuição da variável aleatória t . Supondo Z uma variável aleatória de distribuição normal padrão com $\mu = 0$ e $\sigma = 1$, V é uma variável aleatória com distribuição χ^2 e ν graus de liberdade, se Z e V forem independentes, então a distribuição da variável aleatória t será:

$$t = \frac{Z}{\sqrt{V/\nu}} \quad (3.69)$$

Assim, vemos que t é uma quantidade "crucialmente útil" no caso em que a média e a variância são parâmetros desconhecidos. Se as classes forem bem separadas no plot *t-SNE* então podemos esperar uma classificação precisa. Entretanto isso não é estritamente verdade e o conjunto de características com fraca separação entre classes pode ainda ser bem classificado. Em (LOCHNER et al., 2016) cada um dos métodos de extração de características descritos foram plotados com *t-SNE* para uma visualização prévia e posterior. Nos resultados, apresentamos alguns dos plots t-SNE obtidos para a nossa análise.

3.6.6 Ondaletas (*Wavelet*)

A primeira menção sobre *wavelets* foi feita em 1909 na tese de doutorado de *Alfred Haar* (1885 – 1933). Apesar de a *Wavelet de Haar* apresentar uma forma compacta, não foi extensivamente aplicada devido ao fato de não ser continuamente diferenciável. A técnica de ondaletas ou *wavelet*, realmente se consolidou na década de 80, principalmente através dos trabalhos de *A. Grossmann and J. Morlet* (MORLET, 1983; GROSSMANN; MORLET, 1984; GROSSMANN; MORLET, 1985; GROSSMANN, 1988). Há algum tempo estamos familiarizados com os efeitos da aplicação dessa técnica seja no processamento de sinais naturais (como em nosso estudo) ou no processamento de sinais digitais (como o processamento de imagens e sons). A maioria das inovações tecnológicas que imaginamos em nosso cotidiano são consequências de estudos envolvendo matemática, física e engenharias, no contexto de *transformadas multiescalares*.

Uma das principais ferramentas da teoria *Wavelet* é a transformada *Wavelet*. Ela atua como um microscópio matemático cuja a ampliação é dada pelo inverso do parâmetro de dilatação, designado por s e a capacidade óptica, pela função *Wavelet* analisadora escolhida, designada por ψ (OLIVEIRA, 2007). A transformada *Wavelet* não tem apenas uma função-base como na transformada de Fourier. Na realidade, como veremos existe um grande número de funções que podem ser consideradas como *wavelets-mãe*. Independentemente a *WT* é semelhante a transformada de Fourier em muitos aspectos (MORETTIN, 1999) e por isso, começamos nosso estudo com uma breve revisão sobre as transformadas de Fourier.

A análise clássica de Fourier estabelece que um sinal pode ser representado por uma soma infinita de termos em seno e cossenos, mais conhecida como expansão de Fourier. A transformada de Fourier é uma técnica muito utilizada por diversas áreas, no estudo de processamento de sinais e basicamente consiste em transformar um domínio em outro, onde características importantes do sinal analisado, são preservadas. O domínio transformado normalmente é denominado domínio espectral ou frequencial, enquanto o domínio original da função é chamado de domínio temporal ou espacial.

Geralmente a denominação *transformada de Fourier* refere-se à transformada de Fourier para funções contínuas, que é capaz de representar qualquer função integrável $f(t)$ como a soma de exponenciais complexas com frequência angular ω medida em *rad/s* e amplitude complexa $\mathcal{F}(\omega)$:

$$\mathcal{F}(\omega) = \int_{-\infty}^{\infty} f(t)e^{-i\omega t} dt \longrightarrow f(t) = \mathcal{F}^{-1}(\mathcal{F}(\omega)) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathcal{F}(\omega)e^{i\omega t} d\omega, \quad (3.70)$$

onde $f(t)$ deve satisfazer a relação,

$$\int_{-\infty}^{\infty} |f(t)|^2 dt < \infty \quad (3.71)$$

,

Essa restrição demonstra que o uso de funções globais, nem sempre são eficazes na representação local de sinais. Necessitamos portanto, de uma representação que seja mais localizada e combine os domínios frequencial e temporal. Isto é muito importante quando queremos analisar partes do domínio temporal do sinal que são responsáveis por características importantes do espectro. Com esse propósito podemos multiplicar o sinal original f por uma função que seja diferente de zero no intervalo desejado, e zero fora desse intervalo. Normalmente essa função é designada por $g_{u,\omega}(t)$ (conhecida como função janela - *window function*) e também satisfaz à relação $\int_{-\infty}^{\infty} |g_{u,\omega}(t)|^2 dt < \infty$. Assim, o produto $f(t)g_{u,\omega}^*(t)$ contém informações de f perto de $t = u$. Variando o parâmetro u , deslizamos a janela ao longo do eixo temporal de modo que podemos analisar o comportamento local do sinal f em diferentes intervalos. Normalmente dizemos que $f(t)g(t-u)e^{-i\omega t}$ é o sinal *janelado*.

Motivado pela mecânica quântica, *Dennis Gabor* 1(900 – 1979) ([GABOR, 1946](#)) define formas de ondas atômicas elementares em tempo e frequência que apresentam mínimo espalhamento no plano tempo-frequência, através de uma decomposição do sinal sobre as formas de ondas "atômicas". *Gabor* demonstrou a importância do processamento de sinais com localização em tempo e frequência definindo a *transformada janelada de Fourier*

(transformada de Gabor ou *STFT* (*Short-time Fourier Transform*)):

$$\mathcal{F}(u, \omega) = \int_{-\infty}^{\infty} f(t)g(t - u)e^{-i\omega t} dt \quad (3.72)$$

onde g é a *função janela*. Desta forma, diferentemente da transformada de Fourier original, em que o sinal $f(t)$ deve ser conhecido em todo o eixo temporal antes que o espectro em cada frequência possa ser calculado, na *transformada de Fourier janelada*, precisamos de conhecer $f(t)$ apenas no intervalo em que $g(t - u) \neq 0$. Existem várias escolhas para a *função janela*, mas uma vez fixada a janela para a *transformada de Fourier janelada*, a resolução no tempo e na frequência (t e f) permanece constante em todo o plano $t \times f$.

O trabalho de Fourier remete ao princípio da incerteza de Heisenberg (1927), demonstrando que um sinal não pode ser restrito simultaneamente ao tempo e frequência (GABOR, 1946; DAUBECHIES, 1989), isto é, as variações em tempo Δt e frequência $\Delta \omega$ são inversamente proporcionais, ou seja:

$$\Delta t \cdot \Delta \omega \geq \frac{1}{2} \quad (3.73)$$

em que a constante $1/2$ pode assumir outros valores, dependendo da normalização utilizada na transformada de Fourier. Essa equação também é mencionada como o *princípio de incerteza da informação quântica* (OLIVEIRA, 2007). Assim objetivamos encontrar um método que possa fornecer uma boa resolução em tempo e frequência para qualquer localização do plano tempo-frequência, ou seja, queremos uma função janela cujo raio aumente com o tempo (enquanto resolve conteúdos em baixa frequência) e uma função janela que decresce com o tempo (enquanto resolve conteúdos de alta frequência do sinal). Para isso, ao invés de uma interpretação nos domínios de frequência e tempo ($f \times t$), utilizamos as denominações de escala e translação ($s \times \tau$) respectivamente, isto conduz ao desenvolvimento das *funções wavelets* e *transformadas wavelets*.

3.6.6.1 Transformada wavelet contínua (CWT)

A *continuous wavelet transform* (CWT), é uma transformada integral linear muito utilizada no estudo de características de sinais, extraindo informações de variações em bandas de frequências específicas e detectando estruturas locais presentes (ANTOINE et al., 2008). Dado um sinal f , sua transformada integral (OLIVEIRA, 2007) é definida por:

$$\mathcal{W}_f^\psi(s, \tau) = \int_{-\infty}^{\infty} f(t)\psi_{s,\tau}^*(t)dt \quad (3.74)$$

Essa equação mostra como uma função $f(t)$ é decomposta em um conjunto de funções de base $\psi_{s,\tau}(t)$. As wavelets são geradas a partir de uma base única $\psi(t)$ (chamada de

monther wavelet) para escala e translação:

$$\psi_{s,\tau}(t) = \frac{1}{\sqrt{s}} \psi\left(\frac{t-\tau}{s}\right) \implies \mathcal{W}_f^\psi(s, \tau) = \frac{1}{\sqrt{s}} \int_{-\infty}^{\infty} f(t) \psi^*\left(\frac{t-\tau}{s}\right) dt \quad (3.75)$$

onde o parâmetro s ($s > 0$) se refere a escala, τ a translação ou localização da função wavelet ψ e o termo $s^{-1/2}$ é para a normalização da energia em todas as diferentes escalas. O parâmetro s controla a dilatação ou contração da função wavelet analisadora assim, à medida que o parâmetro τ varia, o sinal f é analisado localmente em torno dele e podemos observar aspectos multiescalares do sinal estudado. O ψ^* é o complexo conjugado da função ψ e $\mathcal{W}_f^\psi(s, \tau)$ é o *coeficiente wavelet*. Na representação contínua os parâmetros de escala s e de localização τ assumem valores reais e contínuos. Em algumas análises de sinais discretos, tais valores podem ser discretizados respeitando condições específicas para o propósito.

O tratamento de dados discretos faz uso da *discrete wavelet transforms (DWT)*, que assim como na transformada discreta de Fourier (*FFT*), fornece um rápido algoritmo para o cálculo de transformadas, realizando a síntese exata de conjuntos finitos e discretos de coeficientes. Em (LOCHNER et al., 2016) foi adotada a *stationary wavelet transform (SWT)* que é um algoritmo de transformada, ele foi projetado para superar problemas de invariância-translação da *DWT*. Esse algoritmo é mais conhecido como "algoritmo à trous" em referência a inserção de zeros em dados faltantes. Para mais detalhes sobre a *DWT* sugerimos (MALLAT, 2009).

Importante notar que nas equações 3.74, 3.75 não especificamos a função wavelet analisadora. Essa é a diferença básica entre a transformada wavelet e outras transformadas. A teoria de wavelets define apenas as propriedades gerais das wavelets e da transformada wavelet, criando uma estrutura em que podemos projetar wavelets à nossa maneira. Na realidade, já existe um grande número de funções de base que podem ser consideradas como *wavelet-mãe* (harr, daubechies, symlets, coiflets, biorthogonal, reverse biorthogonal, meyer, gaussian, mexican hat, morlet, complex gaussian, shannon, legendre, dentre outras). O termo mãe, vem do fato de que as funções envolvidas são versões expandidas ou contraídas de uma mesma wavelet principal, à *monther wavelet*.

3.6.6.2 Propriedades

As propriedades mais importantes das wavelets, são a admissibilidade e regularidade. A admissibilidade de uma função wavelet $\psi(t)$ é dada por:

$$\int_{-\infty}^{\infty} \psi(t) dt = 0 \quad (3.76)$$

isto implica que a função analisadora tem valor médio nulo. Podemos também expressar no domínio de Fourier:

$$c_\psi = 2\pi \int_{-\infty}^{\infty} \frac{|\psi(\omega)|^2}{|\omega|} d\omega < \infty, \quad (3.77)$$

em que $\psi(\omega)$ é a transformada de Fourier da função wavelet ψ e a constante c_ψ é conhecida como *constante de admissibilidade*. Essa característica ondulatória da origem ao nome wavelet em inglês (na realidade o termo original veio do francês ondelette, proposto pelo físico Alex Grosman).

Nestes casos, a transformada inversa *wavelet* pode ser obtida por:

$$f(t) = \frac{1}{c_\psi} \int_{-\infty}^{\infty} \int_0^{\infty} \mathcal{W}_f^\psi(s, \tau) \psi_{s,\tau}(t) ds d\tau, \quad (3.78)$$

onde $\psi_{s,\tau}(t)$ vem de 3.74. A regularidade de uma função *wavelet* é dada pela propriedade de *momentos nulos* formalizada através de:

$$\int_{-\infty}^{\infty} t^m \psi(t) dt = 0, \quad (3.79)$$

para $0 \leq m \leq M$. Esta propriedade de momentos nulos torna possível a análise da regularidade local de sinais. Se a função *wavelet* analisadora é regular e de ordem $n \geq 1$, tal que seus momentos estatísticos até essa ordem também sejam iguais a zero, suas derivadas de n ordem também são zeros, isto é, se a transformada wavelet de uma função é zero, a função *wavelet* tem momentos nulos.

Outra condição para as funções wavelets analisadoras é energia unitária definida pela relação:

$$\int_{-\infty}^{\infty} |\psi(t)|^2 dt = 1, \quad (3.80)$$

essa condição garante que a função wavelet tenha um suporte compacto, tendo um decaimento rápido de amplitude conhecido como *suporte efetivo*, garantindo a localização espacial.

3.6.6.3 Power spectrum e entropia wavelet

Analogamente ao espectro de potência da transformada de Fourier, é possível fazer a integração do parâmetro τ sobre as amplitudes quadráticas dos coeficientes wavelets fornecendo o que é também conhecido como escalograma global:

$$\mathcal{P}_W(s) = \int_{-\infty}^{\infty} |\mathcal{W}_f^\psi(s, \tau)|^2 d\tau, \quad (3.81)$$

Os resultados do escalograma global podem ser interpretados como a distribuição da energia do sinal por escala s .

Em analogia à definição de entropia na teoria da informação, a entropia wavelet contínua foi definida em (SELLO, 2003), sendo:

$$\mathcal{S}_\psi(s, \tau) = \int_0^\infty -\frac{|\mathcal{W}_f^\psi(s, \tau)|^2}{\int_{-\infty}^\infty |\mathcal{W}_f^\psi(s, \tau)|^2 ds} \ln \left[\frac{|\mathcal{W}_f^\psi(s, \tau)|^2}{\int_{-\infty}^\infty |\mathcal{W}_f^\psi(s, \tau)|^2 ds} \right] ds \quad (3.82)$$

Essa entropia *wavelet* é mínima quando o sinal é caracterizado por uma atividade organizada, e máxima, quando o sinal é constituído de superposições de um grande número de processos.

3.6.6.4 Wavelet de Haar

É a primeira função analisadora wavelet. Ela foi definida conforme a seguir:

$$\psi = \begin{cases} 1 & \text{se } 0 \leq t < 1/2, \\ -1 & \text{se } 1/2 \leq t < 1, \\ 0, & \text{nos outros casos} \end{cases} \quad (3.83)$$

vemos que a função é real e antissimétrica com respeito a $t = 0$, além disso, o intervalo em que ela é diferente de zero é finito. Ou seja, ela possui uma boa localização no espaço temporal, porém não no espaço frequencial. Tal característica reduz significativamente a sua aplicação prática no estudo de sinais. A wavelet de Haar é não regular, sua derivada primeira é descontínua, e seu momento de primeira ordem não é zero.

3.6.6.5 Wavelet de Shannon

Esta função também é chamada de *wavelet sinc* e definida por:

$$\psi(t) = \text{sinc}\left(\frac{t}{2}\right) \cos\left(\frac{3\pi t}{2}\right), \quad (3.84)$$

onde $\text{sinc}(t)$ denota $\sin(\pi t)/(\pi t)$. Existe também a versão complexa *wavelet de Shannon* definida como:

$$\psi(t) = \text{sinc}(t)e^{-i2\pi t}. \quad (3.85)$$

Essa wavelet apresenta uma boa resolução frequencial, e em contrapartida, pobre resolução temporal. Além disso, essa função *wavelet* é infinitamente diferenciável e possui infinitos momentos nulos.

3.6.6.6 Wavelet de Morlet

Foi uma das primeiras transformadas *wavelets*. De uma forma simplificada, essa wavelet-mãe consiste de uma onda plana modulada por uma função gaussiana, descrita por:

$$\psi(t) = \exp\left(\frac{-t^2}{2\sigma^2}\right) \exp(i\omega_0 t) \quad (3.86)$$

Em geral escolhe-se a forma ortogonal (MORLET, 1983), em que $\sigma = 1$ e $\omega_0 = 5$ de tal forma, que a condição de admissibilidade seja aceita. O parâmetro σ fornece a informação de quão boa é a relação entre tempo e frequência. Valores pequenos de σ fazem com que tenhamos uma boa resolução temporal, enquanto valores maiores contribuem na resolução frequencial. Observamos que para $\sigma < 1$ ocorre uma quebra na condição de admissibilidade.

3.6.6.7 Wavelet chapéu mexicano

Normalmente definida por:

$$\psi(t) = \frac{2}{\sqrt{3}}\pi^{-1/4}(1 - t^2) \exp\left(\frac{-t^2}{2}\right). \quad (3.87)$$

Essa função wavelet-mãe é muito utilizada em estudos onde a localização temporal é mais importante que a frequencial.

Aqui relatamos algumas das funções mais concisas, entretanto existem muitas outras *funções wavelet* como já mencionadas anteriormente que podem assumir formas mais complexas, mas deixamos isso para um estudo teórico a parte.

3.7 Avaliando a aprendizagem

Na descrição dos métodos avaliativos de aprendizagem, consideraremos o caso do problema da classificação de supernovas em tipo Ia ou nIa. Escolhemos uma classificação binária como exemplo, por simplicidade (mais podemos generalizar para casos com classificação múltipla). Nesse exemplo, a eficácia dos métodos de aprendizagem utilizados para a tarefa (*ANNs*, *BDTs*, dentre outros), pode ser avaliada utilizando diferentes métricas. Uma das formas de avaliar a eficácia de um classificador é olharmos para a *matriz confusão*. Para isso, primeiro precisamos ter um conjunto de predições de treinamento (predições também podem ser realizadas com conjuntos de testes, mas lembramos que o conjunto de teste deve ser utilizado apenas no final do projeto). Para cada instância em um conjunto de dados, o preditor então calcula a pontuação em relação as classes envolvidas com base em alguma *função de decisão* (equivalente a uma função de ativação). Se a pontuação for maior que certo limiar (*decision threshold*) a instância é atribuída como positiva, caso contrário, ela é atribuída como negativa.

		CLASSE VERDADEIRA	
CLASSE PREDITA		Positivo	Negativo
	Positivo	True Positive (TP)	False Positive (FP)
	Negativo	True Negative (TN)	False Negative (FN)

Em seguida, comparamos as predições e os seus reais valores, os erros e acertos do preditor formam a *matriz confusão*. Os elementos da diagonal dessa matriz contém o número de acertos para cada classe e a matriz ideal apresenta todos os elementos fora da diagonal iguais a zero. A tabela a seguir mostra um exemplo de *matriz de confusão* genérica.

