

Estimativa do redshift fotométrico de galáxias utilizando redes neurais artificiais

Sofia Garcia Telles Brito, Prof^a. Dr^a Flávia Sobreira

Universidade Estadual de Campinas

Resumo

Trabalho apresentado no evento As Astrocientistas: III Encontro Brasileiro de Meninas e Mulheres da Astrofísica, Cosmologia e Gravitação, realizado entre os dias 31 de março e 4 de abril de 2025.

Abstract

This work was presented at the event "The Astro-Scientists: III Brazilian Meeting of Girls and Women in Astrophysics, Cosmology and Gravitation", held between March 31 and April 4, 2025.

Palavras-chave: Galáxias, Redshift, Redes Neurais

Keywords: Galaxies, Redshift, Neural Networks

DOI: [10.47456/Cad.Astro.v7nEspecial.53390](https://doi.org/10.47456/Cad.Astro.v7nEspecial.53390)

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 fótons 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) [2] e Dark Energy Spectroscopic Instrument (DESI) [3] 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 quanti-

dade 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 artigo "*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 artificiais.

1.1 Definições

Aprendizado de Máquina (do inglês, *Machine Learning*) se refere a um conjunto de técnicas de construção de modelos, a partir de entradas (*inputs*) amostrais, cujo objetivo é fazer previsões. Alguns exemplos dessas técnicas são métodos de regressão, métodos de classificação (supervisionada e não-supervisionada), estimadores de máxima verossimilhança e o método Bayesiano [5].

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.

Caracterização é a habilidade de abarcar as propriedades da distribuição de um conjunto de galáxias [6].¹

A **Função Densidade de Probabilidade**, $h(x)$, quantifica a probabilidade de que um valor esteja entre x e $x + dx$, que é igual a $h(x)dx$ [5]. A PDF do *redshift*, $p(z)$, nos dá a melhor representação do resultado de um algoritmo de *redshift* fotométrico [6].

2 ANNz2

No ANNz2, o conjunto de dados original utilizado (do SDSS - *Sloan Digital Sky Survey*) é dividido em treinamento, validação e testagem, usados, respectivamente, para derivar o mapeamento entre entradas e saídas, estimar a convergência durante o treinamento, e avaliar a performance final do modelo. [4]

Os dados de entrada coletados do SDSS são as magnitudes m_u , m_g , m_r , m_i , m_z . O dado de saída é a estimativa feita do *redshift* fotométrico [4]. O valor do *redshift* espectroscópico serve como valor real do *redshift*, para comparar com o valor de saída encontrado para o *redshift* fotométrico.

No ANNz2, os métodos de Aprendizado de Máquina utilizados estão implementados no pacote TMVA do ROOT. O ROOT é uma plataforma (*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 Aprendizado de Máquina, como Redes Neurais Artificiais (do inglês, *Artificial Neural Networks* - ANN), Perceptron Multicamadas (do inglês, *Multilayer Perceptron*), Árvores de Decisão com Boosting (do inglês, *Boosted Decision Trees* - BDT), Máquinas de Vetores de Suporte (do inglês, *Support Vector Machines*) e outros [7]. Os métodos que foram considerados mais adequados foram Redes Neurais Artificiais (ANNs) e Árvores de Decisão com Boosting (BDTs) [4].

2.1 Redes neurais artificiais (ANNs)

Trata-se de um mapeamento entre o conjunto de variáveis de entrada (como magnitudes e cores)

a uma ou mais variáveis de saída, que é feito calculando a soma com pesos da coleção de funções de resposta (*response functions*). As variáveis de entrada, as funções de resposta e as variáveis de saída são chamadas de neurônios [4]. No ANNz2 a ANN utilizada foi a *Multilayer Perceptron*. Nessa rede, os neurônios são organizados em pelo menos 3 camadas: entrada, ocultos e saída. A arquitetura desses neurônios pode ser visualizada na Figura 1, à esquerda.

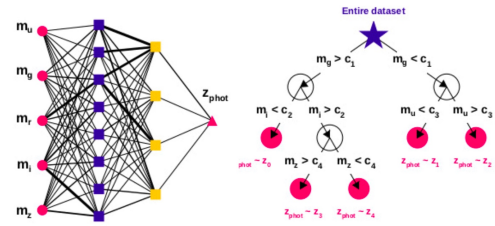


Figura 1: À esquerda se encontra uma representação da Rede Neural Artificial do tipo *Multi-Layer Perceptron* utilizada no artigo em revisão. As variáveis de entrada se tratam das magnitudes UGRIZ e a variável de saída se trata do *redshift* fotométrico [4]. Já à direita, uma representação de Árvores de Decisão com Boosting (*Boosted Decision Trees* - BDTs) utilizada também no artigo em revisão.

