Statistical Challenges in 21st Century Cosmology
Proceedings IAU Symposium No. 306, 2014
A. F. Heavens, J.-L. Starck & A. Krone-Martins, eds.

doi:10.1017/S1743921314010849

ANNz2 - Photometric redshift and probability density function estimation using machine-learning

Iftach Sadeh

Astrophysics Group, Department of Physics and Astronomy, University College London,
Gower Street, London WC1E 6BT, United Kingdom
email: `i.sadeh@ucl.ac.uk`

Abstract. Large photometric galaxy surveys allow the study of questions at the forefront of science, such as the nature of dark energy. The success of such surveys depends on the ability to measure the photometric redshifts of objects (photo-zs), based on limited spectral data. A new major version of the public photo-z estimation software, `ANNz`, is presented here. The new code incorporates several machine-learning methods, such as artificial neural networks and boosted decision/regression trees, which are all used in concert. The objective of the algorithm is to dynamically optimize the performance of the photo-z estimation, and to properly derive the associated uncertainties. In addition to single-value solutions, the new code also generates full probability density functions in two independent ways.

Keywords. techniques: photometric, galaxies: distances and redshifts

1. Introduction

The different approaches to calculate photometric redshifts (photo-zs) can generally be divided into two categories, template fitting methods and empirical machine-learning (see Hildebrandt *et al.* (2010) for a review of the field). Template fitters employ physically motivated models. On the other hand, Machine-learning methods (MLMs) involve deriving the relationship between the photometric observables and the redshift using a so-called training dataset, which includes both the observables and precise redshift information.

This paper presents `ANNz2`†. (Sadeh *et al.* (2015)), a new implementation of the popular code of Collister & Lahav (2003), which uses artificial neural networks (ANNs) to estimate photometric redshifts. The new code incorporates a variety of machine-learning techniques, such as boosted decision/regression trees (BDTs) in addition to ANNs. The various MLMs are provided by the `TMVA` package (Hoecker *et al.* (2007)). It has been designed to calculate both photometric redshifts and probability distribution functions (PDFs), doing so in several different ways. The introduction of photo-z-PDFs has been shown to improve the accuracy of cosmological measurements (Mandelbaum *et al.* (2008)), and is an important feature of the new version of `ANNz`.

2. Description of the operational modes of `ANNz2`

`ANNz2` uses both regression and classification techniques for estimation of single-value photo-z solutions and PDFs. The different configurations are referred to as *single regression*, *randomized regression* and *binned classification*. A short description follows.

† Code publicly available at https://github.com/IftachSadeh/ANNZ

https://doi.org/10.1017/S1743921314010849 Published online by Cambridge University Press

2.1. *Single regression (single-value photo-z estimator)*

In the simplest configuration of `ANNz2`, a single regression is performed, using as the output the spectroscopic redshift, denoted hereafter by $z_{\rm spec}$. Consequently, one may derive a per-galaxy photo-z solution. Associated errors are generated using the nearest neighbours error estimation method, discussed by Oyaizu *et al.* (2008).

2.2. *Randomized regression (single-value photo-z and PDF solutions)*

Instead of choosing a single MLM, it is possible to automatically generate an ensemble of regression methods. The *randomized MLMs* differ from each other in several ways. This includes setting unique random seed initializations, as well as changing the configuration parameters of a given algorithm. To give an example, the latter may refer to using various types and numbers of neurons in an ANN, or to arranging neurons in different layouts of hidden layers; for BDTs, the number of trees and the type of boosting algorithm may be changed, etc. Additionally, `TMVA` provides the option to perform transformations on the input-parameters, including normalization or principal component decomposition. The option is also available to only use a subset of the input parameters, or to train with pre-defined functional combinations of parameters.

Once randomized MLMs are initialized, the various methods are each trained. Subsequently, a distribution of photo-z solutions for each galaxy is generated. A selection procedure is then applied to the ensemble of answers, choosing the subset of methods which achieve optimal performance. The selected methods are used to derive a single photo-z estimator, based on the method with the best performance, denoted by $z_{\rm best}$. The optimized solutions are also used in concert, producing an additional (averaged) single-value solution, referred to as $< z_{\rm best} >$. Finally, the ensemble of estimators are used to derive a complete probability density function, denoted by $\mathrm{PDF}_{\rm tmpl}$. This is done by folding a weighted distribution of the solutions with the corresponding uncertainty estimators of the individual MLMs. The weights are derived dynamically, with the purpose of optimizing the quality of the PDF.