Observamos que podemos estender essa idéia para um caso com multiclases, onde uma matriz de confusão pode ser montada para cada predição de classe. Outra possibilidade, apesar de incomôda devido aos graus de liberdade aumentarem quadraticamente com o número de classes, é montar uma matriz n -dimensional onde n representa o número de classes a serem preditas (VISA et al., 2011). A *matriz de confusão* fornece muitas informações sobre os preditores, sendo que umas são mais concisas do que outras. A seguir mencionamos algumas:

$$acurácia = \frac{TP + TN}{TP + FP + FN + TN}, \quad (3.88)$$

a *acurácia* do preditor é a razão entre os acertos e o total de predições, normalmente ela é calculada através de métodos de *cross-validation*. Outra métrica importante é a *acurácia* das predições positivas, chamada de *precisão* (ou *purity*) do preditor (ou classificador):

$$precisão = \frac{TP}{TP + FP}, \quad (3.89)$$

porém, a *precisão* pode não ser uma medida útil nos casos em que não temos *falsos positivo* (FP). Por isso ela é normalmente empregada junto com outra métrica conhecida como *recall* (também chamada de *sensibilidade*, *true positive rate* (TPR) ou *completeness*):

$$recall = \frac{TP}{TP + FN}. \quad (3.90)$$

Muitas vezes é interessante analisar as curvas para a *precisão* e para o *recall* em função do *threshold* ou o gráfico *precisão*×*recall*. A escolha da *decision threshold* afeta as medições de *precisão* e *recall*, quanto maior o *threshold* maior a discrepância entre eles (o que nitidamente, não é um bom resultado).

Também pode ser conveniente combinar *precisão* e *recall* em uma única métrica, como por exemplo, F_1score (ou $F-score$):

$$F_1score = \frac{2}{\frac{1}{recall} + \frac{1}{precisão}} = 2 \times \frac{precisão \times recall}{precisão + recall} = \frac{TP}{TP + \frac{FN+FP}{2}}. \quad (3.91)$$

A F_1 score favorece preditores que tenham *precisão* e *recall* similares, ajudando a determinar o melhor *decision threshold*. Mas isso nem sempre é o que queremos: em alguns contextos podemos estar interessados na *precisão* e em outros com o *recall* apenas. Infelizmente não podemos ter esse tipo de análise, pois se aumentamos a *precisão* reduzimos o *recall* e vice-versa. Isso é chamado de *precision/recall tradeoff* (GÉRON, 2017).

Além dessas, outras métricas podem ser utilizadas como o *coeficiente de Gini* e *estatísticas Kappa* (POWERS, 2012; VIERA; GARRETT et al., 2005). Por exemplo, em (CHARNOCK; MOSS, 2017) e (LOCHNER et al., 2016) foi empregada uma *figure-of-merit* na comparação de resultados para classificação de supernovas, fornecidos por diferentes métodos de aprendizagem de máquina.

3.7.1 Receiver Operating Characteristic (ROC)

A curva *ROC* foi utilizada pela primeira vez durante a segunda guerra mundial, pelo exército dos Estados Unidos com o objetivo de aumentar a previsão de detecção correta de aeronaves inimigas a partir de seus radares. Desde então ela vêm sendo empregada nas mais diversas áreas e cada vez mais em aprendizagem de máquina e mineração de dados (GREEN; SWETS, 1988).

A curva característica de operação do receptor (*ROC*) é uma representação gráfica que ilustra a capacidade de predição de um classificador binário. Ela é construída plotando a taxa de verdadeiros positivos (*TPR*, outro nome para o *recall*) versus a taxa de falsos positivos (*FPR*, que é igual a um menos *true negative rate* ($1 - TNR$)) para vários valores de *threshold*. A *TNR* é também chamada de *especificidade*, assim a curva *ROC* plota a *sensibilidade (recall)* contra $1 - \text{especificidade}$. Assim, novamente temos *tradeoff* pois, aumentando o *recall (TPR)* mais falsos positivos (*FPR*) são produzidos pelo classificador. Estamos nesses casos, considerando o uso de unidade normalizadas, outra informação importante é a área sob a curva *ROC* (*AUC*). Um preditor perfeito tem *AUC* igual a um, enquanto um preditor puramente aleatório tem *AUC* igual a 0.5.

Podemos ficar na dúvida em quando é melhor utilizar *precisão/recall-curve PRC* ao invés da curva *ROC*. Em geral devemos utilizar a *PRC* sempre que os positivos forem raros ou quando estamos mais preocupados com os falsos positivos do que com os falsos negativos, e a curva *ROC* para os outros casos.

Estas são as idéias fundamentais para o desenvolvimento de classificadores (ou preditores) baseados em aprendizagem de máquina. Treinar o preditor e medir a acurácia com *cross-validation*, selecionar a *precisão/recall* que se encaixa em sua tarefa e realizar comparações entre diferentes modelos utilizando a curva *ROC* e o valor da *AUC*.

4 *Artificial Neural Networks (ANN)*

Uma das mais remotas influências nesse entendimento, talvez seja o "Pato de Vaucason" um autômato idealizado pelo inventor francês *Jacques de Vaucason* (1709 – 1782), sem deixar de mencionar também o dispositivo teórico da "máquina de Turing" (TURING, 1937), concebido pelo matemático britânico *Alan Mathison Turing* (1912 – 1954). Tais idéias influenciaram o imaginário de muitos estudiosos desde então.

As primeiras publicações sobre neurocomputação surgiram em 1943, fruto de trabalhos conjuntos principalmente de Warren McCulloch, um neuroanatomista e psiquiatra, com o matemático Walter Pitts (MCCULLOCH; PITTS, 1943; LANDAHL; MCCULLOCH; PITTS, 1943). Ainda no final da década de 40 e início dos anos 50 tivemos o desenvolvimento dos significativos trabalhos do matemático naturalizado americano *John von Neumann* (1903 – 1957) (RILEY, 1987). Outra importante contribuição na área, foi a publicação do psicólogo canadense *Donald Olding Hebb* (1904 – 1985) (HEBB, 1949), que descreveu um dos primeiros sistemas de aprendizagem atualmente conhecido como *aprendizagem hebbiana*. Durante as décadas de 50 e 60 houveram muitas pesquisas e estudos na área, com destaque para os trabalhos do psicólogo americano *Frank Rosenblatt* (1928 – 1971) (ROSENBLATT, 1958) sobre o *Perceptron*, um modelo cognitivo que consiste em unidades de entrada conectadas a uma única camada de neurônios; e os trabalhos de Bernard Widrow e Marcian E. Hoff (WIDROW; HOFF, 1960) sobre um neurônio artificial baseado no modelo de McCulloch e Pitts, denominado *ADALINE*. Em 1969 ocorreu a "morte" da pesquisa em redes neurais influenciada pela publicação de *Minsky e Papert* (MINSKY; PAPERT, 1969) que expôs as limitações do modelo de Rosenblatt, provando que tais redes não seriam capazes de resolver uma ampla classe de problemas por exemplo, tipo "*exclusive or*" (também conhecidos por *XOR*).

Somente a partir dos anos 80, as redes neurais artificiais passaram a atrair substancial atenção e as pesquisas na área foram retomadas com grande interesse. Essa explosão de interesse foi impulsionada, principalmente, pelo avanço da tecnologia (especialmente em microeletrônica) e também pelo impacto causado pelo artigo publicado pelo físico, biólogo e neurologista americano, chamado John Joseph Hopfield (nascido em Julho de 1933) (HOPFIELD, 1982). Hopfield mostrou a relação entre redes recorrentes auto associativas e sistemas físicos, abrindo espaço para a utilização de teorias correntes da Física na construção de novos modelos.

Como vemos, a teoria sobre redes neurais artificiais é um belo paradigma de programação com inspiração biológica, que permite a um computador aprender a partir de dados. Foi desenvolvida vagamente com base no funcionamento do cérebro humano, que simplificada-

mente podemos dizer, ter cerca de 10^{11} neurônios e apresentar uma topologia altamente complexa onde cada neurônio se conecta a cerca de 10^5 outros neurônios por processos chamados sinapses. Basicamente, a sinapse ativa um transmissor químico, quando a soma dos transmissores atinge um limite faz com que o neurônio dispare uma sinapse excitatória ou inibitória. Com base nestas idéias, um primeiro modelo para o neurônio artificial foi desenvolvido:

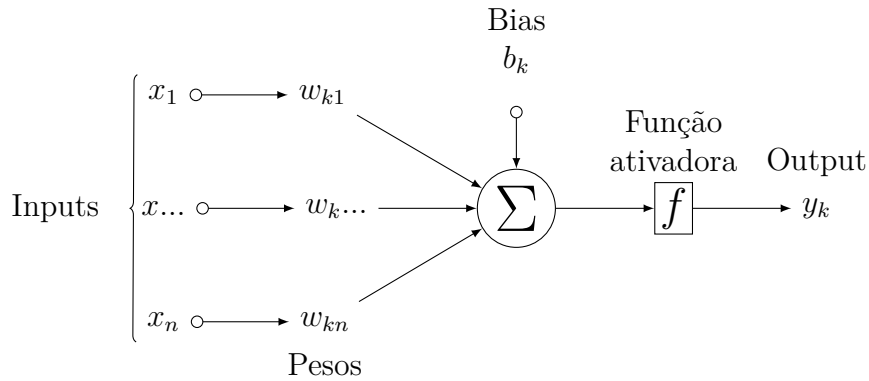


Figura 3 – Diagrama de um neurônio artificial (*Linear Threshold Unit - LTU*) (GOOSSENS; RAHTZ; MITTELBAACH, 1997)

Matematicamente o comportamento de um neurônio artificial pode ser representado pela equação a seguir:

$$y_k = f\left(\sum_{i=1}^n w_{ki}x_i + b_k\right) \quad (4.1)$$

onde x_i é o i ésimo dado de entrada, w_{ki} é o i ésimo peso do neurônio k , b^k é o *bias* para o neurônio k , f é a função ativadora do neurônio e y^k é a saída do neurônio k . Normalmente os pesos e os *bias*es são os parâmetros estimados por uma rede neural artificial durante o processo de treinamento. Após o treinamento, os valores de tais parâmetros são fixados e não havendo atualizações, eles permanecem constantes.

A determinação da função ativadora é um dos fatores determinantes para o bom funcionamento da rede neural. Normalmente as mais utilizadas são funções tipo degrau, linear, sigmoide (logística), tangente hiperbólica e ReLU. A tabela a seguir relaciona algumas dessas e outras funções ativadoras:

Nome	Função Ativadora
Degrau (<i>Heaviside</i>)	$f(u) = \begin{cases} 1 & \text{se } u \geq 0 \\ 0 & \text{se } u < 0 \end{cases}$
Linear	$f(u) = \begin{cases} 1 & \text{se } u \geq 1/2 \\ u & \text{se } -1/2 < u < 1/2 \\ 0 & \text{se } u \leq -1/2 \end{cases}$
Sigmoide (logística)	$f(u) = \sigma(u) = (1 + \exp(-\alpha u))^{-1}$
Tangente hiperbólica	$f(u) = \tanh(u)$
Arco tangente	$f(u) = \tan^{-1}(u)$
<i>Retified Linear Unit</i> (ReLU)	$f(u) = \begin{cases} u & \text{se } u \geq 0 \\ 0 & \text{se } u < 0 \end{cases}$
<i>Parametric Retified Linear Unit</i> (PReLU)	$f(u) = \begin{cases} u & \text{se } u \geq 0 \\ \alpha u & \text{se } u < 0 \end{cases}$
<i>Exponential Linear Unit</i> (ELU)	$f(u) = \begin{cases} u & \text{se } u \geq 0 \\ \alpha(e^u - 1) & \text{se } u < 0 \end{cases}$
<i>Softplus</i>	$f(u) = \log(1 + \exp(u))$

A função sigmoide logística e a função linear, tem como saída valores reais no intervalo entre 0 e 1. O parâmetro α permite que se obtenha diferentes inclinações para a função sigmoide. A função tangente hiperbólica é normalmente usada quando necessitamos que a saída da função esteja entre -1 e $+1$ (apesar de as outras funções mencionadas também poderem ser utilizadas nesses casos, através de normalizações). Já as funções de ativação tipo *ReLU*, *PReLU*, *ELU* e *Softplus* são recomendadas nos casos em que temos *redes neurais profundas* envolvidas. Além dessas, temos também funções ativadoras que combinam todas as saídas em uma estimativa de probabilidade para classificação, como por exemplo a *função Softmax* também conhecida como *exponencial normalizada*. A *Softmax* é muito utilizada em vários métodos de classificação multiclases, como a *regressão logística multinomial* (ou *regressão softmax*), análise de discriminantes lineares (LDA), naive bayes e ANNs. Em geral, a função *Softmax* é descrita por "scores" s_c dados por um vetor de parâmetros referentes a cada classe c envolvida, conforme a seguir:

$$P_c = \sigma(s_c(x)) = \frac{\exp(s_c(x))}{\sum_{j=1}^C \exp(s_j(x))}, \quad (4.2)$$

onde C é o número de classes, $s_c(x)$ é o vetor contendo as pontuações de cada classe c para a instância x , e P_c é a probabilidade estimada de x pertencer a classe c (Géron, 2017).

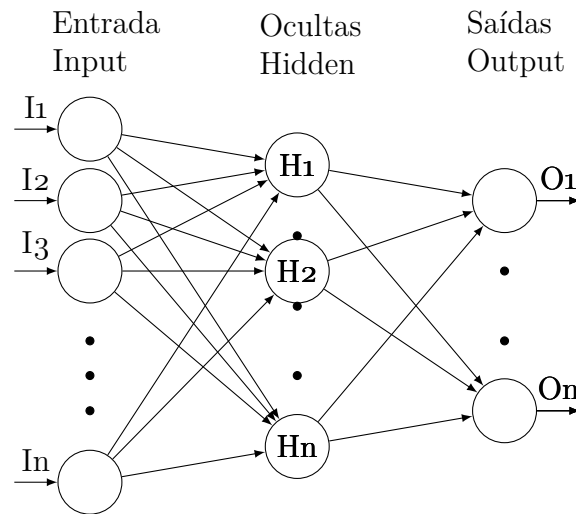


Figura 4 – Diagrama simplificado de uma rede neural MLP.

4.1 Perceptron, MLP e DNN

O Perceptron é uma das mais simples arquiteturas (topologias) de redes neurais artificiais, inicialmente publicada em (ROSENBLATT, 1958). Baseia-se em uma rede constituída com uma única camada de neurônios artificiais (tipo LTU), onde cada neurônio se conecta a todas as entradas associadas aos seus respectivos pesos. Na literatura o termo *Perceptron* também pode fazer referência a uma única unidade de neurônio (LTU) (ou neurônio perceptron). Nesses casos, o treinamento geralmente é realizado através da *regra Delta*, tornando essa rede, basicamente capaz de tratar problemas lineares e separáveis. No entanto, essa limitação pode ser eliminada em redes neurais formadas por múltiplas camadas de Perceptron (*Multi-Layer Perceptron -MLP*), as quais são capazes de resolver problemas mais complexos como do tipo *XOR*.

Nas *MLP*, observamos que geralmente não existem "sinapses" entre neurônios que estão na mesma camada e as conexões ocorrem apenas em um único sentido (*feed-forward*). Isto é, cada neurônio em uma camada direciona conexões para todos os neurônios da camada subsequente. O *teorema da aproximação universal* para redes neurais (HORNIK, 1991), afirma que toda função real contínua pode ser aproximada de forma arbitrária por *MLP* com uma única camada oculta, independentemente da função ativadora adotada. Uma *MLP* com duas ou mais camadas ocultas também é chamada de rede neural profunda (*deep neural network - DNN*).

As redes neurais tipo *MLP* utilizam uma variedade de técnicas de aprendizagem, sendo uma das mais populares o algoritmo de retropropagação *back-propagation* (que é intimamente relacionado ao algoritmo de Gauss-Newton). Basicamente essa técnica de aprendizagem compara os valores de saída da rede com a resposta correta contida no conjunto de

treinamento, calculando alguma função de custo (erro) predefinida, como por exemplo a *entropia cruzada*. O valor da função de custo é retornado pela rede e usando essa informação, o algoritmo ajusta os pesos de cada conexão para minimizar o valor do erro. Após repetir esse processo para um número suficiente de ciclos de treinamento, a rede geralmente converge para algum estado em que o erro é mínimo. Para isso, a rede calcula a derivada da função de erro com relação aos pesos da rede e altera os pesos de forma a reduzir o erro. Por essa razão, a retropropagação só pode ser aplicada em redes neurais com funções de ativação diferenciáveis (como qualquer otimizador baseado em gradiente). Existem várias topologias para as redes neurais, uma das primeiras publicações sobre o tema pode ser vista em (LECUN et al., 1989). Aqui, introduzimos alguns dos principais fundamentos de arquiteturas para ANNs.

4.2 Convolutional Neural Network (CNN)

Esse tipo de arquitetura é superior as *DNNs* em tarefas como classificação de imagens e sons. (GÉRON, 2017). As redes neurais convolucionais são semelhantes as *DNNs*, porém os neurônios em uma rede *CNN* geralmente são dispostos em 3 dimensões. Além disso, eles não são totalmente conectados, isto é, neurônios em camadas ocultas são conectados apenas com uma parte dos neurônios da camada posterior. Esse padrão de conectividade entre neurônios é inspirado na organização do cortex visual dos animais. Neurônios corticais individuais respondem a estímulos apenas em regiões restritas do campo de visão conhecidas como campos receptivos. Os campos receptivos de diferentes neurônios se sobrepõem parcialmente de forma a cobrir todo o campo de visão. Outra característica importante, é que geralmente as *CNNs* apresentam uma função de ativação do tipo *Softmax* na última camada (totalmente conectada), para a realização de tarefas de classificação, nos casos de regressão essa função de ativação normalmente não é aplicada. Na construção da arquitetura de uma *CNN* geralmente são utilizados três tipos de grupos de camadas conhecidas como *convolutional layer*, *pooling layer* e *fully connected layer* (CS231n - <<http://cs231n.github.io/convolutional-networks/>>).

4.2.0.1 Convolutional layers (CL)

A camada convolucional é o bloco fundamental de construção de qualquer CNN, para lidar com entradas em alta dimensão, ela conecta cada neurônio da camada seguinte a uma pequena parte do volume total de entradas na camada anterior. Simplificadamente as combinações de cada neurônio na camada seguinte, representam um "mosaico" da entrada original. Os hiperparâmetros da *CL* geralmente consistem em um conjunto de filtros (ou *kernels*) que podem ser descobertos. Quando uma *CNN* aprende um kernel que pode detectar um determinado recurso, ele pode detectar esse recurso em qualquer lugar da

imagem. Em contraste, quando uma *DNN* aprende um recurso em um local, ele pode detectá-lo apenas nesse local específico. Como as imagens normalmente têm recursos muito repetitivos, as *CNNs* são capazes de generalizar muito melhor do que as *DNNs* para tarefas de processamento de imagens, como classificação, usando menos exemplos de treinamento. Cada filtro é pequeno espacialmente (ao longo da largura e altura), mas se estende por toda a profundidade do volume da rede. Por exemplo, uma *CNN* pode ter tamanho $5 \times 5 \times 3$ (isto é, 5 pixels de largura e altura, e 3 para os canais de cores *RGB*).

