Estimativa do *redshift* fotométrico de galáxias utilizando redes neurais artificiais

Palavras-chave: [Galáxias], [Redshift], [Redes Neurais]
Discente: Sofia Garcia Telles Brito - IMECC, Unicamp
Docente: Profa. Dra. Flávia Sobreira - IFGW, Unicamp

Agosto 2024

Resumo

Esse trabalho consiste na revisão do *paper ANNz2 - Photometric Redshift and Probability Distribution Function Estimation Using Machine Learning* from SADEH, I., ABDALLA, F. B., LAHAV. O *ANNz2* é uma ferramenta que utiliza Redes Neurais Artificiais no treinamento de um algoritmo para que este consiga fazer estimativas do *redshift* fotométrico, z_{photo}, muito próximas do *redshift* espectroscópico, z_{spec}. Foram coletados os dados das magnitudes m_u, m_g, m_r, m_i e m_z, assim como seus respectivos erros associados, e dos *redshifts* espectroscópicos do catálogo do *Sloan Digital Sky Survey* (SDSS), do *Release* 10 (DR10), a fim de testar dois modos de operação do *ANNz2 Single Regression* e *Random Classification*.

Resumo das atividades

1 Introdução

Desde a descoberta da expansão cósmica em 1920 por Edwin Hubble, a cosmologia se tornou uma ciência de *big data* com a construção de telescópios modernos capazes de observar galáxias cada vez mais distantes. Uma informação importante é extraída do espectro da radiação destas galáxias. Quando se compara as linhas de absorção ou emissão da radiação destes objetos com as de vários compostos químicos na Terra, percebe-se um aumento no comprimento de onda dos fotóns caracterizado por um desvio na direção do vermelho. Isso indica que as galáxias estão se afastando de nós [1].

Redshifts espectroscópicos são estimados utilizando as linhas espectrais de objetos observados. Isto é possível em dados de telescópios como o Sloan Digital Sky Survey (SDSS) [5] e Dark Energy Spectroscopic Instrument (DESI) [6] que observam o espectro de uma galáxia.

Usando dados de telescópios espectroscópicos, podemos estimar os *redshifts* dos objetos com muita precisão. Porém, para observar o espectro inteiro de uma galáxia, é necessário um tempo

de observação grande. Já telescópios fotométricos, embora percam a precisão na estimativa dos *redshifts*, conseguem observar uma quantidade muito maior de galáxias, aumentando a precisão da análise estatística feita com os dados. Se sabemos modelar o erro na estimativa de *redshift* fotométrico, então é vantajoso usar estes dados.

Nesse projeto, o código ANNz2 do *paper ANNz2 - Photometric Redshift and Probability Distribution Function Estimation Using Machine Learning* [4] foi escolhido como ferramenta para fazer estimativas do *redshift* fotométrico utilizando redes neurais e diversos métodos de *Machine Learning*.

O objetivo, portanto, do projeto de Iniciação Científica é aprender conceitos de *Machine Learning* e redes neurais, revisar o *paper* citado acima [4] e aprender a trabalhar com os dados das magnitudes e dos *redshifts* espectroscópicos presentes na base de dados do SDSS (*Sloan Digital Sky Survey*), a fim de utilizar esses dados para, com base no que foi feito no *paper* [4], obter resultados utilizando os diversos métodos de treinamento nele apresentados.

2 Machine Learning

Machine Learning (Aprendizado de Máquina) se refere a um conjunto de técnicas para interpretação de dados em que se compara esses dados a modelos para o comportamento dos dados. Alguns exemplos dessas técnicas são métodos de regressão, métodos de classificação supervisionada, *maximum likelihood estimators* e o método Bayesiano [8].

2.1 Alguns conceitos de Machine Learning

2.1.1 Performance

Performance é a habilidade de prever o *redshift* de uma galáxia individual de forma precisa, ou seja, com uma pequena incerteza quando comparado com o verdadeiro *redshift*, que aqui escolheremos como sendo o *redshift* espectroscópico [2].

2.1.2 Caracterização

Caracterização é a habilidade de abarcar as propriedades da distribuição de um conjunto de galáxias [2].

2.1.3 Função Densidade de Probabilidade (PDF)

A Função Densidade de Probabilidade (PDF), $h(x)$, quantifica a probabilidade de que um valor esteja entre x e $x + dx$, que é igual a $h(x)dx$ [8]. A PDF do *redshift*, $p(z)$, nos dá a melhor representação do resultado de um algoritmo de *redshift* fotométrico [2]. A integral da PDF é chama Função Distribuição Cumulativa e é dada por [8]

$$H(x) = \int_{-z\infty}^{x} h(x')dx'.$$

3 ANNz2

No ANNz2, o conjunto de dados original utilizado (do SDSS - *Sloan Digital Sky Survey*) é separado em três partes: treinamento, validação e testagem. O conjunto de treinamento é usado para derivar o mapeamento entre *inputs* e *outputs*, enquanto que, a cada etapa do treinamento, o conjunto de validação é utilizado para estimar a convergência da solução comparando o resultado da estimativa com o valor do *output*. Já o conjunto de testagem é usado após o treinamento para analisar a performance deste. [4]

Os dados de *input* coletados do SDSS são as magnitudes m_u, m_g, m_r, m_i, m_z. O dado de *output* é a estimativa feita do *redshift* fotométrico [4]. Se olharmos para a Seção 2.3, que trata

sobre Regressão, os *inputs* das magnitudes seriam as componentes do vetor $\mathbf{x}$ e o *output* seria o y. O valor do *redshift* espectroscópico serve como valor real do *redshift*, para comparar com o valor de *output* encontrado para o *redshift* fotométrico.