2.3. *Binned classification (PDF solution)*

`ANNz2` may also be run in classification-mode, employing an algorithm similar to that used by Gerdes *et al.* (2010). The first step of the calculation involves dividing the redshift range of the input samples into many small bins. Within the redshift bounds of a given bin, the *signal sample* is defined as the collection of galaxies for which $z_{\rm spec}$ is within the bin. Similarly, the *background sample* includes all galaxies with $z_{\rm spec}$ outside the confines of the bin.

The algorithm proceeds by training a different classification MLM for each redshift bin. The output of a trained method in a given bin is translated to the probability for a galaxy to have redshift which falls inside that bin. The distribution of probabilities from all of the bins is normalized to unity, accounting for possible varying bin width. It then stands as the photo-z-PDF of the galaxy, denoted in the following as $\mathrm{PDF}_{\rm cls}$.

3. Performance of the algorithm

We use data included in data releases 10, DR10, (Ahn *et al.* (2014)) of the Baryon Oscillation Spectroscopic Survey (BOSS) (Dawson *et al.* (2013)), which is part of the Sloan Digital Sky Survey (SDSS), from its current incarnation, SDSS-III (Eisenstein *et al.* (2011)). The data are publicly available as a FITS catalogue, containing flux measurements, which were transformed into magnitudes. Objects in the catalogue were subjected to quality cuts, ensuring that only galaxies with flux measurements in all five bands, `u`,

https://doi.org/10.1017/S1743921314010849 Published online by Cambridge University Press

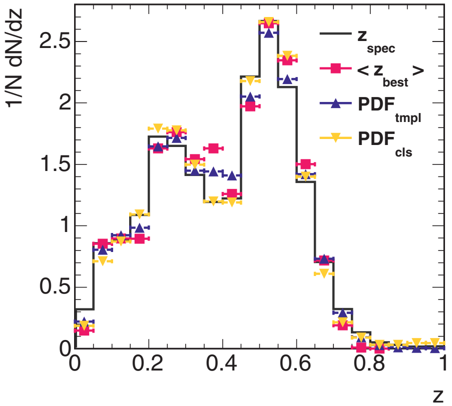

Figure 1. Differential distributions, denoted collectively by z, of the spectroscopic redshift, $z_{\rm spec}$, and of the respective photometric redshift estimated using ANNz2 and described in the text, $< z_{\rm best} >$, $\rm PDF_{tmpl}$ and $\rm PDF_{cls}$, as indicated.

g, r, i and z, which had magnitude errors below 1, were accepted. The five magnitudes served as inputs for training in ANNz2, along with the spectroscopic redshift, which was taken as the true redshift value. The entire sample, consisting of $\sim 150,000$ galaxies, was split into three sub-samples for training, testing and validation purposes; the first sub-sample was used to train the MLMs; the second was used to optimized the performance and derive PDF-weights; using the third sub-sample, the performance was validated by comparing the spectroscopic redshift distribution, with that of the photometric solutions provided by ANNz2.

For this study, ANNz2 was run in randomized regression mode, using 100 ANNs, and in binned classification mode, using collections of BDTs. The results are shown in Fig. 1. One may observe that the PDF solutions provide a better description of the spectroscopic redshift distribution, compared to the single-value photo-zs. Additional tests showing the per-galaxy bias, scatter, outlier fraction and other metrics were performed, and are presented in Sadeh *et al.* (2015).

References

Ahn, C. P., *et al.* 2014, *ApJS*, 211, 17

Collister & Lahav 2003, *Publ.Astron.Soc.Pac.*, 116, 345-351

Dawson, K., *et al.* 2013, *Astron.J.*, 145, 10

Eisenstein, D. J., *et al.* 2011, *Astron.J.*, 142, 72

Gerdes, D., *et al.* 2010, *Astrophys.J.*, 715, 823-832

Hildebrandt, H., *et al.* 2010, *A&A*, 523, A31

Hoecker, A., *et al.* 2007, *PoS*, ACAT, 040

Mandelbaum, R., *et al.* 2008, *Mon.Not.Roy.Astron.Soc.*, 386, 781-806

Oyaizu, H., *et al.* 2008, *Astrophys.J.*, 689, 709-720

Sadeh, I., *et al.* 2015, In preparation

https://doi.org/10.1017/S1743921314010849 Published online by Cambridge University Press