Normalmente três hiperparâmetros controlam o tamanho do volume de saída: **profundidade**, **passo** e o **preenchimento com zeros**.

profundidade: corresponde ao número de filtros (ou kernels) que gostaríamos de utilizar, cada um aprendendo a procurar algo diferente nos dados de entrada.

passo: define como os filtros percorrem o plano da imagem original. Por exemplo, se o passo for 1 os filtros percorrem um pixel por vez, se for 2 (raramente 3 ou mais), então os filtros percorrem 2 pixels por vez, o que produz volumes de saídas menores espacialmente.

preenchimento com zeros (Dropout): preenche o volume de entrada com zeros em regiões onde as informações são insignificantes. Basicamente ele permite controlar o tamanho do volume de saída, podendo ser usado para preservar a mesma largura e altura do volume de entrada.

A camada convolucional calcula cada fatia do volume como uma convolução entre os pesos de cada neurônio e o volume de entrada, por isso normalmente nos referimos aos conjuntos de pesos como filtros (ou um kernel) que são convolvido com a entrada.

4.2.0.2 Pooling layers (PL)

As camadas de agrupamento (PL), são responsáveis por reduzir a resolução das dimensões espaciais da entrada. configurando o hiperparâmetro passo para 2 ou 3 (e mais raramente, valores de pool maiores que 3). A forma mais comum de arquitetura CNN empilha camadas *CLs* intercaladas com *PLs*. A função de uma camada de agrupamento é reduzir progressivamente o tamanho espacial da representação para reduzir a quantidade de parâmetros calculados pela rede e controlar o overfitting (*CS231n*). Geralmente a redução é realizada através de uma função *max pooling*, mas as pooling layers também podem executar outras funções como *average pooling* ou mesmo *L2-norm* (para um exemplo de *pooling* com *Wavelet*, veja (WILLIAMS; LI, 2018)).

Algumas publicações questionam a eficiência das operações em camadas pooling layers. Por exemplo, em (SPRINGENBERG et al., 2014) é proposto o descarte das camadas de agrupamento (PL) em prol de uma arquitetura apenas com camadas convolucionais empilhadas, para reduzir o tamanho da representação, eles sugerem utilizar um valor de *passo* maior em cada camada *CL*, de vez em quando. O descarte de camadas de agrupamento

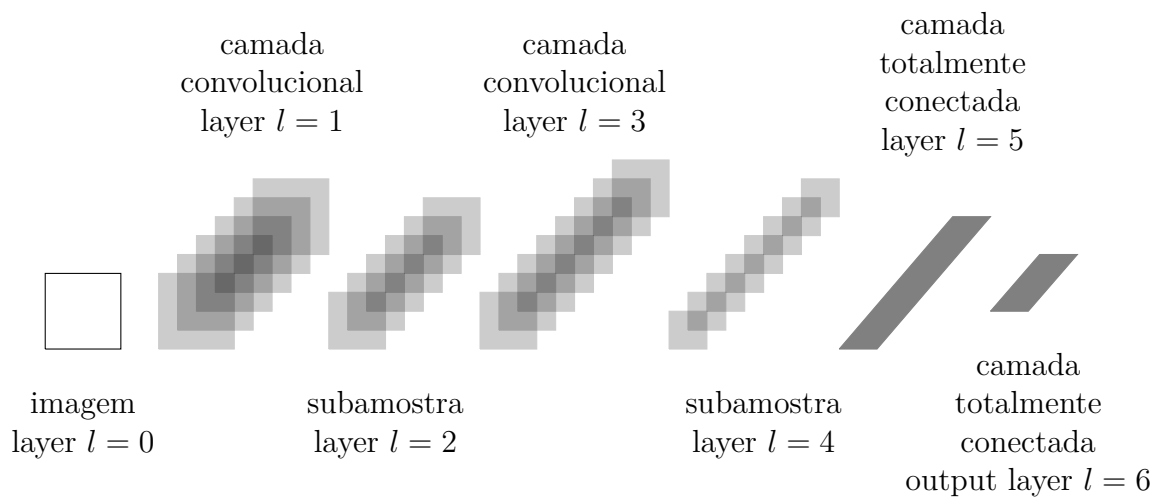


Figura 5 – Arquitetura de uma rede neural convolucional, como originalmente proposta por LeCun (LECUN et al., 1989), alterna entre camadas convolucionais, incluindo não linearidades hiperbólicas e camadas de subamostragem. Na figura, as camadas convolucionais já incluem não-linearidades e, assim, uma camada convolucional na verdade representa duas camadas. Nos casos de classificação, normalmente utiliza-se uma função softmax na camada de saída.

também foi considerado importante no treinamento de bons modelos geradores, como autoencodificadores variacionais (VAEs) ou redes geradoras de adversários (GANs). Parece provável que as futuras arquiteturas apresentem muito poucas ou nenhuma camada de agrupamento (CS231n).

4.2.0.3 Fully connected layers (FC)

Neurônios em uma camada totalmente conectada têm conexões completas com todos os neurônios da camada anterior, assim como visto nas DNNs. Elas são comumente empregadas por classificadores baseado na função de ativação *Softmax*, minimizando a entropia cruzada entre distribuições estimadas. Camadas *FC* podem ser convertidas em *CL* e vice-versa, para mais detalhes e exemplos de aplicações práticas, veja CS231n.

Uma rede neural convolucional é geralmente complexa, e começar do zero sua construção nem sempre é viável. Por isso, muitas vezes a melhor opção é modificar alguma arquitetura já existente, como por exemplo, *LeNet*, *AlexNet*, *ZF Net*, *GooLeNet*, *VGGNet*, *ResNet*, entre outros (veja CS231n). A *CNN* também pode ser combinada com uma *RNN* como mencionamos a seguir.

4.3 Recurrent Neural Network (RNN)

RNN é uma classe de redes neurais que inclui conexões ponderadas dentro de uma mesma camada (em contraste com as redes de feed-forward tradicionais, que conectam alimentação

apenas para camadas subsequentes). Como as *RNNs* incluem loops, elas podem armazenar informações ao processar novas entradas. Essa memória as tornam ideais para tarefas de processamento onde as entradas anteriores devem ser consideradas (como por exemplo, dados de séries temporais).

As Redes neurais recorrentes não são uma única arquitetura de rede, mas sim, uma coleção de topologias que podem ser aplicadas a uma variedade de problemas. uma das primeiras e mais simples arquiteturas com natureza recorrente, são as redes de *Hopfield* (HOPFIELD, 1982). Outros tipos mais sofisticados e eficientes de *RNNs* têm sido propostos, incluindo a *long-short term memory (LSTM)* (HOCHREITER; SCHMIDHUBER, 1997) e *gated recurrent units (GRU)*, que é uma otimização da *LSTM* (CHUNG et al., 2014). O formato *LSTM* revolucionou o desenvolvimento nos reconhecimentos de sons e imagens, foi utilizado no *Google Android* e na pesquisa de voz do Google. Além disso, a combinação de *LSTM* e *CNN* otimizou o reconhecimento de legendas automáticas e imagens (VINYALS et al., 2015). Em (CHARNOCK; MOSS, 2017) são aplicadas redes *LSTM* unidirecionais e bidirecionais a classificação fotométrica de supernovas. As *RNNs* bidirecionais usam um conjunto de treinamento (por exemplo, sinais temporais) para prever ou rotular cada elemento com base em seu comportamento passado e futuro. Isso é feito concatenando as saídas de duas *RNNs*, uma processando a sequência da esquerda para a direita e outra da direita para a esquerda. Os resultados combinados são as previsões dos sinais.

4.4 Processos de aprendizagem com ANNs

Como vimos até então, a aprendizagem de uma *ANN* acontece a partir de processos interativos nos quais os parâmetros livres da rede são ajustados. O tipo de aprendizagem é determinado pela maneira com que estes parâmetros são ajustados a cada interação. Basicamente, as diferenças entre os tipos de aprendizagem dizem respeito a maneira como é feito o ajuste do peso sináptico de um neurônio. Essas diferenças, resultam em vários algoritmos para o treinamento (regras de aprendizagem), que podem ser agrupados em dois paradigmas principais: *aprendizado supervisionado* e *aprendizado não supervisionado* (BRAGA; CARVALHO; LUDERMIR, 2014). Mencionaremos agora algumas das mais conhecidas regras de aprendizagem.

4.4.1 Regra Delta

Essa regra de aprendizagem foi desenvolvida por Widrow e Hoff (WIDROW; HOFF, 1960), também conhecida como método *least mean square* (LMS). Para um dado vetor de entrada, o vetor de saída é comparado com a resposta correta, se a diferença é 0 não ocorre aprendizado caso contrário temos um ajuste de "pesos". Por exemplo, podemos supor que conhecemos as saídas desejadas para cada neurônio k da camada de saída de uma rede

neural, de forma que o erro ($\epsilon_k = \hat{y}_k - y_k$) possa ser calculado através da diferença entre o valor esperado ($\hat{y}_k$) e a saída do neurônio (y_k). A correção do erro é realizada através de atualizações a cada interação. Se considerarmos n como um dos passos do processo iterativo e a atualização do peso (w_{ki}) de um neurônio k , cuja entrada i resulta no erro (ϵ_k), podemos escrever:

$$w_{ki}(n+1) := w_{ki}(n) + \Delta w_{ki}(n), \quad (4.3)$$

onde a variação do peso Δw_{ki} é dada por:

$$\Delta w_{ki} = \eta x_i \epsilon_k. \quad (4.4)$$

x_i é o i -ésimo dado de entrada e η é conhecido como *taxa de aprendizagem* que tem valor real positivo no intervalo entre 0 e 1. Essa regra é derivada da minimização do erro na saída de uma rede neural através do *gradiente descendente*. Por exemplo, o erro total de uma rede neural artificial com k saídas pode ser definido como:

$$E = \sum_j \frac{1}{2} (\hat{y}^j - y^j)^2 \quad (4.5)$$

Assim desejamos analisar todos os valores possíveis de pesos para cada neurônio da rede, proporcionalmente a função erro *gradiente descendente*. Isso é feito calculando a derivada parcial do erro em relação a cada peso ($\partial E / \partial w_{ki}$). O resultado que obtemos é exatamente a equação 4.4.

A constante η tem grande influência na estabilidade e convergência do processo de aprendizagem iterativo. Essa regra de aprendizagem é utilizada nas redes neurais artificiais *Perceptron* e também nas *Perceptron Multicamadas*. O algoritmo *back-propagation* surgiu como uma generalização da regra delta para redes neurais multicamadas (RUMELHART; HINTON; WILLIAMS, 1986).

4.4.2 Aprendizagem com memória

Na aprendizagem baseada em memória um amplo conjunto de treinamento deve ser armazenado explicitamente em uma grande memória que é representada por um conjunto de pares ordenados $(\mathbf{x}_i, y_i)$ formado pelo vetor de parâmetros de cada entrada ($\mathbf{x}_i$) e as suas respectivas respostas (y_i). Assim, por exemplo, a classificação de um novo vetor de entrada pode ser feita através de uma pesquisa em torno de sua vizinhança local, com base nos dados já "memorizados".

4.4.3 Regra de Hebb

A aprendizagem hebbiana é baseada na força da sinapse em função da sincronia de neurônios que estão em ambos os lados dessa sinapse. Essa aprendizagem é chamada hebbiana em homenagem a seu idealizador Donald Hebb (HEBB, 1949) que propôs um postulado inspirado na base de aprendizagem associativa (ANDERSON, 1972; KOHONEN, 1972). Esse postulado foi proposto no contexto neurobiológico, mas sua aplicação em redes neurais artificiais pode ser descrito como uma regra em duas partes:

1. A "força" da sinapse aumenta se os dois neurônios que estão envolvidos são ativados sincronamente.
2. A "força" da sinapse reduz se os dois neurônios envolvidos são ativados assincronamente.

Basicamente, a forma mais simples de aprendizagem hebbiana pode ser definida por:

$$\Delta w_{ki}(n) = \eta y_k(n) x_i(n), \quad (4.6)$$

onde n representa cada interação. Essa regra foi inicialmente aplicada em (HOPFIELD, 1982). No entanto essa regra apresenta grande instabilidade e por isso os modelos de redes neurais artificiais normalmente empregam uma modificação dessa regra, conhecida como *algoritmo hebbiano generalizado* (GHA) ou *regra de Sanger* e definido por:

$$\Delta w_{ki} = \eta \left(\hat{y}_k x_i - \hat{y}_k \sum_k w_{ki} y_k \right) \quad (4.7)$$

onde w_{ki} define o peso sináptico ou "força" da conexão entre a i -ésima entrada e o k -ésimo neurônio de saída. Assim o aprendizado é um processo de única camada, isto é, um peso sináptico muda apenas dependendo da resposta das entradas e saídas para a camada de neurônios, dispensando o uso de redes neurais multicamadas. Essa regra de aprendizagem é muito utilizada em aplicações onde é feita a análise de componentes principais.

4.4.4 Aprendizagem competitiva

Na regra de aprendizagem competitiva os neurônios competem entre si e o vencedor ganha o direito de disparar enquanto os outros permanecem inertes. Basicamente ela é formada em redes neurais com camadas ocultas (também chamadas de camadas competitivas), onde cada neurônio responde de forma diferente aos padrões de entrada e apenas o neurônio que apresentar o melhor resultado (medido através de alguma função de similaridade) é disparado. Assim, individualmente os neurônios da rede podem aprender a se especializar em conjuntos de padrões semelhantes e tornam-se potenciais detectores de recursos ou características principais.

4.4.5 Aprendizagem de Boltzmann

Esse método de aprendizagem baseia-se na *Boltzmann Machine* desenvolvida em 1985 por Geoffrey Hinton e Terrence Sejnowski ([Géron, 2017](#)). Assim como as redes Hopfield ([Hopfield, 1982](#)) elas são *ANNs* totalmente conectadas e baseadas em *neurônios estocásticos*, ao invés de utilizarem uma função de passo (ou ativação) determinística para decidir o valor de saída. Esses neurônios apresentam uma probabilidade entre 0 e 1, dentre outras formas. A função de probabilidade que tais *ANNs* utilizam, são baseadas na distribuição de Boltzmann, para mais detalhes, ([Géron, 2017](#)).

5 Resultados

Aqui descrevemos as modificações no código de (LOCHNER et al., 2016) e configurações feitas para as simulações com o SNANA (KESSLER, 2018). Ao final apresentamos alguns dos resultados gráficos preliminares obtidos para o conjunto simulado de $\sim 2 \times 10^4$ supernovas.

5.1 Modificações em snmachine e configuração do SNANA

Como já mencionado anteriormente, os dados simulados foram obtidos através do software SNANA, tanto para a pesquisa padrão (*Wide-Fast-Deep WFD*) quanto profunda (*Deep Drilling Fields - DDF*). Geramos 20.860 simulações fotométricas de supernovas entre os tipos Ia, II, IIP, IIL, Ib, Ic e Iib. Dessas, 1.732 são do tipo Ia. Utilizamos o padrão fornecido por R. Biswas, D. Cinabro e R. Kessler <<https://github.com/rbiswas4/OpSimSummary>>, considerando os dois arquivos SIMLIB: LSST_DDF_COADD.SIMLIB e LSST_WFD_COADD.SIMLIB. Estes arquivos públicos foram gerados utilizando o código OpSim (baseado no padrão de simulação de operações do LSST - OpSim <www.lsst.org/scientists/simulations/opsim>) e refletem a atual cadência esperada para a pesquisa, podendo ainda ocorrerem modificações futuras. Na geração desses arquivos SIMLIB, as noites de observação são estimadas utilizando o código SNANA `simlib_coadd.exe`, o céu e o PsF (*Progressive segmented Frame*) são calculados utilizando 'filtSkyBrightness' e 'FWHMeff'. A extinção devido a nossa galáxia é simulada utilizando a fração de incerteza em $E(B - V)$ através da escolha apropriada da chave de entrada para simulação OPT_MWEBV (KESSLER, 2018). A correção-K é feita através do arquivo em `$SNDATA_ROOT/kcor/LSST/2017-04_approx/kcor_LSST.fits`.

Outra característica importante de nossas simulações é que enquanto modelos de curvas de luz para supernovas tipo Ia são baseadas em alguma equação paramétrica ou modelo fotométrico, as simulações de supernovas não-Ia são baseadas em suavização de curvas de luz e modelos espectrais. O modelo espectral corresponde a uma particular supernova não-Ia (bem observada), onde o "espectro composto" é deformado para coincidir com a fotometria observada. Em (PIEREL et al., 2018) é proposto um método para a extensão do modelo SALT2 nas bandas UV e IR , evidenciando a importância desse estudo, para as futuras pesquisas sinópticas, como o LSST. Além disso, SN não-Ia apresentam características que podem ser utilizadas para simular peculiares Ia, modelos teóricos, AGN ou qualquer outro transiente (KESSLER, 2018). A lista de modelos avaliados de *NON1ASED* é fornecida em `$SNDATA_ROOT/snsed/NON1A/NON1A.LIST`.

O tipo Ia foi simulado utilizando o modelo *SALT2.JLA - B14*, enquanto os tipos não-Ia,

com o *NON1ASED*. O modelo *NON1ASED* assume que para cada quadro de descanso (rest frame) a distribuição espectral de energia (*SED*) deve ser deformada para corresponder à fotometria da curva de luz subjacente. Além disso, utilizamos `INPUT_FILE_INCLUDE: $SNDATA_ROOT/snsed/NON1A/SIMGEN_INCLUDE_NON1A_J17-beforeAdjust.INPUT`, para mais informações sobre a simulação de supernovas tipo não-Ia, consulte (KESSLER, 2018).

O SNANA é um software para simulações cosmológicas muito abrangente, nesse trabalho preliminar desenvolvemos as simulações apenas baseando-se em ajustes dos arquivos *.INPUT* fornecidos para o LSST no diretório de instalação `$SNDATA_ROOT/sample_input_files/LSST`. No futuro, pretendemos ser capazes de gerar dados simulados atualizados, a medida que o software e as informações sobre a pesquisa forem disponibilizados. Além disso, pretendemos ser capazes de realizar a identificação de qualquer transiente através da curva de luz, independente de ser uma SN ou não.

O padrão SNANA utiliza como formato para os dados de saída *.FITS* (`FORMAT_MASK : 32`). Nós configuramos para o formato *.DAT* através da opção `FORMAT_MASK : 2` dentro dos arquivos de entrada *SIMGEN_SALT2.INPUT* e *SIMGEN_NON1A.INPUT*, para o executável *snlc_sim.exe*. A escolha do formato *.DAT* para os arquivos de output foi motivada pelo código *snmachine* (LOCHNER et al., 2016), que apesar de apresentar uma exclusiva classe para a leitura de simulações no formato *.FITS* (`OpSimDataset - sndata.py`), foi originalmente desenvolvido para análise com o conjunto de dados simulados do SPCC (KESSLER et al., 2010b) (21.319 supernovas simuladas no formato *.DAT*).

O código *snmachine* foi desenvolvido através de um conjunto de sub-rotinas construídas em python: *sndata.py* (contém as rotinas para leitura dos dados de entrada), *snfeatures.py* (rotinas para extração de características), *parametric_models.py* (rotinas para os modelos de Newling (NEWLING et al., 2010) e Karpenka (KARPENKA; FERROZ; HOBSON, 2012)), *tsne_plot.py* (responsável pela visualização prévia da separação entre tipos de supernovas após a etapa de extração de características) e o *snclassifier.py* (rotinas para os métodos de classificação). O *snmachine* depende de diferentes bibliotecas públicas e dentre elas, destacamos: *numpy*, *cython*, *scipy*, *matplotlib*, *astropy*, *sncosmo*, *scikit-learn*, *pymultinest*, *emcee*, *iminuit* e *pyWavelets*.