No ANNz2, os métodos de *Machine Learning* utilizados estão implementados no pacote TMVA do ROOT. O ROOT é um *framework* de processamento de dados do CERN (*European Organization for Nuclear Research*. O TMVA se trata de uma biblioteca do ROOT que contém diversas implementações de técnicas de *Machine Learning*, como *Neural Networks*, *Deep Networks*, *Multilayer Perceptron*, *Boosted/Bagged Decision Trees*, *Support Vector Machines*(CVM) e outros [9]. Os métodos que foram considerados mais adequados foram redes neurais artificais e *boosted decision trees* [4].

3.1 Redes neurais artificiais (ANNs)

Trata-se de um mapeamento entre o conjunto de variáveis de *input* (como magnitudes e cores) a uma ou mais varíáveis de output, que é feito calculando a soma com pesos da coleção de funções de resposta (*response functions*). As variáveis de *input*, as funções de resposta e as variáveis de *output* são chamadas de neurônios [4]. No ANNz2 a ANNs utilizada foi a *Multilayer Perceptron*. Nessa rede, os neurônios são organizados em pelo menos 3 camadas: *input*, ocultos e *output*.

O aprendizado ocorre pela mudança de pesos inter-neuroniais após cada elemento do conjunto de dados ter sido processado, usando um algoritmo de *back propagation* [4].

3.2 *Boosted decision trees (BDTs)*

Se trata de uma árvore binária na qual as decisões são tomadas para uma variável por vez, até que o critério de parada seja satisfeito. Os vários nós de *output* da árvore são chamados de folhas (*leafs*).

3.3 Definição de métricas e notações

Algumas definições de métricas e notações importantes feitas no *paper* [4] são:

-Viés fotométrico: $\delta_{gal} = z_{phot} - z_{spec}$ [4];

-Espalhamento fotométrico: desvio padrão de δ_{gal} para um conjunto de galáxias [4];

-σ_{68} denota a meia largura da área que abrange o pico do $68^{\underline{o}}$ percentil da distribuição de δ_{gal} [4];

-Fração atípica da distribuição do viés: $f(\alpha\sigma)$, definida como a porcentagem de objetos que têm viés maior que um fator, α, de σ ou σ_{68} [4];

-Fração atípica combinada: $f(2, 3\sigma_{68} = \frac{1}{2}(f(2\sigma_{68}) + f(3\sigma_{68})$ [4].

3.4 Modos de treinamento do ANNz2

O ANNz2 utiliza tanto técnicas de regressão quanto técnicas de classificação. Os modos de treinamento do ANNz2 são *Single Regression*, *Random Regression*, *Binned Classification*, *Single Classification* e *Random Classification*. Esse modos de treinamento são descritos nas subseções a seguir.

3.4.1 Single Regression

Essa é a configuração mais simples do ANNz2. Nesse caso, uma única regressão é feita [7].

3.4.2 Random Regression

Na Regressão Aleatória (*Random Regression*), um conjunto de métodos de regressão é automaticamente gerado, sendo que esse métodos de *Machine Learning* diferem uns dos outros de diversas maneiras, como por exemplo no conjunto de parâmetros de entrada usado no treinamento. Assim que o treinamento ocorre, a otimização é realizada, obtendo-se uma distribuição de soluções do photo-z para cada galáxia e, então, essas soluções são analizadas para se determinar quais foram

os métodos que atingiram performance ótima. Esses métodos selecionados então são acrescidos de suas respectivas incertezas e um conjunto de PDFs é gerado, casa uma sendo contruída por um conjunto diferente de pesos relativos associados às componentes dos métodos. Por fim, torna-se possível selecionar a melhor solução de todos os métodos randomizados. [7]

4 Base de dados

A base de dados utilizada é a do SDSS (*Sloan Digital Sky Survey*) [3]. Essa base de dados consiste, basicamente, de tabelas cujas colunas representam as variáveis observadas e as linhas são correspondentes cada uma a um objeto observado. Cada objeto observado é identificado com um dado de uma coluna chamada RA (Ascensão Reta) e de uma coluna chamada DEC (Declinação), o que é melhor explicado na Seção 4.1. Para cada objeto, temos também incluídos os valores das magnitudes m_u, m_g, m_r, m_i, m_z e do *redshift* espectroscópico z_{esp}.

5 Resultados

Após realizar o treinamento no modo *Single Regression*, obteve-se o gráfico mostrado na Figura 1 de z_{best} (z fotométrico) vs z_{true} (z espectroscópico). O coeficiente angular da reta vermelha feita após o ajuste linear foi de aproximadamente 0.99. Logo, o resultado do treinamento foi bom, dado que o resultado ideal seria a reta $z_{best} = z_{true}$.

