RNAAS RESEARCH NOTES OF THE AAS

SpyderZ: An Efficient Support Vector Machine Library for Photometric Redshift Estimation and Redshift Probability Information

To cite this article: Vikhyat Agarwal *et al* 2024 *Res. Notes AAS* **8** 126

Draft version May 5, 2024
Typeset using LaTeX default style in AASTeX631

SpyderZ: an efficient support vector machine library for photometric redshift estimation and redshift probability information

Vikhyat Agarwal [1], Jack Singal [1], and Christine Gyure[1]

[1] *University of Richmond*
410 Westhampton Way
Richmond, VA 23173

ABSTRACT

We present SpyderZ, a Python-based library for photometric redshift estimation using support vector machines (implemented with scikit-learn). Our approach discretizes redshift values into uniformly-sized bins and uses one-vs-one support vector classifiers with voting strategies to produce effective probability density functions (ePDFs) over redshift for each galaxy. These ePDFs, which are not constrained to be Gaussian or any other shape, allow for our model's predictions to be used quantitatively with uncertainty analysis methods, and have been shown to enable reliable catastrophic outlier detection. Adapted from the previous IDL package SpiderZ, SpyderZ offers training and evaluation speed optimizations on the order of 10^2, along with support for parallelization across CPU cores. Our library also offers in-built data sanity checks, result visualizations, metric calculations, cross validation, batch evaluations, and parallelized hyperparameter search (grid search and random search).

Keywords: Redshift surveys(1378) — Photometry(1234) — Support vector machine (1936)

1. INTRODUCTION

Redshift (denoted z) is a key property of a galaxy that encodes its distance from us, as well as how far back in time it is from when we observe it. Traditionally, redshift is determined by observing redshifted absorption and emission lines in the source's spectrum. However such time-consuming spectroscopic measurements are not possible for large scale surveys which observe millions of galaxies. As an alternative, photometric redshift estimation (photo-z) utilizes the measured flux of sources in a small number of broadband filters, each covering a subrange of the spectral energy distribution (SED).

Several machine learning (ML) approaches have been used to obtain photo-z estimations, as discussed in Salvato et al. (2019). Popular ML models include artificial neural networks, random forests, and support vector machines (SVMs). In previous work, (Jones & Singal 2017) created SpiderZ, a support vector machine (SVM) library for photo-z estimation. SpiderZ provided excellent redshift point estimate accuracy, but also, importantly, naturally provided an effective probability distribution function (ePDF) over redshift for each galaxy, as discussed extensively in (Jones & Singal 2020) and (Wyatt & Singal 2021). This was implemented in IDL, a data analysis language previously popular within astrophysics, but now increasingly lacking support and documentation.

We present SpyderZ, an adaption of SpiderZ, implemented in Python using the machine learning library scikit-learn (Pedregosa et al. 2011) to utilize modern computational optimizations like parallelization and memory-caching.

2. METHODS

In SpyderZ, we divide the range of feasible z values (e.g.: $0 < z < 4$) for a dataset into equal sized bins of fixed width (e.g.: 0.1). Each of these bins is a class, and we train the SVM classifiers to predict which class a given galaxy belongs to. Each bin of redshift is paired against another bin of redshift, and the responsible classifier choses which bin a datapoint is more likely to be in, based on which side of the separating hyperplane it lies.

By counting the 'votes' for each bin across all classifiers, SpyderZ naturally determines an effective probability distrubution (EPDF) over redshift for each galaxy. The most probable bin can be selected as the discrete photo-z estimate, but the EPDF contains valuable information for uncertainty and outlier flagging.