Algumas modificações foram realizadas nas sub-rotinas *sndata.py*, *snfeatures.py*, *tsne_plot.py* e *classifier.py*, com o intuito de uma análise similar a que foi feita em (LOCHNER et al., 2016), utilizando nossos dados simulados. Em *sndata.py* alteramos as variáveis *sntypes*, *markers* e *labels*, também modificamos o padrão de filtros na classe *Dataset* para o padrão LSST (TYSON; STRAUSS, 2009). Em *Dataset.get_lightcurve*, mudamos a referência *'HOST_GALAXY_PHOTO - Z :'* para *'REDSHIFT_HELIO'* : (isto significa, que estamos considerando o redshift heliocêntrico e não o redshift fotométrico da galáxia hospedeira, como referência). No SNANA, a simulação do redshift da galáxia hospedeira

Priors on the SALT2 Model Parameters	
Parameter Name	Prior
z	$\mathcal{U}(0.01, 1.5)$
t_0	$\mathcal{U}(-100, 100)$
x_0	$\mathcal{U}(-10^{-3}, 10^{-3})$
x_1	$\mathcal{U}(-3, 3)$
c	$\mathcal{U}(-0.5, 0.5)$

Priors on the Model 1 Parameters	
Parameter Name	Prior
$\log(A)$	$\mathcal{U}(0, 10)$
ϕ	$\mathcal{U}(-60, 100)$
$\log(\sigma)$	$\mathcal{U}(-3, 4)$
$\log(k)$	$\mathcal{U}(-4, 4)$
$\log(\psi)$	$\mathcal{U}(-6, 10)$

Priors on the Model 2 Parameters	
Parameter Name	Prior
$\log(A)$	$\mathcal{U}(\log(10^{-5}), \log(1000))$
$\log(B)$	$\mathcal{U}(\log(10^{-5}), \log(100))$
t_0	$\mathcal{U}(0, 100)$
t_1	$\mathcal{U}(0, 100)$
T_{rise}	$\mathcal{U}(0, 100)$
T_{fall}	$\mathcal{U}(0, 100)$

Figura 6 – Priors para *SALT2* e modelos paramétricos. Os mesmos que foram utilizados em (LOCHNER et al., 2016).

de cada *SN* é feita através de um arquivo .HOSTLIB (KESSLER, 2018).

No arquivo *snfeatures.py* ativamos a função *registerBands*, com a biblioteca *sncosmo* na classe *TemplateFeatures()*, para um *lsst_dir* que contém os arquivos de bandas de passagem obtidos de <github.com/lsst/throughputs/>. A classe *TemplateFeatures()* fornece a opção de escolha da utilização de diferentes modelos (*mlcs2k2*, *salt2-extended*, *nugent-sn2n*, *nugent-sn2l*, *nugent-sn2p* e *nugent-sn1bc*) no ajuste da curva de luz subjacente aos dados, porém, assim como feito em (LOCHNER et al., 2016), apenas testamos com o *SALT2*. Também podemos optar entre três diferentes métodos: *leastsq* (mínimos quadrados), *emcee* (*python ensemble sampling toolkit for affine-invariant MCMC*) e *multinest* (*efficient and robust bayesian inference*) que se baseia no algoritmo de amostragem aninhada (*nested sampling*) que é uma abordagem computacional para o problema de comparação entre modelos em estatística bayesiana (SKILLING, 2004a; SKILLING, 2004b; SIVIA; SKILLING, 2006).

Uma outra importante modificação que foi feita no código original *snfeatures.py* é a execução de processos gaussianos com a biblioteca *python george* para processos gaussianos de regressão <dfm.io/george>. Em *tsne_plot.py* incluímos os tipos de supernovas IIB, IIP, IIL, Ib, Ic e II (o tipo II compreende supernovas com características do tipo II, mas que não se enquadram em nenhum dos subtipos anteriores). Além disso, na validação cruzada foi alterado o parâmetro *cv* em *grid_search.GridSearchCV* de 5 para 10. O *grid_search* é um módulo do pacote *scikit-learn* para buscas exaustivas por específicos valores de parâmetros para um estimador.

Na figura 6, fornecemos os priors utilizados na etapa de extração de característica. A seguir, fornecemos os resultados preliminares obtidos em nosso estudo.

5.2 Resultados gráficos

Nesta seção apresentamos inicialmente os t-SNE plots para cada um dos métodos de extração de características, em seguida, as curvas ROC para avaliação da classificação realizada através dos referidos métodos combinados com aprendizagens de máquina e redes neurais artificiais (ANN). Por último fornecemos alguns exemplos de ajustes de curvas obtidos.

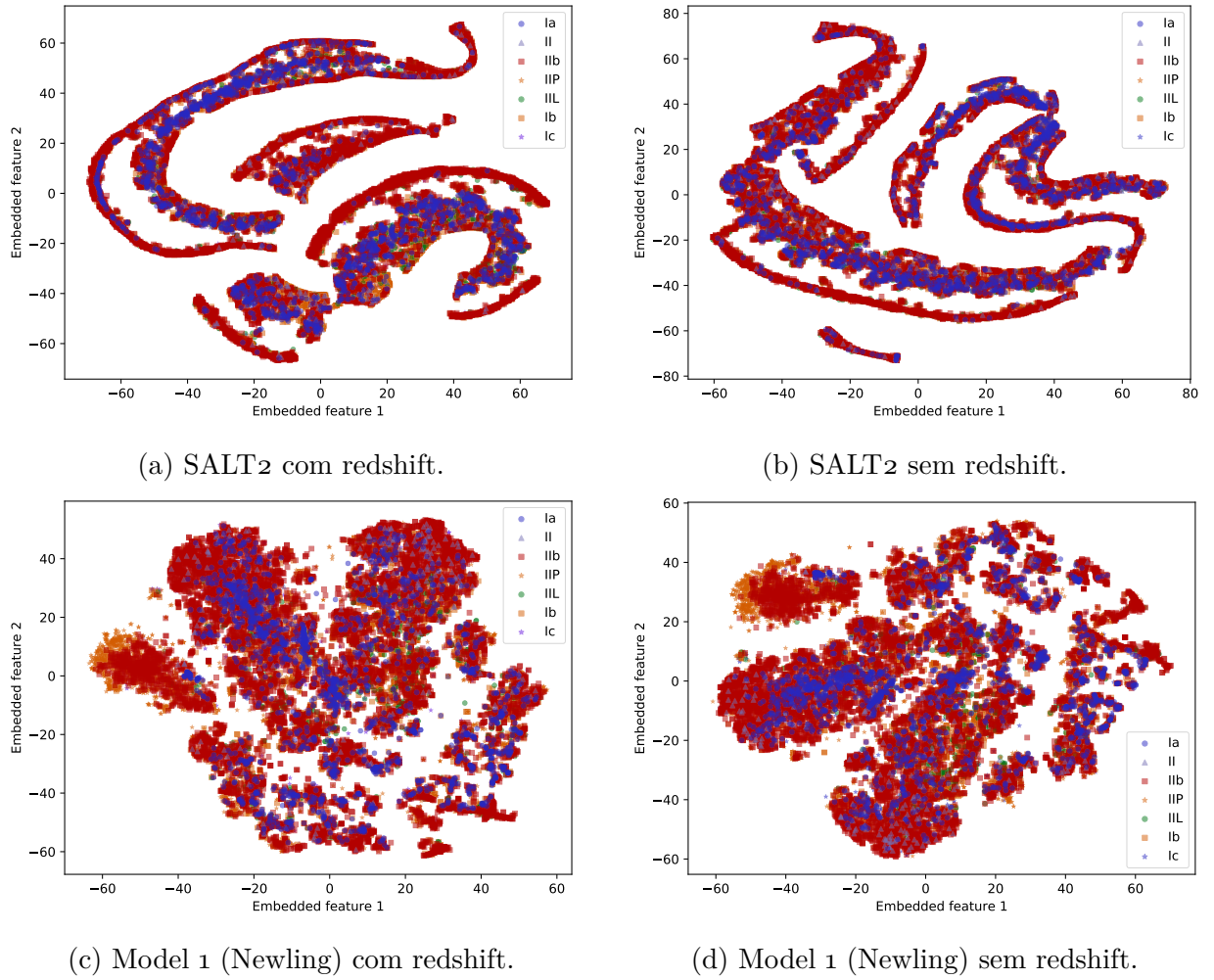
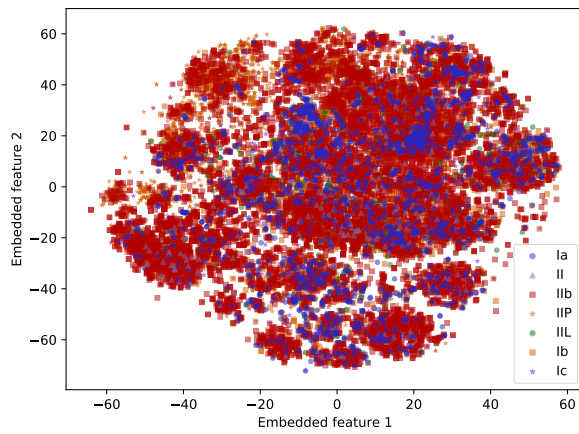


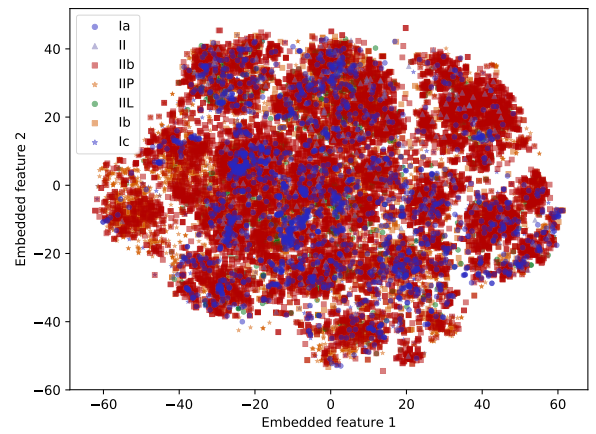
Figura 7 – Visualizações t-SNE para os métodos de extração de características, com e sem a informação do redshift. SALT2 e Modelo1.

	<i>SALT2</i>	Modelo 1	Modelo 2	Wavelets
<i>NB</i>	...	...	...	...
<i>KNN</i>	$n_neighbors = 141$	$n_neighbors = 101$	$n_neighbors = 46$	$n_neighbors = 121$
<i>SVM</i>	$C = 0.56, \gamma = 1.78$	$C = 1780, \gamma = 0.0032$	$C = 31.6, \gamma = 0.0032$	$C = 0.56, \gamma = 1.78$
<i>RF</i>	$n_estimators = 400$	$n_estimators = 600$	$n_estimators = 600$	$n_estimators = 600$
<i>BDT</i>	$n_estimators = 40$	$n_estimators = 75$	$n_estimators = 65$	$n_estimators = 70$
<i>ANN</i>	$neurons = 110$	$neurons = 110$	$neurons = 115$	$neurons = 80$

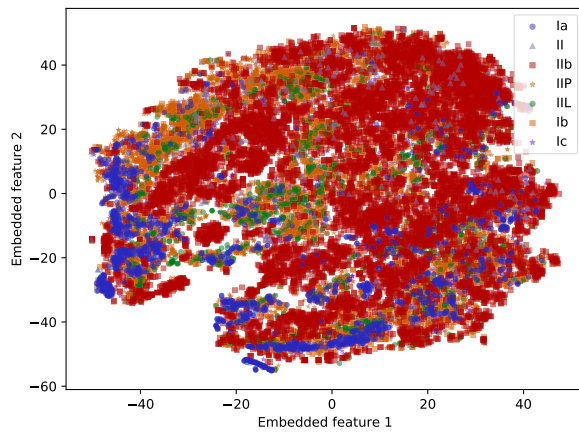
Tabela 1 – Hiperparâmetros para cada algoritmo de extração de característica e aprendizagem de máquina, respectivamente.



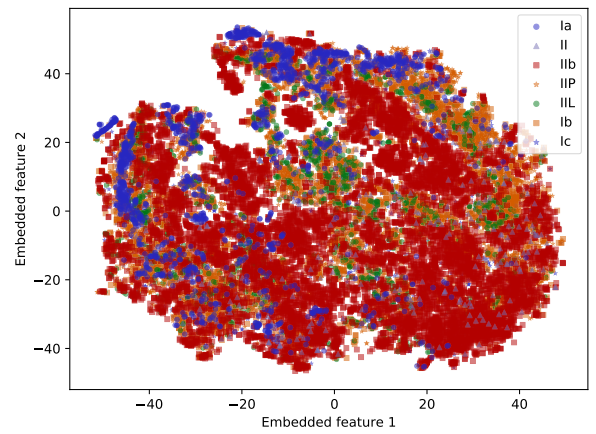
(a) Model2 (Karpenka) com redshift.



(b) Model2 (Karpenka) sem redshift.

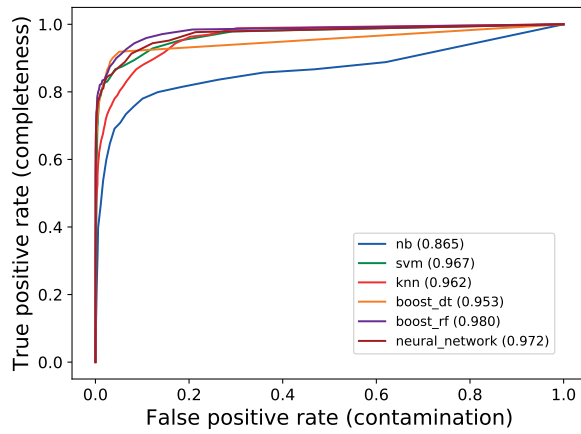


(c) Wavelet com redshift.

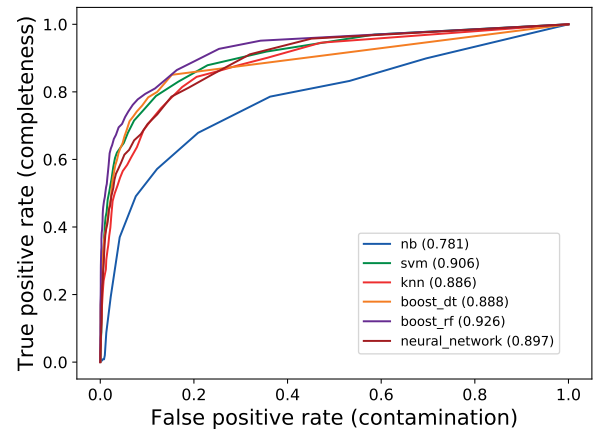


(d) Wavelet sem redshift.

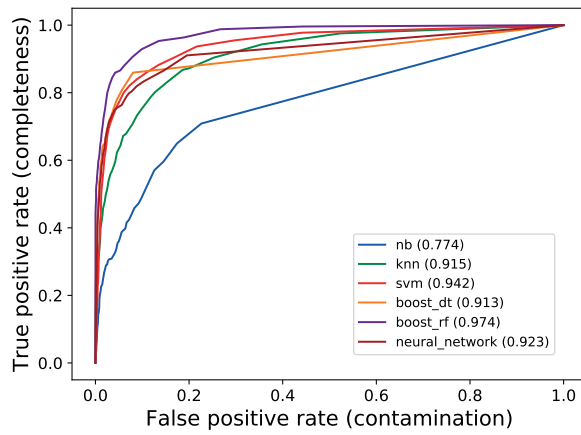
Figura 8 – Visualizações t-SNE para os métodos de extração de características, com e sem a informação do redshift. Modelo2 e Wavelet.



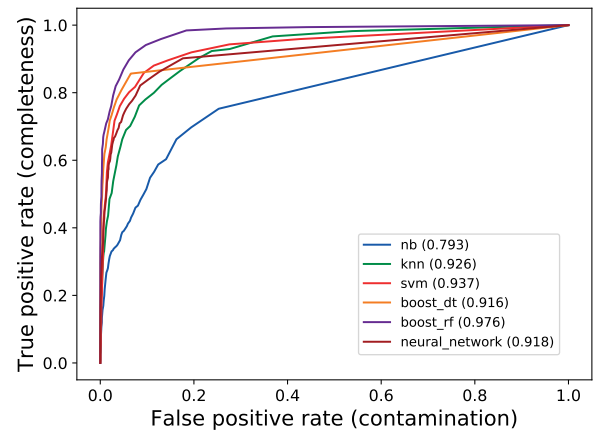
(a) SALT2 com redshift.



(b) SALT2 sem redshift.

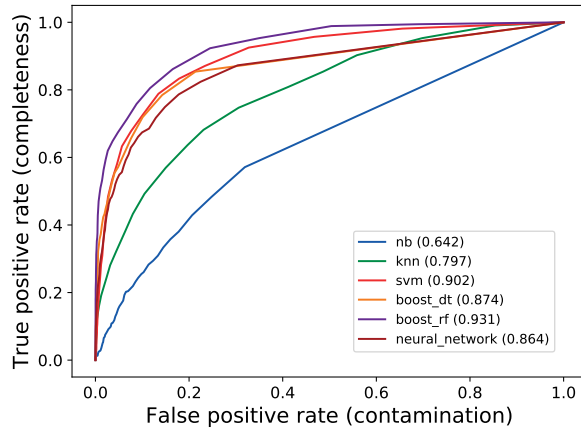


(c) Model1 (Newling) com redshift.

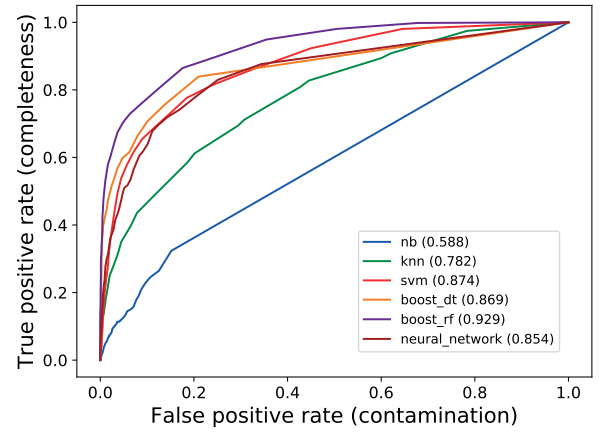


(d) Model1 (Newling) sem redshift.

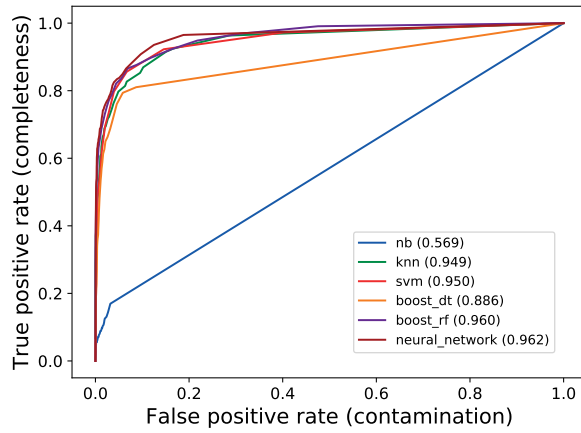
Figura 9 – Curva ROC com e sem a informação do redshift para SALT2 e Modelo1. Cada curva representa um algoritmo de aprendizagem de máquina diferente, com sua respectiva AUC informada na legenda.