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

2.2 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 saída da árvore são chamados de folhas (*leaves*). Uma representação esquemática desse método pode ser vista na Figura 1, à direita.

2.3 Definição de métricas e notações

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

1. Viés fotométrico (δ): $\delta = z_{phot} - z_{spec}$ [4];
2. Espalhamento fotométrico (σ): desvio padrão de δ para um conjunto de galáxias [4];

¹As definições de Performance e Caracterização estão dadas no contexto do *redshift* fotométrico de galáxias.

3. 68º percentil de meia-largura da distribuição do viés fotométrico (σ_{68}): denota a meia-largura da área que abrange o pico do 68º percentil da distribuição de δ [4];
4. Fração de *outliers* ($f(\alpha\sigma)$): definida como a porcentagem de objetos que têm viés maior que um fator, α , de σ ou σ_{68} [4];
5. Coeficiente de determinação (R^2):

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2},$$

onde $\bar{y}$ é a média dos valores observados, y_i são os valores observados e $\hat{y}_i$ são os valores preditos. Observe que quanto mais próximo R^2 estiver de 1, melhor o ajuste feito pelo modelo.

2.4 Modos de treinamento do ANNz2

O ANNz2 utiliza tanto técnicas de regressão quanto técnicas de classificação. Os modos de treinamento do ANNz2 são Regressão Simples (do inglês, *Single Regression*), Regressão Aleatória (do inglês, *Random Regression*), Classificação por faixas (do inglês, *Binned Classification*), Classificação Simples (do inglês, *Single Classification*) e Classificação Aleatória (do inglês, *Random Classification*). Neste trabalho, foram testados apenas os modos de Regressão Simples (*Single Regression* - SR) e de Regressão Aleatória (*Random Regression* - RR), que estão descritos nas subseções a seguir.

2.4.1 Regressão Simples (*Single Regression*)

Essa é a configuração mais simples do ANNz2. Nesse caso, uma única regressão é feita [8].

2.4.2 Regressão Aleatória (*Random Regression*)

Na Regressão Aleatória, um conjunto de métodos de regressão é automaticamente gerado, sendo que esses métodos de Aprendizado de Máquina 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 analisadas 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, cada uma sendo construída 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. [8]

3 Metodologia

3.1 Base de dados

A base de dados utilizada foi a do SDSS (*Sloan Digital Sky Survey*) [9], especificamente os dados fotométricos e espectroscópicos de galáxias do Release 10 (DR10), que foram acessados utilizando a interface CasJobs¹. O conjunto de dados foi extraído usando uma consulta (*query*) SQL contendo apenas as galáxias com *redshifts* espectroscópicos válidos ($z_{\text{Warning}} = 0$) e pertencente ao intervalo $0.02 \leq z_{\text{spec}} \leq 0.8$, com magnitude na banda r entre 14 e 22 e com erros nas magnitudes UGRIZ entre 0.001 e 0.5. Foram extraídos os dados fotométricos e espectroscópicos fazendo a correspondência com os números identificadores dos objetos ($s.\text{bestObjID} = p.\text{objID}$). Essa consulta resultou em um conjunto de 500 mil objetos, que, após cortes de qualidade (como remoção de duplicatas e de valores NaN), passou a conter 499985 objetos. A fim de evitar um alto custo de tempo computacional, foi selecionado randomicamente 20% do conjunto total de objetos, mantendo a representatividade da distribuição original de *redshifts*. Esse novo conjunto, de cerca de 100 mil objetos, foi subdividido entre conjunto de treino (60%, 60 mil objetos), conjunto de validação (20%, 20 mil objetos) e conjunto de teste (20%, 20 mil objetos).

Os dados fotométricos coletados foram as magnitudes UGRIZ (m_u , m_g , m_r , m_i , m_z), seus erros associados e quatro cores fotométricas, calculadas como diferenças entre bandas adjacentes ($color_{X-Y} = m_X - m_Y$), com erros propagados por $\sigma_{color_{X-Y}} = \sqrt{\sigma_{m_X}^2 + \sigma_{m_Y}^2}$. Das 18 colunas do arquivo de entrada, foram utilizadas

¹<https://skyserver.sdss.org/casjobs/>

como características (*features*) no treinamento as 5 magnitudes e as 4 cores, totalizando 9 variáveis. As cores não foram utilizadas no artigo original, tendo sido incluídas neste trabalho como aperfeiçoamento da metodologia.

3.2 Treinamento, otimização e validação

O código ANNz2 (versão 2.3.2, ROOT versão 6.36.04) foi utilizado para estimar o *redshift* fotométrico através de dois dos métodos de treinamento disponibilizados pelos autores: SR e RR. Foi usado o modo ANN (Redes Neurais Artificiais), sendo que para o modo SR foi usado um único MLM (Método de Aprendizado de Máquina) e para o modo de RR foram usados 50 MLMs. Além disso, as métricas de performance utilizadas seguem as definições da Seção 2.3.