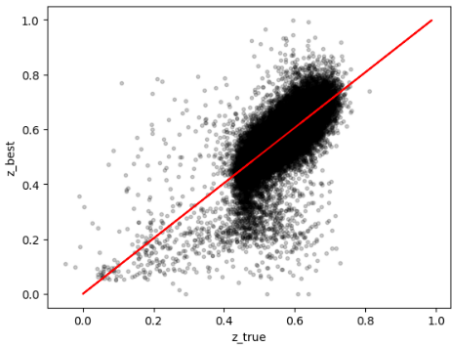

Figure 1: Gráfico z_best (z fotométrico) vs z_true (z espectroscópico) após treinamento do *ANNz2* no modo *Single Regression*.

Após realizar o treinamento no modo *Random Regression*, obteve-se o gráfico mostrado na Figura 2 de z_{best} (z fotométrico) vs z_{true} (z espectroscópico). O coeficiente angular da reta vermelha feita após o ajuste linear foi de aproximadamente 0.9. Logo, o resultado do treinamento foi bom, dado que o resultado ideal seria a reta $z_{best} = z_{true}$.

Outras informações

Durante a execução do projeto, ocorreram alterações nas metas e objetivos, uma vez que somente foi possível testar os modos de treinamento *Single Regression* e *Random Regression*. Além disso, foi encontrado um obstáculo técnico. O código ANNz (referência [3], disponível no link http://www.ast.cam.ac.uk/ aac) não está mais disponível para download. Desse modo, optamos por utilizar o código *ANNz2* (referência [4]), do *paper ANNz2 - photometric redshift and probability distribution function estimation using machine learning*, que se trata de uma implementação mais recente do código *ANNz* anterior (referência [3]). O código está disponível para download com um guia para instalação no link https://github.com/IftachSadeh/ANNZ.Assim, apesar de esse desafio

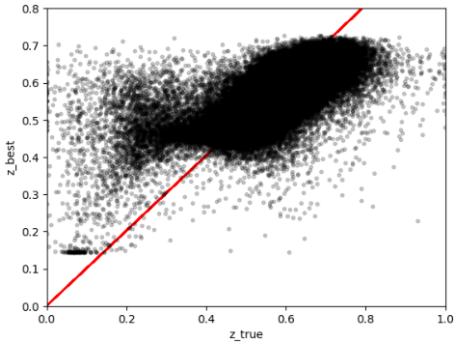

Figure 2: Gráfico z_{best} (z fotométrico) vs z_{true} (z espectroscópico) após treinamento do *ANNz2* no modo *Random Regression*.

ter surgido, foi possível contorná-lo com tranquilidade utilizando um *paper* mais recente sobre o mesmo tema.

O cronograma de atividades propunha fazer a revisão bibliográfica completa em setembro e outubro, estudar redes neurais e a estrutura interna do *ANNz2* em novembro, dezembro e janeiro; e selecionar amostras de galáxias com *redshifts* espectroscópicos conhecidos e realizar o treinamento de redes neurais em fevereiro e março. Todas essas atividades foram realizadas completamente, exceto pela última que ainda está em andamento no mês de março.

Agradecemos o apoio financeiro oferecido pela PIBIC na forma de Bolsa de Iniciação científica.

References

[1] OLIVEIRA, Kepler de. Astronomia e Astrofísica. 3ª ed. São Paulo: Editora Livraria da Física, 2014;

[2] NEWMANN, Jeffrey A., GRUEN, Daniel. Photometric Redshifts for Next-Generation Surveys. June, 2022. Disponível em: https://arxiv.org/pdf/2206.13633.pdf

[3] COLLISTER, Adrian A., LAHAV, Ofer. ANNz: Estimating Photometric Redshifts Using Artificial Neural Networks. University of Cambridge, Cambridge, UK, p. (1,6), February, 2004. Disponível em: https://arxiv.org/pdf/astro-ph/0311058.pdf

[4] SADEH, I., ABDALLA, F. B., LAHAV, O. ANNz2 - Photometric Redshift and Probability Distribution Function Estimation Using Machine Learning. University College London, UK; Rhodes University, PO, p. (1,22), June, 2016. Disponível em: https://arxiv.org/abs/1507.00490

[5] D.G. York et al. The Sloan Digital Sky Survey: Tecnical Summary. AJ, 120:1579, 2000.

[6] B. Abareshi et al. Overview of the Instrumentation for the Dark Energy Spectroscopic Instrument. AJ, 164:207.

[7] Código e documentação do ANNz2 disponível para instalação. Disponível em: https://github.com/IftachSadeh/ANNZ

[8] IVEZIC, Z., CONNOLLY, A., VANDERPLAS, J., GRAY, A. Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data. Princeton University Press, 2014.

[9] ROOT - *Data Analysis Framework*, 2024. Disponível em: [https://root.cern/manual/tmva/]. Acesso em: 9, março, 2024.