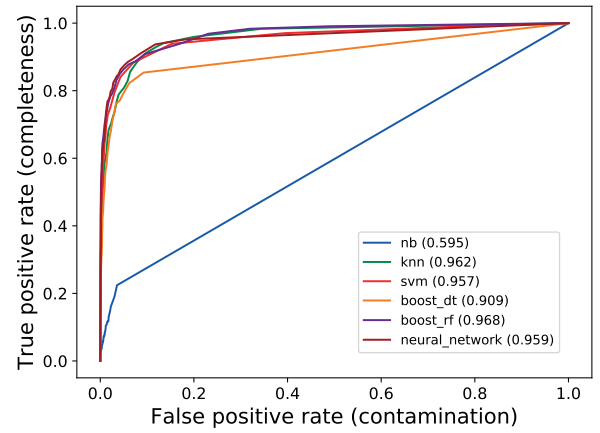
(a) Modelo2 (Karpenka) com redshift.



(b) Modelo2 (Karpenka) sem redshift.

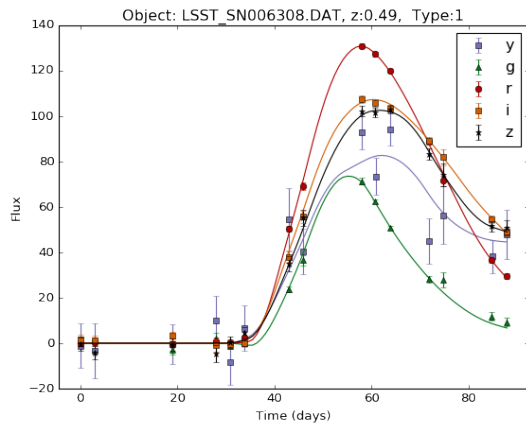


(c) Wavelet com redshift.

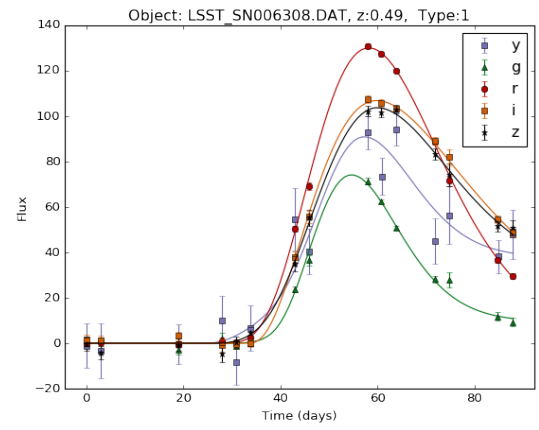


(d) Wavelet sem redshift.

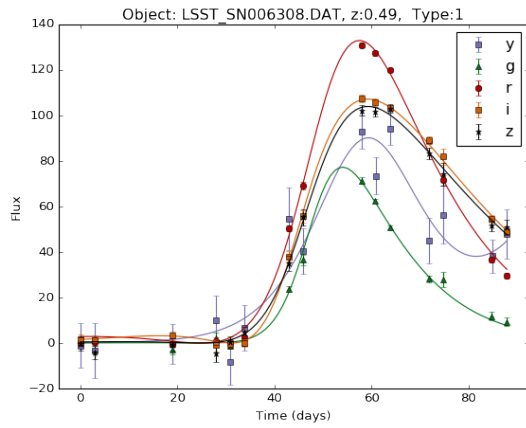
Figura 10 – Curva ROC com e sem a informação do redshift para Modelo2 e Wavelet. Cada curva representa um algoritmo de aprendizagem de máquina diferente, com sua respectiva AUC informada na legenda.



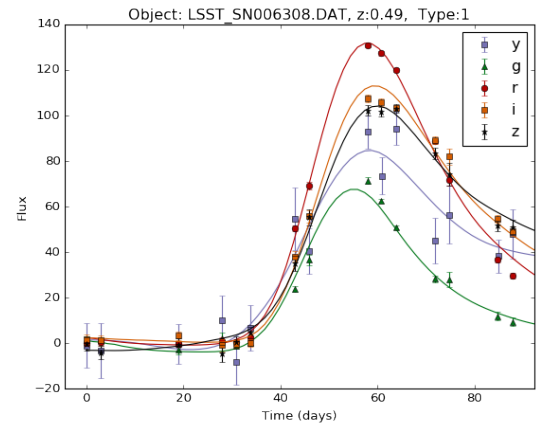
(a) SALT2



(b) Newling.

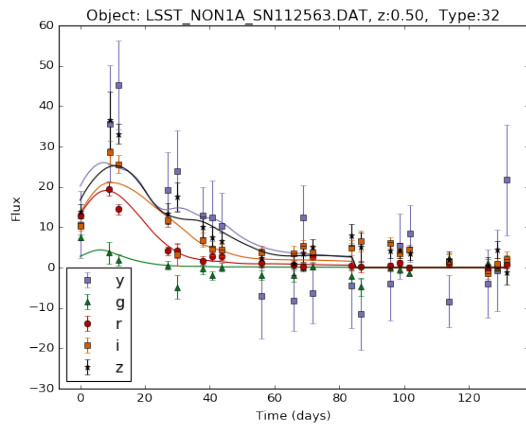


(c) Karpenka.

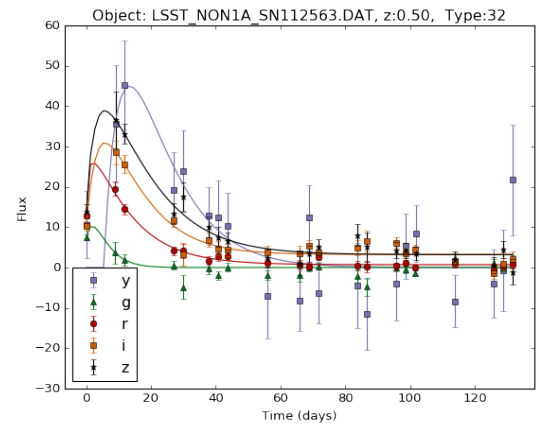


(d) Wavelet.

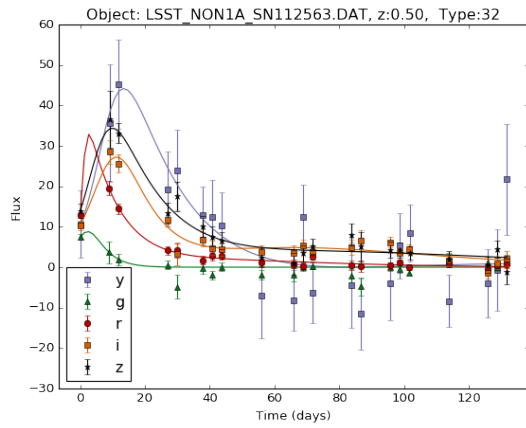
Figura 11 – Exemplos de ajustes para curvas de luz obtidas através dos métodos de extração de característica, supernova tipo Ia.



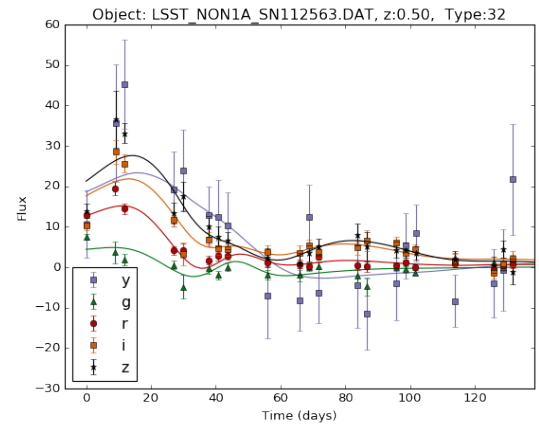
(a) SALT2



(b) Newling.

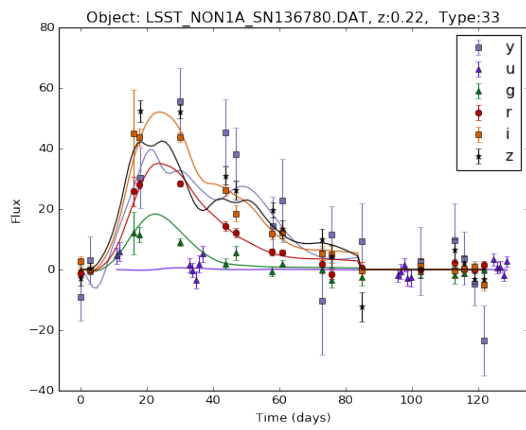


(c) Karpenka.

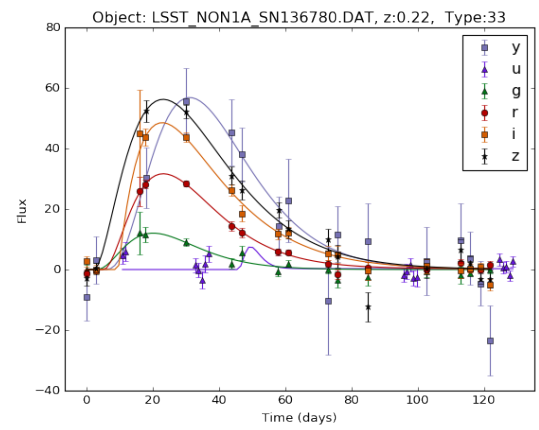


(d) Wavelet.

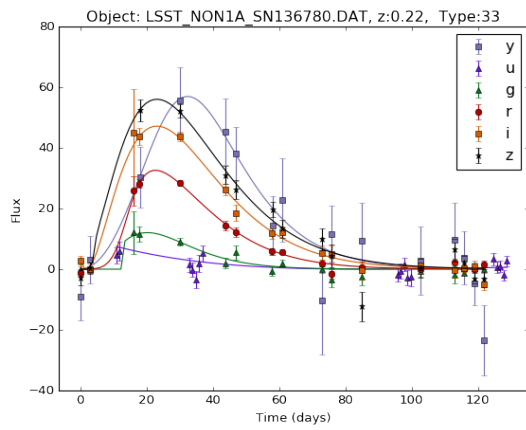
Figura 12 – Exemplos de ajustes para curvas de luz obtidas através dos métodos de extração de característica, supernova tipo Ib.



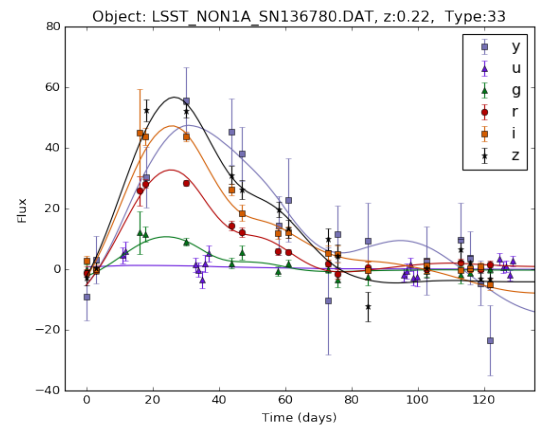
(a) SALT2



(b) Newling.

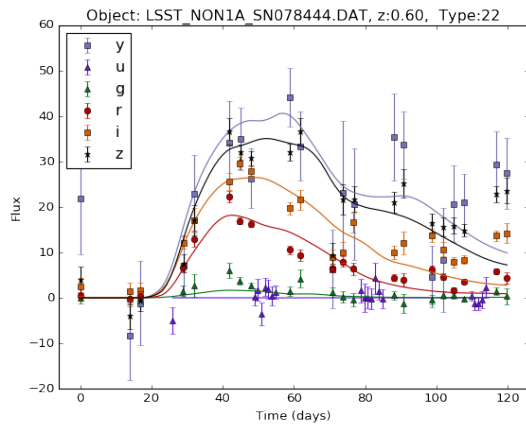


(c) Karpenka.

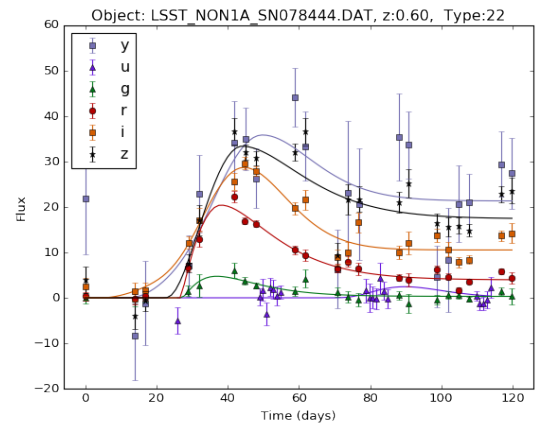


(d) Wavelet.

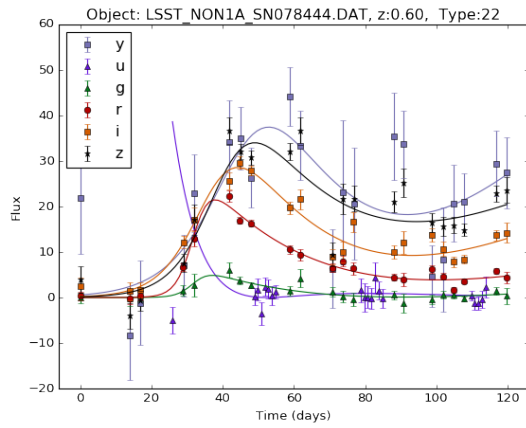
Figura 13 – Exemplos de ajustes para curvas de luz obtidas através dos métodos de extração de característica, supernova tipo Ic.



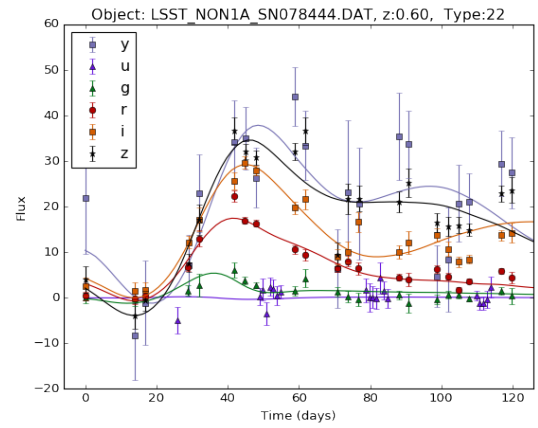
(a) SALT2



(b) Newling.

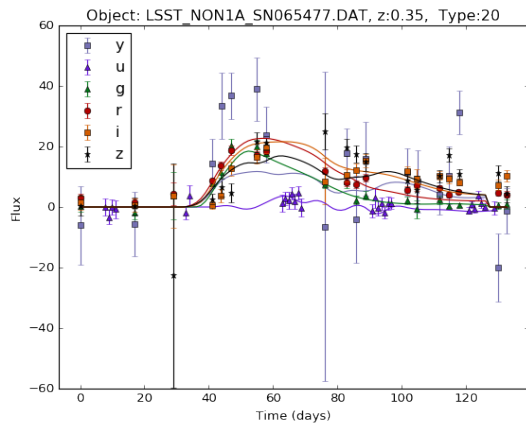


(c) Karpenka.

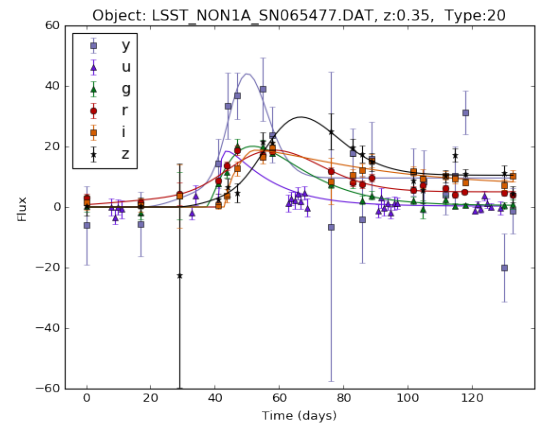


(d) Wavelet.

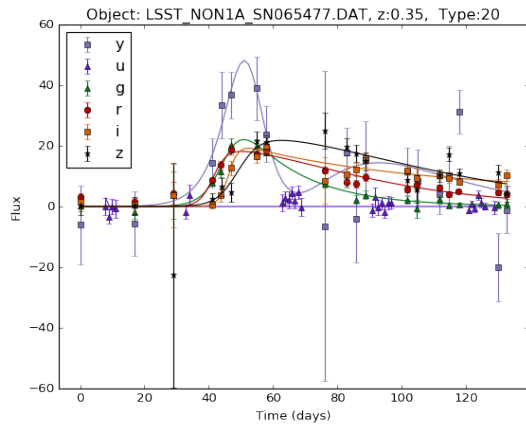
Figura 14 – Exemplos de ajustes para curvas de luz obtidas através dos métodos de extração de característica, supernova tipo IIL.



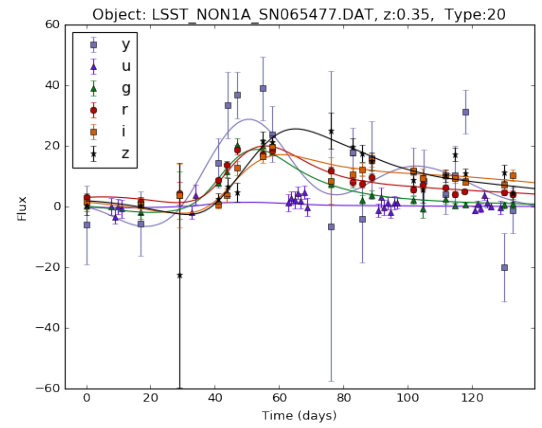
(a) SALT2



(b) Newling.



(c) Karpenka.



(d) Wavelet.

Figura 15 – Exemplos de ajustes para curvas de luz obtidas através dos métodos de extração de característica, supernova tipo IIP.

6 Conclusão

Apesar das visualizações em t-SNE, terem sido feitas diferenciando todos tipos de supernovas simulados (visualizados através de diferentes cores e formas de kernels nas imagens), em nossa pesquisa tratamos apenas da classificação binária, entre supernovas Ia e não-Ia. Verificamos, que mesmo assim o método não supervisionado de redução de dimensionalidade, t-SNE, é capaz de encontrar características correlacionadas entre os subtipos de supernovas gerais. Isso pode ser melhor evidenciado, através da separação de cores nas figuras 8(c) e 8(d) (t-SNE Wavelet).