4 Resultados

4.1 Correlação entre z_{phot} e z_{spec}

Após realizar os treinamentos nos modos com SR e RR, foram construídos os gráficos necessários para avaliar os resultados.

Os gráficos (a) das Figuras 2 e 3 apresentam as correlações entre z_{phot} e z_{spec} para SR e RR, respectivamente, e os ajustes lineares resultaram em

$$z_{phot} = 0.933z_{spec} + 0.011 \quad (R^2 = 0.9349),$$

$$z_{phot} = 0.941z_{spec} + 0.010 \quad (R^2 = 0.9395),$$

demonstrando um excelente poder preditivo dos modelos, uma vez que os valores de R^2 e os coeficientes angulares se aproximam de 1 e os valores dos coeficientes lineares se aproximam de 0.

4.2 Métricas

Nos gráficos (b), (c) e (d) das Figuras 2 e 3 e na Tabela 1, podemos visualizar as métricas obtidas.

Tabela 1

Modo	$\langle \delta \rangle$	σ	σ_{68}	$f(2\sigma_{68})$	$f(3\sigma_{68})$
SR	0.001	0.03	0.019	0.101	0.039
RR	0.001	0.03	0.018	0.102	0.039

Os gráficos (b) nas Figuras 2 e 3 mostram o viés fotométrico δ_b por intervalo de *redshift*. Para os intervalos em z_{spec} , tanto para SR quanto para

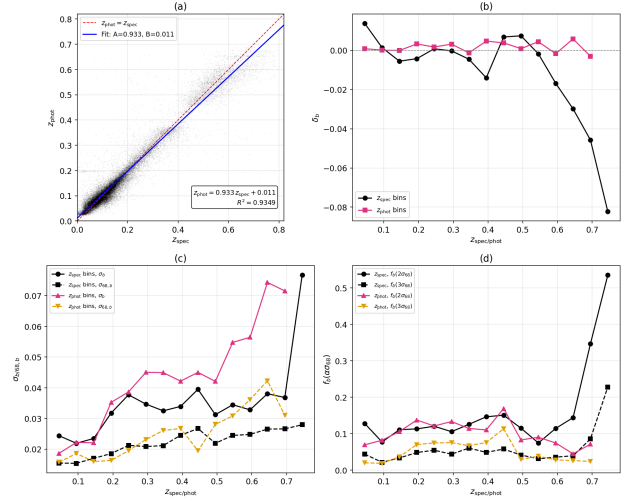


Figura 2: Resultados do treinamento utilizando ANNs para o modo SR.

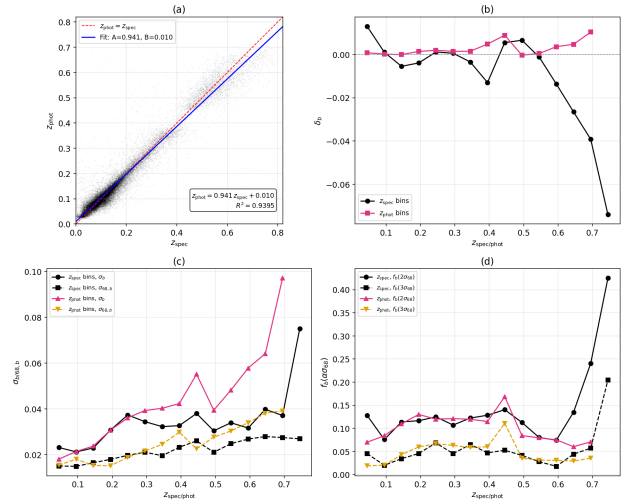


Figura 3: Resultados do treinamento utilizando ANN para o modo RR.

RR, o viés permanece próximo de zero aproximadamente no intervalo de $0.1 \leq z \leq 0.5$, porém apresenta uma queda significativa para $z \geq 0.6$, o que indica que os modelos subestimam o *redshift* de galáxias com z_{spec} elevado. Os intervalos em z_{phot} apresentam viés menor e mais estável, próximo de zero, por todo o intervalo de *redshift*.

Os gráficos (c) mostram que σ_{68} se mantém estável para $z < 0.6$, com aumento acentuado acima desse valor. Os gráficos (d) mostram comportamento similar para as frações *outliers* $f(2\sigma_{68})$ e $f(3\sigma_{68})$ nos intervalos (*bins*) em z_{spec} , enquanto nos intervalos em z_{phot} essas frações permanecem estáveis e abaixo de 0.2 em todo o intervalo.