Features of SpyderZ:

- object oriented interface with easy model initialization, training, and evaluation
- parallelizes batch training and prediction on the n(n-1)/2 separate classifiers across CPU threads
- generates ePDFs, which can be used for further outlier flagging methods
- generates result visualizations
- calculates relevant photo-z estimation metrics, as per (Ilbert et al. 2006)
- enables hyperparameter search for C and gamma

3. RESULTS

We train and evaluate SpyderZ using a data set of 286,401 galaxies with five-band ugriz photometry and reliable spectroscopic redshifts from the Hyper-Suprime Cam (HSC) Public Data Release 2 by (Aihara et al. 2019). We use a training set of 4,000 randomly chosen galaxies, and the model is evaluated on the other 282,401 galaxies.

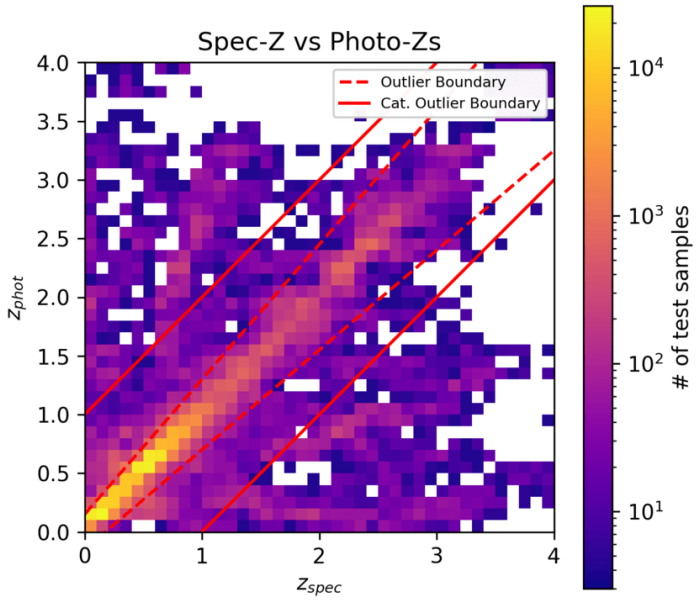

Figure 1. Density plot of photo-z point estimates (taken as the highest probability redshift bin of the EPDF) as determined by SpyderZ, versus the actual redshifts for the HSC test data set.

We get the following performance metrics in our evaluation, when using the error metric $E = \frac{z_{phot} - z_{spec}}{1 + z_{spec}}$:

- Normalised Median Absolute Deviation (NMAD) $= 1.4826 \cdot \text{median}(|E - \text{median}(E)|) =$ **0.049**
- Root Mean Square Error (RMSE) $= \sqrt{\sum E_i^2} =$ **0.222**
- Bias $= \text{median}(z_{phot} - z_{spec}) =$ **-0.005**
- Outlier percentage (galaxies for which $|E| > 0.15$) $=$ **9.5%**
- Catastrophic outlier percentage (galaxies for which $|z_{phot} - z_{spec}| > 1$) $=$ **4.9%**

The SpyderZ library code, along with the Jupyter notebook used to produce the results above, is hosted at https://github.com/Vikhyat2603/SpyderZ and preserved on Zenodo at doi:10.5281/zenodo.11111982 under an MIT license.

Software: scikit-learn (Pedregosa et al. 2011), SpyderZ (Agarwal 2024)

REFERENCES

Agarwal, V. 2024, SpyderZ: an efficient support vector machine library for photometric redshift estimation and redshift probability information, 1.0.0, Zenodo, doi: 10.5281/zenodo.11111982

Aihara, H., AlSayyad, Y., Ando, M., et al. 2019, Publications of the Astronomical Society of Japan, 71, 114

Ilbert, O., Arnouts, S., Mccracken, H. J., et al. 2006, Astronomy & Astrophysics, 457, 841

Jones, E., & Singal, J. 2017, Astronomy & Astrophysics, 600, A113

—. 2020, Publications of the Astronomical Society of the Pacific, 132, 024501

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, the Journal of machine Learning research, 12, 2825

Salvato, M., Ilbert, O., & Hoyle, B. 2019, Nature Astronomy, 3, 212

Wyatt, M., & Singal, J. 2021, Publications of the Astronomical Society of the Pacific, 133, 044504