Destacamos esse resultado, apesar de não ter sido o melhor obtido para os métodos extração de características. Isso porque o método SALT2 é tendencioso, já que o SNANA realiza a simulação das SN Ia com base no mesmo modelo. Os modelos de Newling e Karpenka são paramétricos e capazes de realizar aproximações, porém, não são capazes de descreverem a real variabilidade entre tipos e subtipos de supernovas. Desta forma, evidenciamos que o método Wavelet combinado com o plot t-SNE, é uma poderosa ferramenta para a extração de características. Também avaliamos a combinação com diferentes aprendizagens de máquinas para a classificação. Através das comparações entre os valores obtidos para a área sob a curva ROC (AUC), podemos eleger o método *Boosted Random Forest* (boost_rf) como o que apresentou melhores resultados para a classificação independentemente do método de extração de características utilizado. Observamos que esse resultado ainda deverá ser reavaliado utilizando hiperparâmetros diferentes dos fornecidos na tabela 1, portanto, análises futuras poderão evidenciar melhores resultados com outros métodos, que podem se destacar em relação ao boost_rf, como por exemplo, *artificial neural network* (ANN) que em nossos resultados preliminares foi eleito como o segundo melhor método para classificação.

Outra importante conclusão de nossos resultados, é que a informação do redshift não aparenta significativa relevância para a classificação (de acordo com (LOCHNER et al., 2016)). As diferenças são muito pequenas entre os resultados da AUC para análises com e sem o redshift, exceto no caso do modelo SALT2, figura 9(a) e (b) (ROC SALT2), isso porque o redshift é um prior para esse modelo.

Por fim, os resultados preliminares obtidos com a análise de nosso conjunto simulado, são comparáveis aos encontrados em (LOCHNER et al., 2016) para os métodos de extração de características e classificação.

Esperamos que esse trabalho introduza alunos de pós-graduação e desperte o interesse pelo uso de métodos de aprendizagem de máquina e redes neurais artificiais, envolvendo também outros problemas de análise.

Desejamos continuar nosso estudo e ampliar a análise preliminar para um conjunto de dados simulados com 10% dos dados esperados para o LSST ($\sim 2 \times 10^5$ Supernovas) na obtenção dos resultados finais. Temos a expectativa de aumentar o envolvimento com demais problemas de classificação para o projeto LSST e futuras pesquisas observacionais.

Referências

AMENDOLA, L. Large-scale inhomogeneities may improve the cosmic concordance of supernovae. *Physical Reviews*, v. 105, 2010. Citado na página 32.

ANDERSON, J. A. A simple neural network generating an interactive memory. *Mathematical biosciences*, Elsevier, v. 14, n. 3-4, p. 197–220, 1972. Citado na página 100.

ANTARES-COLLABORATION et al. Machine learning-based brokers for real-time classification of the lsst alert stream. mar 2018. Disponível em: <<http://arxiv.org/abs/1801.07323>>. Citado 2 vezes nas páginas 17 e 43.

ANTOINE, J.-P. et al. *Two-dimensional Wavelets and Their Relatives*. [S.l.]: Cambrigde University, 2008. Citado na página 84.

ARNETT, W. D. A possible model of supernovae: Detonation of ^{12}C . *Astrophysics and Space Science*, v. 5, 1969. Disponível em: <<https://link.springer.com/article/10.1007/BF00650291>>. Citado na página 33.

ARNETT, W. D.; BRANCH, D.; WHEELER, J. C. Hubble's constant and exploding carbon-oxygen white dwarf models for type I supernovae. *Nature*, v. 314, 1985. Citado na página 33.

BAADE, W. The absolute photographic magnitude of supernovae. *Astrophysical Journal*, v. 88, 1938. Disponível em: <<http://adsabs.harvard.edu/full/1938ApJ....88..285B>>. Citado na página 32.

BAADE, W.; ZWICKY, F. Cosmic rays from super-novae. *National Academy of Sciences*, v. 20, 1934. Disponível em: <<http://www.pnas.org/content/20/5/259>>. Citado na página 32.

BARBARY, K.; CONTRIBUTORS. Sncosmo documentation - release 1.6.0. Abril 2018. Disponível em: <<https://media.readthedocs.org/pdf/sncosmo/v1.6.x/sncosmo.pdf>>. Citado na página 58.

BARBON, R.; CIATTI, F.; ROSINO, L. On the light curve and properties of type I supernovae. *Astronomy and Astrophysics*, v. 25, 1973. Disponível em: <<http://adsabs.harvard.edu/full/1973A%26A....25..241B>>. Citado na página 34.

BAZIN, G. et al. The core-collapse rate from the supernova legacy survey. *Astronomy and Astrophysics*, v. 499, p. 653–660, 2009. Citado na página 56.

BERTOLA, F. The supernovae in NGC 1073 and NGC 1058. *Annales d'Astrophysique*, v. 27, 1964. Citado na página 29.

BESSEL, M. S. The late-M dwarfs. *Astronomical Journal*, v. 101, 1991. Disponível em: <<http://adsabs.harvard.edu/full/1991AJ....101..662B>>. Citado na página 37.

BOHLIN, R. C. Hubble space telescope calspec flux standards: Sirius (and Vega). *Astrophysical Journal*, v. 147, 2014. Citado na página 39.

- BOSER, B. E.; GUYON, I. M.; VAPNIK, V. N. A training algorithm for optimal margin classifiers. In: ACM. Proceedings of the fifth annual workshop on Computational learning theory. [S.l.], 1992. p. 144–152. Citado na página 63.
- BRAGA, A. de P.; CARVALHO, A. P. de Leon F. de; LUDERMIR, T. B. Redes Neurais Artificiais teoria e aplicações. 2. ed. [S.l.]: LTC-GEN, 2014. Citado na página 98.
- BRANCH, D. Type Ia supernovae as standard candles. Annu. Rev. Astron. Astrophys., v. 30, 1992. Citado na página 33.
- BREIMAN, L. Bagging predictors. Machine learning, Springer, v. 24, n. 2, p. 123–140, 1996. Citado 2 vezes nas páginas 64 e 69.
- BREIMAN, L. Pasting small votes for classification in large databases and on-line. Machine learning, Springer, v. 36, n. 1-2, p. 85–103, 1999. Citado na página 69.
- BUHLMANN, P.; HOTHORN, T. Boosting algorithms: Regularization, prediction and model fitting. Statistical Science, v. 22, n. 4, 2007. Disponível em: <arxiv.org/pdf/0804.2752>. Citado na página 66.
- BURFIELD, C. Floyd-warshall algorithm. Massachusetts Institute of Technology, 2013. Citado na página 76.
- BURNS, C. R. et al. The carnegie supernova project: Light curve fitting with snoopy. Publications of the Astronomical Journal, v. 141, n. 1. Disponível em: <<http://iopscience.iop.org/article/10.1088/0004-6256/141/1/19/meta>>. Citado na página 57.
- CARDELLI, J. A.; CLAYTON, G. C.; MATHIS, J. S. The relationship between infrared, optical and ultraviolet extinction. Astrophysical Journal, v. 345, 1989. Disponível em: <<http://articles.adsabs.harvard.edu/full/1989ApJ...345..245C/0000246.000.html>>. Citado na página 50.
- CHANDRASEKHAR, S. The maximum mass of ideal white dwarfs. The Astrophysical Journal, v. 74, p. 81, 1931. Citado na página 26.
- CHARNOCK, T.; MOSS, A. Deep recurrent neural networks for supernovae classification. Astrophysical Journal, v. 837, 2017. Disponível em: <<https://doi.org/10.3847/2041-8213/aa603d>>. Citado 3 vezes nas páginas 17, 90 e 98.
- CHOUDHURI, A. R. Astrophysics for Physicists. [S.l.]: Cambridge University, 2010. Citado 4 vezes nas páginas 14, 20, 26 e 28.
- CHUNG, J. et al. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555, 2014. Citado na página 98.
- CLARET, A.; DÍAS-CORDOVÉS, J.; GIMÉNEZ, A. Linear and non-linear limb-darkening coefficients for the photometric bands $R I J H K$. Astronomy e Astrophysics Supplement Series, v. 114, 2007. Disponível em: <<http://iopscience.iop.org/article/10.1086/518000/pdf>>. Citado na página 37.
- COELHO, R. C. V. et al. Standardization of type Ia supernovae. 2014. Disponível em: <<http://arxiv.org/abs/1411.3596>>. Citado 2 vezes nas páginas 35 e 36.

- COLGATE, S. A.; MCKEE, C. Early supernova luminosity. *Astrophysics and Space Science*, v. 157, 1969. Disponível em: <<http://adsabs.harvard.edu/full/1969ApJ...157..623C>>. Citado na página 33.
- CONLEY, A. et al. Supernova constraints and systematic uncertainties from the first three years of the supernova legacy survey. *The Astrophysical Journal Supplement Series*, v. 192, mar 2010. Citado na página 32.
- CONNOLLY, N.; CONNOLLY, B. A bayesian approach to classifying supernovae with color. 2009. Disponível em: <<http://arxiv.org/abs/0909.3652>>. Citado na página 31.
- CORASANITI, P. S. The impact of cosmic dust on supernova cosmology. *Monthly Notices of the Royal Astronomical Society*, v. 372, 2006. Citado na página 32.
- CORTES, C.; VAPNIK, V. Support-vector networks. *Machine learning*, Springer, v. 20, n. 3, p. 273–297, 1995. Citado na página 63.
- DAI, M. et al. Photometric classification and redshift estimation of lsst supernovae. *Monthly Notices of the Royal Astronomical Society*, Oxford University Press, v. 477, n. 3, p. 4142–4151, 2018. Citado 2 vezes nas páginas 17 e 58.
- DAUBECHIES, I. Orthonormal bases of compactly supported wavelets. AT e T Bell Laboratories, mar 1989. Citado na página 84.
- DAVIS, T. M. The effect of peculiar velocities on supernova cosmology. *Astrophysical Journal*, v. 741, 2011. Citado na página 31.
- DEUSTUA, S. E. et al. Wfc3/uvis update 2017 chip-dependent inverse sensitivity values. 2017. Disponível em: <http://www.stsci.edu/hst/wfc3/phot_zp_lbn>. Citado na página 39.
- DOGGETT, J. B.; BRANCH, D. A comparative study of supernova light curves. *Astronomical Journal*, v. 90, 1985. Disponível em: <<http://adsabs.harvard.edu/full/1985AJ.....90.2303D>>. Citado na página 33.
- DRESSLER, A. Galaxy morphology in rich clusters-implications for the formation and evolution of galaxies. *The Astrophysical Journal*, v. 236, p. 351–365, 1980. Citado na página 62.
- FABRIS, J. C. et al. *Programa mínimo de Cosmologia*. [S.l.]: JAVÁ editora - CBPF, 2010. Citado na página 18.
- FEINDT, U.; KOWALSKI, M.; PAECH, K. The self-calibrating hubble diagram. *Journal of Cosmology and Astroparticle Physics*, v. 04.001, 2012. Citado na página 31.
- FEROZ, F.; HOBSON, M. Multimodal nested sampling: an efficient and robust alternative to markov chain monte carlo methods for astronomical data analyses. *Monthly Notices of the Royal Astronomical Society*, v. 384, p. 449–463, 2008. Disponível em: <academic.oup.com/mnras/article/384/2/449/1023701>. Citado na página 56.
- FEROZ, F.; HOBSON, M.; BRIDGES, M. Multinest: an efficient and robust bayesian inference tool for cosmology and particle physics. *Monthly Notices of the Royal Astronomical Society*, v. 398, p. 1601–1614, 2009. Disponível em: <academic.oup.com/mnras/article/398/4/1601/981502>. Citado na página 56.

- FILIPPENKO, A. V. Type ia supernovae in elliptical and spiral galaxies - possible differences in photometric homogeneity. *Astronomical Society of the Pacific*, v. 101, Jun. 1989. Disponível em: <<http://adsabs.harvard.edu/full/1989PASP..101..588F>>. Citado na página 45.
- FOLEY, R. J. A mismatch in the ultraviolet spectra between low-redshift and intermediate-redshift type ia supernovae as a possible systematic uncertainty for supernova cosmology. *Astrophysical Journal*, v. 143, 2012. Disponível em: <<http://arxiv.org/abs/1010.2749>>. Citado na página 31.
- FOWLER, R. H. On dense matter. *Monthly Notices of the Royal Astronomical Society*, v. 87, p. 114–122, 1926. Citado na página 23.
- FREUND, Y.; SCHAPIRE, R. A decision-theoretic generalization of on-line learning and an application to boosting. *AT and T. Lab*, 1996. Citado na página 65.
- FREUND, Y.; SCHAPIRE, R. Experiments with a new boosting algorithm. In *Machine Learning: Proceedings of the Thirteenth International Conference*, 1996. Citado 2 vezes nas páginas 65 e 67.
- FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* (2001), v. 29, n. 5, 1999. Citado 2 vezes nas páginas 67 e 68.
- FRIEDMAN, J. H.; HASTIER, T.; TIBSHIRANI, R. Additive logistic regression: a statistical view of boosting (with discussion). *The Annals of Statistics*, v. 28, 2000. Citado na página 66.
- GABOR, D. Theory of communication. part 1: The analysis of information. *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering, IET*, v. 93, n. 26, p. 429–441, 1946. Citado 2 vezes nas páginas 83 e 84.
- GASTWIRTH, J. L. A general definition of the lorenz curve. *Econometrica: Journal of the Econometric Society*, JSTOR, p. 1037–1039, 1971. Citado na página 64.
- GILKS, W. R.; RICHARDSON, S.; SPIEGELHALTER, D. *Markov chain Monte Carlo in practice*. [S.l.]: CRC press, 1995. Citado na página 74.
- GIUNT, C.; KIM, C. W. *Fundamentals of Neutrino Physics and Astrophysics*. 1. ed. New York (USA): Oxford University Press Inc., 2007. Citado 2 vezes nas páginas 14 e 29.
- GLAZEBROOK, K. et al. An imaging k-band survey – i. the catalogue, star and galaxy count. *Monthly Notices of the Royal Astronomical Society*, v. 266, 1994. Citado na página 43.
- GOLDHABER, G. et al. Timescale stretch parameterization of type ia supernova b-band light curves. *Publications of the Astrophysical Journal*, v. 558, n. 1. Disponível em: <<https://arxiv.org/pdf/astro-ph/0104382>>. Citado na página 57.
- GOOSSENS, M.; RAHTZ, S.; MITTELBAACH, F. *Illustrating documents with TEX and PostScript*. [S.l.]: Addison-Wesley, 1997. Citado 2 vezes nas páginas 8 e 92.
- GREEN, D. M.; SWETS, J. A. *Book. Signal detection theory and psychophysics*. Reprint ed. Los Altos, Calif. : Peninsula Publishing, 1988. ISBN 0932146236. Disponível em: <<https://trove.nla.gov.au/work/7432279>>. Citado na página 90.

- GROSSMANN, A. Wavelet transforms and edge detection. In: Stochastic processes in physics and engineering. [S.l.]: Springer, 1988. p. 149–157. Citado na página 82.
- GROSSMANN, A.; MORLET, J. Decomposition of hardy functions into square integrable wavelets of constant shape. SIAM journal on mathematical analysis, SIAM, v. 15, n. 4, p. 723–736, 1984. Citado na página 82.
- GROSSMANN, A.; MORLET, J. Decomposition of functions into wavelets of constant shape, and related transforms. In: Mathematics+ Physics: Lectures on Recent Results (Volume 1). [S.l.]: World Scientific, 1985. p. 135–165. Citado na página 82.
- GUILLOCHON, J. et al. An open catalog for supernova data. 1993. Disponível em: <http://arxiv.org/abs/1605.01054>. Citado na página 45.
- GUO, S.-L. et al. Improved dijkstra algorithm based on fibonacci heap for solving the shortest path problem with specified nodes. World Scientific. Citado na página 76.
- GUY, J. Salt2: using distant supernovae to improve the use of type ia supernovae as distance indicators. Publications of the Astronomical Society of the Pacific, v. 466, 2007. Citado 6 vezes nas páginas 8, 51, 52, 53, 54 e 55.
- GUY, J. et al. Salt: A spectral adaptive light curve template para type ia supernovae. Astronomy and Astrophysics, v. 443, 2005. Citado na página 51.
- GUY, J. et al. Salt: A spectral adaptive light curve template fo type ia supernovae. 2008. Citado 3 vezes nas páginas 51, 54 e 57.
- GÉRON, A. Hands-On Machine Learning with Scikit-Learn and TensorFlow. 1. ed. [S.l.]: O'Reilly Media Inc., 2017. Citado 6 vezes nas páginas 59, 69, 90, 93, 95 e 101.
- HAMUY, M. et al. The morphology of type ia supernovae light curves. sep 1996. Disponível em: <https://arxiv.org/pdf/astro-ph/9609063.pdf>. Citado na página 49.
- HAYDEN, B.; GARNAVICH, P.; SURVEY, S.-I. S. The rise and fall of sdss-ii supernovae. Publications of the American Astronomical Society, v. 41. Disponível em: <http://adsabs.harvard.edu/abs/2009AAS...21342513H>. Citado na página 57.
- HAYES, D. S.; LATHAM, D. W. A rediscussion of the atmospheric extinction and the absolute spectral-energy distribution of vega. Astrophysical Journal, v. 197, 1975. Disponível em: http://articles.adsabs.harvard.edu/full/1975ApJ...197..593H&data_type=PDF_HIGH&type=PRINTER&filetype=.pdf. Citado na página 37.
- HEBB, D. O. The organization of behavior: A neurophysiological approach. [S.l.]: John Wiley, 1949. Citado 2 vezes nas páginas 91 e 100.
- HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. Neural computation, MIT Press, v. 9, n. 8, p. 1735–1780, 1997. Citado na página 98.
- HOFLICH, P.; KHOKHLOV, A. Explosion models for type ia supernovae: A comparison with observed light curves, distances, ho and qo. feb 1996. Disponível em: <http://arxiv.org/abs/9602025>. Citado na página 34.

HOFLICH, P.; MULLER, E.; KHOKHLOV, A. Light curve models for type ia supernovae: Physical assumptions, their influence and validity. *Astronomy and Astrophysics*, v. 268, 1993. Disponível em: <<http://adsabs.harvard.edu/full/1993A%26A...268..570H>>. Citado na página 33.

HOPFIELD, J. J. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America*, v. 79, p. 2554–2558, Abril 1982. Disponível em: <<http://www.pnas.org/content/pnas/79/8/2554.full.pdf>>. Citado 4 vezes nas páginas 91, 98, 100 e 101.

HORNIK, K. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, v. 4, p. 251–257, 1991. Disponível em: <https://web.njit.edu/~usman/courses/cs675_fall17/hornik-nn-1991.pdf>. Citado na página 94.

HOYLE, F.; FOWLER, W. A. Nucleosynthesis in supernovae. *Astrophysical Journal*, v. 132, 1960. Disponível em: <<http://adsabs.harvard.edu/full/1960ApJ...132..565H>>. Citado na página 33.

IVEZIC, Z. et al. *Statistical, Data Mining, and Machine Learning in Astronomy. A Practical Python Guide for the Analysis of Survey Data*. 1. ed. New Jersey (USA): Princeton University Press, 2014. Citado 2 vezes nas páginas 62 e 75.

JACKSON, J. D. *Classical Electrodynamics*. 3. ed. [S.l.]: John Wiley and Sons, 1999. Citado na página 21.

JEFFREYS, H. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, The Royal Society, v. 186, n. 1007, p. 527–529, 1946. ISSN 00804630. Disponível em: <<http://www.jstor.org/stable/97878>>. Citado na página 130.