4.3 Comparação com trabalho original

O artigo original [4] apresenta na Figura 10 os valores de $\langle\delta\rangle$, σ_{68} , $f(2\sigma_{68})$ e $f(3\sigma_{68})$ para z_{phot} , presentes na Tabela 2. É possível perceber que os valores de $f(2\sigma_{68})$ e $f(3\sigma_{68})$ foram muito similares aos obtidos na Tabela 1. Já os valores de δ e σ_{68} foram diferentes dos obtidos na Tabela 1, sendo δ do artigo mais baixo e σ_{68} mais alto.

Tabela 2

$\langle\delta\rangle$	σ_{68}	$f(2\sigma_{68})$	$f(3\sigma_{68})$
0.0002	0.031	0.1	0.04

5 Conclusão

Neste trabalho, foi realizada uma revisão do artigo “ANNz2 - Photometric Redshift and Probability Distribution Function Estimation Using Machine Learning” [4], na qual foram explorados, na opção de ANN (Redes Neurais Artificiais), dois modos de treinamento, Regressão Simples (*Single Regression*) e Regressão Aleatória (*Random Regression*). Foi utilizada uma amostra de galáxias do *Sloan Digital Sky Survey*, Release 10 (DR10), com faixa de *redshift* $0.02 \leq z \leq 0.8$.

Os resultados obtidos demonstram que ambos os modos de treinamento apresentam excelente poder preditivo, com coeficientes de determinação $R^2 > 0.93$ e $\sigma_{68} < 0.02$. O modo RR apresentou performance ligeiramente superior ao de SR, comportamento consistente com o esperado dada a maior simplicidade do último.

Como trabalho futuro, propõe-se a comparação entre os métodos de aprendizado de máquina Árvores de Decisão com Boosting (BDTs) e Redes Neurais Artificiais (ANNs) nos modos SR e RR, bem como a exploração dos demais modos de treinamento disponíveis no ANNz2: Classificação por faixas (*Binned Classification*), Classificação Simples (*Single Classification*) e Classificação Aleatória (*Random Classification*).

6 Agradecimentos

Agradecemos à Pró-Reitoria de Pesquisa (PRP) da Universidade Estadual de Campinas (UNICAMP) pela contemplação desse projeto de

Iniciação Científica no Programa Institucional de Bolsas de Iniciação Científica (PIBIC) e pelo financiamento concedido.

Sobre as autoras

Sofia Garcia Telles Brito (sofiagarcia@unifil.com.br) é Bacharela em Matemática Aplicada e Computacional pela Universidade Estadual de Campinas (UNICAMP) e possui interesse em Astronomia, Cosmologia, Aprendizado de Máquina e Processamento de Imagem. (Currículo Lattes: <https://lattes.cnpq.br/9686182889855838>)

Dra. Flavia Sobreira (sobreira@ifi.unicamp.br) é professora e pesquisadora da Unicamp. Trabalha na área de Cosmologia com ênfase na pesquisa sobre a aceleração do universo. (Currículo Lattes: <https://lattes.cnpq.br/2432352931378666>)

Referências

- [1] K. d. Oliveira, *Astronomia e Astrofísica* (Editora Livraria da Física, São Paulo, 2014), 3 ed.
- [2] D. G. York et al., *The Sloan Digital Sky Survey: Technical Summary*, *Astronomical Journal* **120**, 1579 (2000).
- [3] B. Abareschi et al., *Overview of the Instrumentation for the Dark Energy Spectroscopic Instrument*, *Astronomical Journal* **164**, 207 (2022).
- [4] I. Sadeh, F. B. Abdalla e O. Lahav, *ANNz2 - Photometric Redshift and Probability Distribution Function Estimation Using Machine Learning*, arXiv:1507.00490 1–22 (2016). University College London, UK; Rhodes University.
- [5] Z. Ivezić et al., *Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data* (Princeton University Press, 2014).
- [6] J. A. Newman e D. Gruen, *Photometric Redshifts for Next-Generation Surveys*, arXiv:2206.13633 (2022). Disponível em: <https://arxiv.org/pdf/2206.13633.pdf>.

- [7] ROOT Collaboration, *ROOT - Data Analysis Framework*, ROOT TMVA Manual (2024). Disponível em: <https://root.cern/manual/tmva/>. <https://github.com/IftachSadeh/ANNZ>.
- [8] I. Sadeh, *ANNz2 Code and Documentation*, GitHub repository (2024). Disponível em:
- [9] A. A. Collister e O. Lahav, *ANNz: Estimating Photometric Redshifts Using Artificial Neural Networks*, arXiv preprint astro-ph/0311058 1–6 (2004). University of Cambridge, Cambridge, UK.