4300430: Introdução à Cosmologia Física

Final Project: Photometric Redshifts (Photo-zs)

(2024)

In this project you will compute photometric redshifts (photo-zs) from the magnitudes of galaxies, using an Artificial Neural Network (ANN) code. Read the paper by Collister & Lahav 2004 (https://arxiv.org/abs/astro-ph/0311058) and download the corresponding `ANNz` code (http://www.homepages.ucl.ac.uk/ ucapola/annz.html). Read the README file, compile the code and run it on the example file.

Download the following files for training sets here and here and validation sets here and here (to be used for training the network). Download the testing set here (to be used for applying the network solution and quantifying photo-z errors). All these files contain magnitudes (u, g, r, i, z), magnitude errors $(\sigma_u, \sigma_g, \sigma_r, \sigma_i, \sigma_z)$ and spectroscopic redshift $(z^{\rm spec})$.

Basic Steps:

- Start using the **representative** training and validation files above to train and validate the network solution.
- Run the code to train and validate your photo-z solution and apply this solution to the testing set to compute photo-zs $z_i^{\rm phot}$ for $i = 1...N$ where N is the number of galaxies in the testing set.
- Define metrics to quantify the quality of the solution, e.g. the photo-z bias $z^{\rm bias}$

$$z^{\rm bias} = \frac{1}{N}\sum_{i=0}^{N}(z_i^{\rm phot} - z_i^{\rm spec}) \tag{1}$$

and the photo-z scatter σ_z

$$\sigma_z^2 = \frac{1}{N-1}\sum_{i=0}^{N}(z_i^{\rm phot} - z_i^{\rm spec})^2 \tag{2}$$

- Plot z^{phot} versus z^{spec} and indicate in the plot the $y = x$ line as well as the values of the metrics above.

- The metrics above can also be defined in bins of z^{spec} or z^{phot}. Do this calculation in multiple bins of z^{spec} and z^{phot} and plot z^{bias} and σ_z as a function of z^{spec} and z^{phot}.

- Plot on a single figure the redshift distributions $N(z^{\text{spec}})$ and $N(z^{\text{phot}})$ as histograms for the testing set in bins of redshift (e.g. 20 bins of $\Delta z = 0.1$ from $z = 0$ to $z = 2$).

Investigate how metrics vary as a function of the following changes:

- Changes in the number of multiple network solutions from different seeds that get combined into the final photo-z solution. Until now you had probably used a single solution.

- Changes in the network configuration, by changing the number of hidden layers and the number of nodes in these hidden layers.

- Changes in the size of the training and validation set used in the training/validation process. Consider fractions of these sets, e.g. 100%, 70%, 50%, 20%, 5% of the points.

- Changes in the maximum redshift used in the training/validation process.

- For the best case found, repeat the computation for **non-representative** training and validation sets. This represents a more realistic situation for real data analysis.

- Up to this point, you used `ANNz` to output a single number, i.e the galaxy photo-z z^{phot}. `ANNz` can also produce multiple outputs, for instance multiple points in the *redshift distribution* $p(z_i)$ of a single galaxy. The overall redshift distribution $N(z)$ can be estimated by summing $p(z_i)$'s of all galaxies. Think of how you could use the training and validation sets to estimate $p(z_i)'s$ with `ANNz`. (*Hint:* What is the redshift distribution of a perfect spectroscopic redshift z^{spec} in the training set?)