JEFFREYS, H. *Scientific Inference*. 1. ed. [S.l.]: ST John's College Cambridge, 1931. Citado na página 130.

JEFFREYS, H. *The Theory of Probability*. 3. ed. [S.l.]: Oxford University Press, 1961. Citado na página 130.

JHA, S.; RIESS, A. G.; KIRSHNER, R. P. Improved distances to type ia supernovae with multicolor light-curve shapes: Mlcs2k2. *Astrophysical Journal*, v. 659, 2007. Citado 2 vezes nas páginas 50 e 57.

JOHNSON, H. L.; MORGAN, W. W. Fundamental stellar photometry for standards of spectral type on the revised system of the yerkes spectral atlas. *Astrophysical Journal*, v. 117, 1953. Disponível em: <http://adsbit.harvard.edu/cgi-bin/nph-iarticle_query?1953ApJ...117..313J&defaultprint=YES&page_ind=0&filetype=.pdf>. Citado na página 35.

KARPENKA, N. V.; FERROZ, F.; HOBSON, M. P. A simple and robust method for automated photometric classification of supernovae using neural networks. *Monthly Notices of the Royal Astronomical Society*, 2012. Disponível em: <<https://arxiv.org/pdf/astro-ph/12081264>>. Citado 3 vezes nas páginas 9, 56 e 103.

- KELLY, P. L. Hubble residuals of nearby type ia supernovae are correlated with host galaxy masses. *Astrophysical Journal*, v. 715, 2010. Citado na página 32.
- KESSLER, R. First-year sloan digital sky survey-ii supernova results: Hubble diagram and cosmological parameters. *The Astrophysical Journal Supplement Series*, v. 185, 2009. Citado 2 vezes nas páginas 31 e 42.
- KESSLER, R. Snana user's manual:simulation, lightcurve fitters and cosmology fitters. Department of Astronomy and Astrophysics - University of Chicago, 2018. Disponível em: <http://snana.uchicago.edu/doc/snana_manual.pdf>. Citado 5 vezes nas páginas 57, 58, 102, 103 e 104.
- KESSLER, R. et al. Snana: A public software package for supernova analysis. *Publications of the Astronomical Society of the Pacific*, v. 121, 2009. Citado na página 57.
- KESSLER, R. et al. Results from the supernova photometric classification challenge. *Astronomical Society of the Pacific*, v. 122, p. 1415–1431, Dec. 2010. Citado na página 56.
- KESSLER, R. et al. Supernova photometric classification challenge. Jan 2010. Disponível em: <<http://arxiv.org/pdf/1001.5210>>. Citado 3 vezes nas páginas 55, 57 e 103.
- KIRSHINER, R. P. et al. The spectra of supernovae. *Astrophysical Journal*, v. 185, 1973. Disponível em: <<http://adsabs.harvard.edu/full/1973ApJ...185..303K>>. Citado na página 33.
- KNIGHTS, M. et al. Extending beams to incorporate correlated systematic uncertainties. mar 2014. Disponível em: <<http://arxiv.org/abs/1205.3493>>. Citado na página 136.
- KOHONEN, T. Correlation matrix memories. *IEEE transactions on computers*, IEEE, v. 100, n. 4, p. 353–359, 1972. Citado na página 100.
- KOWAL, C. T. Absolute magnitudes of supernovae. *Astronomical Journal*, v. 73, 1968. Disponível em: <<http://adsabs.harvard.edu/full/1968AJ.....73.1021K>>. Citado na página 33.
- KUNZ, M.; BASSET, B. A.; HLOZEK, R. A. Bayesian estimation applied to multiple species: Towards cosmology with a million supernovae. mar 2007. Citado 2 vezes nas páginas 16 e 136.
- LANDAHL, H. D.; MCCULLOCH, W. S.; PITTS, W. A statistical consequence of the logical calculus of nervous nets. *The bulletin of mathematical biophysics*, v. 5, p. 115–133, Dec 1943. Disponível em: <link.springer.com/article/10.1007%2F02478260?LI=true>. Citado na página 91.
- LANDOLT, A. U. *UBVRI* photometric standard stars around the sky at - 50 declination. *Astronomical Journal*, v. 133, 2007. Disponível em: <<http://iopscience.iop.org/article/10.1086/518000/pdf>>. Citado na página 37.
- LAPLACE, P. S. *Mémoire sur la probabilité des causes par les événements*. [S.l.]: Mémoire de l'Académie royale des sciences, 1878. 621 p. Citado na página 130.
- LECLERCQ, F. Bayesian optimisation for likelihood-free cosmological inference. *astro-ph*, arXiv, 2018. Disponível em: <<https://arxiv.org/pdf/astro-ph/1805.07152>>. Citado na página 31.

LECUN, Y. et al. Generalization and network design strategies. Connectionism in perspective, Citeseer, p. 143–155, 1989. Citado 2 vezes nas páginas 95 e 97.

LEIBUNDGUT, B. et al. The dilation in the light curve of the distant type ia supernova sn 1995k. *Astrophysical Journal*, v. 466, 1996. Disponível em: <<http://iopscience.iop.org/article/10.1086/310164/pdf>>. Citado na página 50.

LEIBUNDGUT, B.; TAMMANN, G. A. Supernova studies iii. the calibration of the absolute magnitude of supernovae of type ia. *Astronomy and Astrophysics*, v. 230, 1989. Disponível em: <<http://adsabs.harvard.edu/full/1990A%26A...230...81L>>. Citado na página 33.

LI, M.; VITÁNYI, P. M. B. AN INTRODUCTION TO KOLMOGOROV COMPLEXITY AND ITS APPLICATIONS. 3. ed. Massachusetts (USA): Springer, 2014. Citado na página 131.

LIVNE, E. Sucessive detonations in accreting white dwarfs as an alternative mechanism for type i supernovae. *Astrophysical Journal*, v. 354, 1990. Citado na página 33.

LIVNE, E.; ARNETT, D. Explosion of sub-chandrasekhar mass white dwarfs in two dimensions. *Astrophysical Journal*, v. 452, 1995. Disponível em: <<http://adsabs.harvard.edu/abs/1995ApJ...452...62L>>. Citado na página 34.

LOCHNER, M. et al. Photometric supernova classification with machine learning. mar 2016. Disponível em: <<http://arxiv.org/abs/1603.00882>>. Citado 11 vezes nas páginas 8, 12, 17, 58, 82, 85, 90, 102, 103, 104 e 115.

MACDONALD, E. C. et al. The oxford–dartmouth thirty degree survey – i. observations and calibration of a wide-field multiband survey. *Monthly Notices of the Royal Astronomical Society*, v. 352, 2004. Citado na página 43.

MALLAT, S. A Wavelet Tour of Signal Processing. 3. ed. [S.l.]: Academic Press, 2009. Citado na página 85.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, v. 5, p. 115–133, Dec 1943. Disponível em: <<https://link.springer.com/article/10.1007/BF02478259>>. Citado na página 91.

MEZZACAPPA, A.; FULLER, G. M. Open Issues in Core Collapse Supernova Theory. 1. ed. New Jersey (USA): World Scientific Publishing Co. Pte. Ltd., 2005. Citado na página 13.

MILLER, R. G. The jackknife-a review. *Biometrika*, Oxford University Press, v. 61, n. 1, p. 1–15, 1974. Citado na página 60.

MINKOWSKI, R. Spectra of supernovae. *Astronomical Society of the Pacific*, v. 53, p. 224, 1941. Disponível em: <<http://adsabs.harvard.edu/full/1941PASP...53..224M>>. Citado 2 vezes nas páginas 29 e 32.

MINKOWSKI, R. Supernovae and supernova remnants. *Annual Review of Astronomy and Astrophysics*, v. 2, 1964. Disponível em: <<http://adsabs.harvard.edu/full/1964ARA%26A...2..247M>>. Citado na página 33.

- MINSKY, M.; PAPERT, S. Perceptron (expanded edition). [S.l.]: MIT Press, Cambridge, MA. Original edition published in, 1969. Citado na página 91.
- MORÉ, J. J. The levenberg-marquardt algorithm: implementation and theory. In: Numerical analysis. [S.l.]: Springer, 1978. p. 105–116. Citado na página 74.
- MORETTIN, P. A. Ondas e Ondaletas Vol. 23. [S.l.]: Edusp, 1999. Citado na página 82.
- MORLET, J. Sampling theory and wave propagation. In: Issues in acoustic Signal—image processing and recognition. [S.l.]: Springer, 1983. p. 233–261. Citado 2 vezes nas páginas 82 e 88.
- NEWLING, J. Parameter estimation with bayesian estimation applied to multiple species in the presence of biases and correlations. *Monthly Notices of Royal Astronomical Society*, v. 421, 2012. Citado na página 136.
- NEWLING, J. et al. Statistical classification techniques for photometric supernova typing. *Monthly Notices of the Royal Astronomical Society*, 2010. Disponível em: <<https://arxiv.org/abs/1010.1005>>. Citado 5 vezes nas páginas 9, 55, 56, 103 e 136.
- NG, A. Cs229 lecture notes. CS229 Lecture notes, v. 1, n. 1, 2000. Citado 2 vezes nas páginas 76 e 79.
- NOMOTO, K.; THIELEMANN, F.-K.; YOKOI, K. Accreting white dwarf models for type i supernovae iii. carbon deflagration supernovae. *Astrophysical Journal*, v. 286, 1984. Citado na página 33.
- NUGENT, P.; KIM, A.; PERLMUTTER, S. K-correction and extinction correction for type ia supernovae. sep 2002. Disponível em: <<https://arxiv.org/pdf/astro-ph/0205351.pdf>>. Citado 2 vezes nas páginas 50 e 51.
- OKE, J. B.; GUNN, J. Secondary standard stars for absolute spectrophotometry. *Astrophysical Journal*, v. 266, 1983. Disponível em: <<http://adsabs.harvard.edu/full/1983ApJ...266..713O7>>. Citado 2 vezes nas páginas 38 e 41.
- OKE, J. B.; SANDAGE, A. Energy distributions, k corrections, and the stebbins-whithford effect for giant elliptical galaxies. *Astrophysical Journal*, v. 154, 1968. Disponível em: <<http://adsabs.harvard.edu/abs/1968ApJ...154...21O>>. Citado na página 50.
- OKE, J. B.; SEARLE, L. The spectra of supernovae. *Annual Review of Astronomy and Astrophysics*, v. 12, 1974. Disponível em: <<http://adsabs.harvard.edu/full/1974ARA%26A..12..315O>>. Citado na página 33.
- OLIVEIRA, H. M. de. Análise de Fourier e Wavelets: Sinais Estacionários e Não Estacionários. 1. ed. Recife - PE (BR): Editora Universitária UFPE, 2007. Citado 2 vezes nas páginas 82 e 84.
- PADMANABHAN, N. et al. An improved photometric calibration of the sloan digital sky survey imaging data. *The Astrophysical Journal*, v. 674, 2008. Citado na página 44.
- PALANQUE-DELABROUILLE, N. Photometric redshifts for type ia supernovae in the supernova legacy survey. *Astronomy and Astrophysics*, v. 514, 2010. Citado na página 31.

- PATHRIA, R. K. Statistical Mechanics. [S.l.]: Butterworth-Heinemann, 1996. Citado na página 21.
- PERLICK, V. Distance measures in cosmology. oct 2000. Disponível em: <https://arxiv.org/abs/astro-ph/9905116>. Citado na página 36.
- PERLICK, V. Gravitational lensing from spacetime perspective. oct 2010. Disponível em: <https://arxiv.org/pdf/1010.3416.pdf>. Citado na página 36.
- PHILLIPS, M. M. The absolute magnitudes of type ia supernovae. *Astrophysical Journal*, v. 413, 1993. Disponível em: <http://articles.adsabs.harvard.edu/full/1993ApJ...413L.105P>. Citado 3 vezes nas páginas 30, 34 e 46.
- P.HOFLICH; KHOKHLOV, A. M.; WHELLER, J. C. Delayed detonation models for normal and subluminous type ia supernovae: Absolute brightness, light curves and molecule formation. *Astrophysical Journal*, v. 444, 1995. Disponível em: <http://adsabs.harvard.edu/abs/1995ApJ...444..831H>. Citado 2 vezes nas páginas 33 e 34.
- PIEREL, J. et al. Extending supernova spectral templates for next-generation space telescope observations. *arXiv preprint arXiv:1808.02534*, 2018. Disponível em: <http://adsabs.harvard.edu/abs/2018AAS...23124504R>. Citado na página 102.
- POWERS, D. M. The problem with kappa. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. [S.l.], 2012. p. 345–355. Citado na página 90.
- RIESS, A. G.; PRESS, W. H.; KIRSHNER, R. P. A precise distance indicator: Type ia supernova multicolor light curve shapes. *Astrophysical Journal*, 1996. Disponível em: <http://arxiv.org/abs/astro-ph/9604143v1>. Citado 2 vezes nas páginas 45 e 50.
- RILEY, H. N. The von neumann architecture of computer systems. Computer Science Department, California State Polytechnic University, 1987. Citado na página 91.
- ROBERT, C. P.; CHOPIN, N.; ROUSSEAU, J. Harold jeffreys's theory of probability revisited. *Statistical Science*, v. 24, n. 2, p. 141–172, 2009. Disponível em: <https://arxiv.org/pdf/0804.3173.pdf>. Citado na página 130.
- ROSENBLATT, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, v. 65, p. 386–408, Nov 1958. Disponível em: <http://psycnet.apa.org/doiLanding?doi=10.1037%2Fh0042519>. Citado 2 vezes nas páginas 91 e 94.
- RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. *nature*, Nature Publishing Group, v. 323, n. 6088, p. 533, 1986. Citado na página 99.
- SAKO, M. The sloan digital sky survey-ii supernova survey: Search algorithm and follow-up observations. *Astrophysical Journal*, v. 135, 2008. Disponível em: <http://arxiv.org/abs/1010.2749>. Citado na página 32.

- SANDAGE, A. et al. The cepheid distance to ngc 5253: Calibration of $m(\text{max})$ for the type ia supernovae sn 1972e and sn 1895b. *Astrophysical Journal*, v. 423, 1979. Disponível em: <<http://adsabs.harvard.edu/abs/1994ApJ...423L..13S>>. Citado na página 48.
- SAVAGE, B. D.; MATHIS, J. S. Observed properties of interstellar dust. *Astrophysical Journal*, v. 17, 1979. Disponível em: <<http://adsabs.harvard.edu/abs/1979ARA26A..17...73S>>. Citado na página 47.
- SCHAPIRE, R.; SINGER, Y. Improved boosting algorithms using confidence-rated predictions. *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, 1998. Citado na página 67.
- SCHMIDT, B. P. et al. The high-z supernova search : Measuring cosmic deceleration and global curvature of the universe using type ia supernovae. *Astrophysical Journal*, v. 507, 1998. Disponível em: <<http://iopscience.iop.org/article/10.1086/306308/pdf>>. Citado na página 36.
- SCHNEIDER, D. P.; GUNN, J. E.; G, H. J. Ccd photometry of abell clusters. i. magnitudes and redshifts for 84 brightest cluster galaxies. *Astrophysical Journal*, v. 264, 1983. Disponível em: <<http://articles.adsabs.harvard.edu/full/1983ApJ...264..337S/0000337.000.html>>. Citado na página 38.
- SELLO, S. Wavelet entropy and the multi-peaked structure of solar cycle maximum. *New Astronomy*, Elsevier, v. 8, n. 2, p. 105–117, 2003. Citado na página 87.
- SHAPIRO, S. L.; TEUKOLSKY, S. A. *Black Holes, White Dwarfs, and Neutron Stars*. [S.l.]: John Wiley and Sons, 1983. Citado 2 vezes nas páginas 20 e 28.
- SHAPLEY, H.; CURTIS, H. D. The scale of the universe. *Bulletin of the National Research Council*, v. 2, 1921. Disponível em: <<http://adsabs.harvard.edu/abs/1921BuNRC...2..171S>>. Citado na página 32.
- SHMILOVICI, A. Support vector machines. In: *Data mining and knowledge discovery handbook*. [S.l.]: Springer, 2009. p. 231–247. Citado na página 62.
- SIVIA, D.; SKILLING, J. *Data Analysis: A Bayesian Tutorial*. [S.l.]: Oxford University Press, 2006. Citado 2 vezes nas páginas 56 e 104.
- SKILLING, J. Nested sampling. *American Institute of Physics Conference Series*, v. 119, p. 1211–1232, 2004. Citado 2 vezes nas páginas 56 e 104.
- SKILLING, J. Nested sampling. In: *AIP Conference Proceedings*. [S.l.], 2004. v. 735, n. 1, p. 395–405. Citado na página 104.
- SPRINGENBERG, J. T. et al. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014. Citado na página 96.
- STUBBS, C. W. et al. Toward more precise survey photometry for panstarrs and lsst measuring directly the optical transmission spectrum of the atmosphere. *Astronomical Society of the Pacific*, v. 119, 2007. Disponível em: <<http://iopscience.iop.org/article/10.1086/522208/meta>>. Citado na página 43.

- STUBBS, C. W. et al. An overview of the large synoptic survey telescope (lsst) system. *Bulletin of the American Astronomical Society*, v. 36, 2004. Disponível em: <<http://adsabs.harvard.edu/abs/2004AAS...20510802S>>. Citado na página 42.
- STUBBS, C. W.; TONRY, J. L. Toward 1% photometry: End-to-end calibration of astronomical telescopes and detectors. *Astrophysical Journal*, v. 646, 2006. Disponível em: <<http://iopscience.iop.org/article/10.1086/505138/meta>>. Citado na página 42.
- TENENBAUM, J. B. Mapping a manifold of perceptual observations. In: *Advances in neural information processing systems*. [S.l.: s.n.], 1998. p. 682–688. Citado na página 76.
- TENENBAUM, J. B.; SILVA, V. D.; LANGFORD, J. C. A global geometric framework for nonlinear dimensionality reduction. *science*, American Association for the Advancement of Science, v. 290, n. 5500, p. 2319–2323, 2000. Citado na página 76.
- THUAN, T. X.; GUNN, J. E. A new four-color intermediate-band photometric system. *Astronomical Society of the Pacific*, v. 88, 1983. Disponível em: <<http://iopscience.iop.org/article/10.1086/129982/pdf>>. Citado na página 38.
- TONRY, J. L. et al. The pan-starrs1 photometric system. *Astrophysical Journal*, v. 750, 2012. Disponível em: <<http://iopscience.iop.org/article/10.1088/0004-637X/750/2/99/pdf>>. Citado na página 38.
- TURATTO, M. Classification of supernovae. 2003. Disponível em: <<http://arxiv.org/abs/0301107>>. Citado 2 vezes nas páginas 16 e 29.
- TURATTO, M.; BENETTI, S.; PASTORELLO, A. Supernova classes and subclasses. 2007. Disponível em: <<http://arxiv.org/abs/0706.1086>>. Citado 2 vezes nas páginas 16 e 29.
- TURING, A. M. On computable numbers, with an application to the entscheidungsproblem. *London Mathematical Society*, 1937. Disponível em: <<https://londmathsoc.onlinelibrary.wiley.com/doi/abs/10.1112/plms/s2-42.1.230>>. Citado na página 91.
- TYSON, A.; STRAUSS, Z. I. M. A. *LSST Science Book*. 1. ed. Arizona (USA): LSST Corporation., 2009. Citado 8 vezes nas páginas 16, 31, 40, 41, 42, 43, 44 e 103.
- UOMOTO, A.; KIRSHNER, R. P. Peculiar type i supernovas. *Astronomy and Astrophysics*, v. 149, 1985. Citado na página 33.
- VIERA, A. J.; GARRETT, J. M. et al. Understanding interobserver agreement: the kappa statistic. *Fam Med*, v. 37, n. 5, p. 360–363, 2005. Citado na página 90.
- VINYALS, O. et al. Show and tell: A neural image caption generator. In: *IEEE. Computer Vision and Pattern Recognition (CVPR)*, 2015 IEEE Conference on. [S.l.], 2015. p. 3156–3164. Citado na página 98.
- VISA, S. et al. Confusion matrix-based feature selection. In: *MAICS*. [s.n.], 2011. p. 120–127. Disponível em: <https://www.researchgate.net/profile/Atsushi_Inoue/publication/220833227_Page_Ranking_Refinement_Using_Fuzzy_Sets_and_Logic/links/54b743480cf24eb34f6e9e80.pdf#page=126>. Citado na página 89.

- WADE, R. A.; HOESSEL, J. g.; ELIAS, J. H. A two-color photometric system for the near infrared. *Astronomical Society of the Pacific*, v. 91, 1979. Disponível em: <<http://iopscience.iop.org/article/10.1086/130435/pdf>>. Citado na página 38.
- WHEELER, J. C.; BENETTI, S. *Allen's Astrophysical Quantities*. 4. ed. New York (USA): Springer Science+Business Media New York, 2002. Citado na página 29.
- WHEELER, J. C.; HARKNESS, R. P. Type i supernovae. *IOPscience*, v. 53, 1990. Citado na página 33.
- WHEELER, J. C.; MAZUREK, T. J. Thermonuclear explosions in stars. *Fundamentals of Cosmic Physics*, v. 5, 1980. Citado na página 33.
- WHELLER, J. C. et al. Supernovae, supernebulae and nucleosynthesis. *Astronomical Society of the Pacific*, v. 98, 1986. Citado na página 33.
- WHELLER, J. C.; LEVREAU, R. The peculiar type i supernova in ngc 991. *Astronomy and Astrophysics*, v. 294, 1985. Citado na página 33.
- WIDROW, B.; HOFF, M. E. *Adaptive switching circuits*. [S.l.], 1960. Citado 2 vezes nas páginas 91 e 98.
- WILLIAMS, T.; LI, R. Wavelet pooling for convolutional neural networks. *The Sixth International Conference on Learning Representations*, 2018. Disponível em: <https://www.researchgate.net/profile/Travis_Williams3/publication/324090115_Wavelet_Pooling_for_Convolutional_Neural_Networks/links/5abd121645851584fa6fb0fb/Wavelet-Pooling-for-Convolutional-Neural-Networks.pdf>. Citado na página 96.
- WOOLEY, S. E.; WEAVER, T. A. Sub-chandrasekhar mass models for type ia supernovae. *Astrophysical Journal*, v. 423, p. 371–379, 1994. Disponível em: <<http://adsabs.harvard.edu/abs/1994ApJ...423..371W>>. Citado na página 34.
- YAMAGUCHI, H.; KOYAMA, K.; UCHIDA, H. Suzaku view of the supernova remnant rcw 86:x-ray studies of newly-discovered fe-rich ejecta. 2012. Disponível em: <<http://arxiv.org/abs/1202.1594>>. Citado na página 28.

APÊNDICE A – Resumo de estatística

Iniciamos esse capítulo com uma introdução histórica sobre o tema, em seguida abordamos conteúdos básicos da estatística como probabilidades, o teorema de Bayes, verossimilhança (*likelihood*) e suas aplicações na comparação entre modelos.

Dados históricos indicam que a estatística Bayesiana surgiu primeiramente a partir de questionamentos de James Bernoulli (1654–1705), ao notar que existia uma diferença entre a lógica dedutiva (utilizada na matemática) e a lógica indutiva (que estamos acostumados a utilizar cotidianamente). Para Bernoulli, a questão estava em como utilizar a lógica dedutiva, bem conhecida desde a Grécia antiga, para nos ajudar a resolver problemas que exigissem raciocínio indutivo. A resposta veio em um trabalho postumamente publicado de Thomas Bayes (1702 – 1761), por seus filhos em 1763. No entanto, a formulação dessas idéias foi aprimorada por Pierre Simon Laplace (1749 – 1827), que por volta de 1812, redescobriu as idéias de Bayes, aplicando-as em diversas áreas do conhecimento, como medicina, física e ciências jurídicas (LAPLACE, 1878). Um famoso exemplo em Astrofísica foi o uso da teoria Bayesiana para estimar a massa de Saturno, utilizando dados observacionais e as leis físicas, seu resultado foi tão bom que em mais de 200 anos de observações posteriores, esse valor precisou ser ajustado em apenas 0.5%. Apesar dos inúmeros sucessos obtidos por Laplace, a comunidade matemática de seu tempo não achava que seria possível desenvolver essa teoria sobre uma base matemática sólida. Entretanto no século passado, essa teoria foi reavivada por Jeffreys (1891 – 1989) (JEFREYS, 1931; JEFFREYS, 1946; JEFREYS, 1961) e posteriormente reinterpretada com uma extensão da lógica Aristotélica. Essa teoria é dita como uma linguagem simplista da lógica Booleana, onde as hipóteses possíveis deixam de ser verdadeiro ou falso e passam a um mundo contínuo, no qual cada hipótese tem um grau de verdade. Todas as operações lógicas são aplicáveis nesse novo contexto, formando a base para o teorema de Bayes (ROBERT; CHOPIN; ROUSSEAU, 2009).

A.1 Probabilidades

Trataremos nessa introdução de relacionar as distribuições de probabilidades *marginais*, *conjuntas* e *condicionais*.

O termo probabilidade *marginal* (por, exemplo, $p(X)$) se refere a distribuição de variáveis marginais pertencentes a um subconjunto de nosso espaço de variáveis aleatórias. São denominadas marginais porque costumam ser encontradas por operações matemáticas nas linhas e colunas de uma tabela e escritas nas margens da mesma. Em muitos casos, a

análise de dados começa com um conjunto de variáveis aleatórias que depois é acrescido com novas variáveis (como por exemplo, a soma das variáveis originais). Quando nos concentramos na análise de um subconjunto marginal (como a soma), reduzindo o número de variáveis a serem analisadas para um, dizemos que estamos marginalizando sobre esses dados de interesse. No espaço *multivariado*, distintas análises podem ser feitas, cada uma tratando um subconjunto de variáveis diferentes como marginais.

A probabilidade *conjunta* (por exemplo, $P(X, Y) = P(X \cap Y)$) é a probabilidade dos eventos x e y ocorrerem simultaneamente), se refere a distribuição de variáveis aleatórias em um espaço ao menos bidimensional. No caso de termos apenas duas variáveis aleatórias chamamos de distribuição bivariada, mas esse conceito pode ser generalizado para qualquer número de variáveis aleatórias, sendo chamada de distribuição *multivariada*. Em um espaço de variáveis independentes, a distribuição de probabilidade conjunta é o produto das probabilidades marginais, já no caso de dependência entre as variáveis, veremos que a distribuição conjunta tem uma dependência direta com a *probabilidade condicional* através da *regra do produto*.

Nos referimos a probabilidade *condicional* (por exemplo, $P(X|Y)$ é a probabilidade do evento x ocorrer dado que y ocorra), como um dos conceitos mais fundamentais e importantes na teoria estatística. Pela definição de Kolmogorov (LI; VITÁNYI, 2014) a probabilidade condicional entre dois eventos é definida como o quociente da probabilidade conjunta pela probabilidade marginal do segundo evento ($P(X|Y) = P(X \cap Y)/P(Y)$). Importante observar que essa definição não é um resultado teórico (nos apenas denotamos a quantidade $P(X \cap Y)/P(Y)$ por $P(X|Y)$ e chamamos de probabilidade condicional). Se a distribuição condicional for contínua, podemos escrever uma função de densidade de probabilidade que permite expressar diferentes propriedades de uma distribuição, incluindo os momentos de diferentes ordens, muitas vezes também referido como probabilidade, média, variância, *skewness*, kurtosis dentre outros.

Momento estatístico é uma propriedade importante, descrita através de distribuições muito utilizada na física e estatística. Por exemplo, se representarmos cada um dos dados por um valor de massa, o momento zero é a massa total do sistema, o primeiro momento é o centro de massa e o terceiro momento é a inércia rotacional. Já se os dados representarem uma densidade de probabilidade então o momento zero é a probabilidade total (que é 1), o primeiro momento é a média, o segundo momento é a variância da distribuição, o terceiro é a assimetria (*skewness*) e o quarto momento é a curtose (*kurtosis*). Em geral, estes são os momentos estatísticos mais aplicados, eles são ditos centrais devido ao fato de que são calculados com base no desvio da média (valor esperado) da variável aleatória de interesse. Momentos de ordens superiores também podem ser calculados, mas esse não é nosso objetivo. Importante destacar que os vários momentos estatísticos formam um conjunto de valores através dos quais as propriedades de uma distribuição de probabilidade

podem ser totalmente descritas.

A.2 Probabilidades discretas ou contínuas

Até então, enfatizamos o uso de espaços discretos de hipóteses. Porém se fizermos hipóteses acerca de quantidades contínuas, como massa, distâncias e outras quantidades físicas não quantizadas, devemos propor que nossas hipóteses sejam contínuas. A passagem para um limite contínuo, é feita naturalmente sem qualquer perda de generalização, demonstramos como isso pode ser feito a seguir.

Considere uma hipótese qualquer H , acerca de um parâmetro que desejamos investigar e seja h uma hipótese acerca do valor numérico de H estar em um intervalo $[a, b]$. O operador lógico que delimita essa proposição é $+$ (ou), pois qualquer dos valores no intervalo é um valor válido. Então o que buscamos é a probabilidade da soma de todos os valores compreendidos no intervalo $[a, b]$:

$$P(a < H < b|I) = \sum_i P(H_i|I) \quad (\text{A.1})$$

Precisamos entender como proceder com o somatório na equação (A.1). Para tanto, voltamos as operações lógicas:

Regra do produto:

A e B são independentes: $P(A, B) = P(A)p(B)$

A e B são dependentes: $P(A, B) = P(B)P(A|B) = P(A)P(B|A) = P(B, A)$

Regra da soma:

A e B são independentes (*probabilidade exclusiva*): $P(A + B) = P(A) + p(B)$

A e B são dependentes (*probabilidade não exclusiva*): $P(A + B) = P(A) + P(B) - P(A, B)$

Claramente, esse resultado associado a (A.1) fornece:

$$P(a < H < b|I) = \sum_{i=a}^b P(H_i|I), \quad (\text{A.2})$$

no limite contínuo entre a e b (um caso em que existem infinitos pontos entre os extremos a e b das hipóteses), nesse caso $P(H_i|I)$ é uma função das nossas hipóteses prévias, que podem ser derivadas:

$$P(H|I) = \lim_{\delta h \rightarrow 0} \frac{P[h \leq H \leq (h + \delta h)|I]}{\delta h}. \quad (\text{A.3})$$

Logo nossa proposição h , acerca do valor numérico de H estar em um intervalo $[a, b]$, tem a seguinte evidência:

$$P(h|I) = \int_a^b P(H|I)dH, \quad (\text{A.4})$$

Claramente para o caso de simultâneas e múltiplas hipóteses, também obedecemos a regra do produto, recaindo naturalmente em derivadas e integrais múltiplas:

$$P(X, Y|I) = \lim_{\delta x, \delta y \rightarrow 0} \frac{p(x \leq X \leq x + \delta x, y \leq Y \leq y + \delta y|I)}{\delta x \delta y} \quad (\text{A.5})$$

$$P(h_X, h_Y|I) = \int_y^{y+\delta y} \int_x^{x+\delta x} P(X, Y|I)dXdY \quad (\text{A.6})$$

No espaço contínuo também definimos a densidade de probabilidade ou *função de distribuição de probabilidade (pdf)*, $f(x)$:

$$f(x)dx = P(x) \quad (\text{A.7})$$

A *pdf* tem duas propriedades fundamentais:

$$\int f(x)dx = 1 \quad (\text{A.8})$$

$$f(x) \geq 0 \quad (\text{A.9})$$

Podemos facilmente estender essa idéia para casos de probabilidades conjuntas:

$$f(x, y)dxdy = P(x, y) \quad (\text{A.10})$$

$$f(x, y) \geq 0 \quad (\text{A.11})$$

$$\int f(x, y)dxdy = 1 \quad (\text{A.12})$$

Nos casos em que ambos os eventos são independentes podemos também escrever:

$$P(X, Y) = P(X)p(Y) = \int_{X, X+\Delta X} f(x, y)dx \int_{Y, Y+\Delta Y} f(x, y)dy \quad (\text{A.13})$$

Nestas condições a PDF $f(x, y)$ é dita separável, isto é $f(x, y) = f_1(x)f_2(y)$

$$P(X, Y) = P(X)p(Y) = \int_{X, X+\Delta X} f_1(x)dx \int_{Y, Y+\Delta Y} f_2(y)dy \quad (\text{A.14})$$

obviamente isso, pode ser estendido para N eventos em uma distribuição multivariada.

Nos casos em que existem dependências entre as variáveis, a probabilidade conjunta é definida de forma análoga, através da regra do produto de probabilidades, visto anteriormente.

A.3 Teorema de Bayes

O teorema de Bayes é uma consequência direta da regra do produto. Usualmente descrito conforme a seguir:

$$P(H_i|D, I) = \frac{P(H_i|I)P(D|H_i, I)}{P(D|I)} \quad (\text{A.15})$$

onde H_i são as proposições afirmando a veracidade das hipóteses de interesse, D são as proposições sobre os dados, I são as proposições representando a informação a priori da distribuição, $P(H_i|D, I)$ é a probabilidade posterior das hipóteses, $P(H_i|I)$ é a probabilidade a priori das hipóteses, $P(D|H_i, I)$ é a probabilidade dos dados D (verossimilhança) se H_i e I são verdadeiros e por último, $P(D|I) = \sum_i P(H_i|I)P(D|H_i, I)$ pode ser interpretado como uma constante de normalização.

O teorema de Bayes nos diz que se inserimos nossas hipóteses prévias (H_i) na equação (A.15) e medirmos o seu grau de verdade ($P(H_i|I)$) em seguida, podemos revisar a veracidade de nossas hipóteses à luz dos dados ($P(D|H_i, I)$) gerando novos graus de veracidade corrigidos ($P(H_i|D, I)$), isto é, revisar a nossa teoria com base nos dados. Desta forma, podemos utilizar a distribuição posterior como uma probabilidade a priori, aplicando o teorema de Bayes sucessivas vezes, melhorando a veracidade de nossas hipóteses prévias e aprimorando a teoria. A interpretação do teorema de Bayes desta forma, representa um modelo matemático para o processo de aprendizado supervisionado.

A.4 Montagem da Verossimilhança (Likelihood)

A verossimilhança (em inglês *likelihood*) faz a ligação da teoria com os dados de nossas observações. A idéia básica por trás desse conceito, é tentar quantificar os desvios entre os dados medidos experimentalmente (observáveis) e os valores preditos teoricamente (parâmetros), isso pode ser feito por qualquer função matemática para quantificar desvios.

Quando supomos que todas as observações são independentes, comumente descrevemos os desvios por distribuição gaussiana.

Em termos gerais, $P(D|H_i, I)$ é $p(D|M)$ chamada de verossimilhança ou likelihood, de fato nosso modelo deve conter hipóteses e informações a priori, ela fornece a probabilidade de se obter os dados D se a teoria contida no modelo M for verdadeira, porém não é uma probabilidade propriamente, já que não é normalizada.

Na inferência clássica os parâmetros não são aleatórios, mas desconhecidos. Na inferência Bayesiana tratamos os parâmetros como quantidades aleatórias, seus valores estimados representam não um número, mas sim, uma distribuição de probabilidades. Na realidade $p(D|M)$ é uma função da teoria $p(D|M) \equiv (D|T)$ e não dos dados, por essa razão, muitas vezes a likelihood é escrita como:

$$\mathcal{L}(T) = p(D|T) \quad (\text{A.16})$$

Não há nenhuma maneira de resolver esse problemas dentro da inferência frequentista, somente através da estatística bayesiana pode-se obter $p(D|T)$ em que T é ao mesmo a variável lógica livre e quantidade real. O valor de $\hat{T}$ que maximiza $\mathcal{L}(T)$ é o valor mais provável da teoria, algo bastante razoável. Observe que $\hat{T}$ é em função de D , e portanto é uma variável do tipo dados aleatórios. A distribuição de $\hat{T}$ então pode ser obtidas analítica ou numericamente (via método Monte Carlo) a partir da distribuição de D . A distribuição de $\hat{T}$ fornece os intervalos de confiança para o verdadeiro T .

Pelo teorema de Bayes (A.15), temos duas likelihoods:

$$\mathcal{L}(D) = p(D|I) \quad (\text{A.17})$$

que é um fator de normalização e

$$\mathcal{L}(M) \equiv \mathcal{L}(T) = p(D|T) = p(D|H_i, I) \quad (\text{A.18})$$

que pode ser usada para encontrar o modelo M (as hipóteses(H_i) mais prováveis) com base nas informações prévias e nos dados, como veremos a seguir, ela também é chamada de likelihood global ou evidência.

A.5 Evidência e *Odds ratio*

A likelihood global $\mathcal{L}(T)$ é definida como $p(D|M) \equiv p(D|T) = p(D|H_i, I)$, que no caso contínuo pode ser definido por:

$$\mathcal{E} \equiv \mathcal{L}(M) = \int_i f(H_i|I) f(D|H_i, I) dH \quad (\text{A.19})$$

utilizando esse resultado podemos realizar a comparação entre modelos através do *odds ratio*:

$$\mathcal{O}_{ij} = \frac{p(H_i|D)}{p(H_j|D)} = \frac{p(H_i)}{p(H_j)} \mathcal{B}_{ij}, \quad (\text{A.20})$$

onde o fator de normalização $\mathcal{L}(D)$ é substituído pelo *fator de Bayes* $\mathcal{B}_{ij}$ que representa a razão entre as evidências dos modelos concorrentes,

$$\mathcal{B}_{ij} = \frac{\mathcal{E}_i}{\mathcal{E}_j}. \quad (\text{A.21})$$

A forma da evidência pode ser bem mais complicada em análises estatísticas amplas e multicorrelacionadas, como nos tratamentos de perturbações (*nuisance*) que enquanto não entendidas, são tomadas como ruídos. Para exemplos de aplicação dessas definições em problemas de classificação, veja também (NEWLING et al., 2010; KUNZ; BASSET; HLOZEK, 2007; NEWLING, 2012; KNIGHTS et al., 2014).