Galaxy Finding and Sorting in the Vera Rubin Observatory's LSST and Beyond

Presentation: Prof. Ian Dell'Antonio

Workshop:

1 Overview of the Physics

In Optical and Near-infrared sensing of the Universe you typically measure the intensity of light as a function of position in the sky, frequency or wavelength, and time. However, we usually think of the information as coming from distinct objects distributed in 3-dimensional space. For the light coming from outside the Milky Way, the vast majority of these objects are galaxies. Thus, one of the major steps in astrophysical analyses is to turn our observations into sets of individual galaxies whose position, luminosity, distance, size, etc. are measured. There are two aspects to this problem. The first is to detect individual galaxies (a form of the "segmentation problem". This is very interesting and has been subject to many machine-learning approaches, but is not the topic of this module. The second aspect is to measure the distance from Earth of the galaxies. For all but the closest galaxies, the most accurate measure of distance is the redshift, obtained via spectral analysis of the emission and absorption lines. However, these measurements are hard to make, and thus distances are accurately known for a small number of galaxies. On the other hand, images taken through broad-band filters are efficient in detecting galaxies–the number of galaxies with redshifts numbers in the millions, while the number of galaxies observed in imaging is in the billions. Thus, techniques that can measure the redshift and thus distance of galaxies via imaging are extremely valuable. Deterministic algorithms exist to estimate the redshifts, but they presuppose that the data exactly matches pre-determined templates. More modern approaches use subsets of the data itself to train the measurement.

2 Module Description

2.1 Introduction

In this module we will use the multi-band imaging information and a multi-layer neural network trained on a subset of galaxies with five-band images (u,g,r,i,z, representing images spanning the wavelength to solve a slightly simpler version of the galaxy redshift determination problem. Often, it's more important to know which galaxies to include or exclude in an analysis as a first step rather than extracting the precise redshift. In this module, you'll be using a network trained via stochastic gradient descent to separate galaxies into background (galaxies that are gravitationally lensed by the mass) and potential galaxy cluster members.

2.2 AI Aspects

The network will be trained on the subset of objects with spectroscopic redshifts and multi-band magnitudes. The differences in magnitude form a set of 10 dependent variables that encode the color information, and the network determines both the "close" and "far" bin attributions. The training uses stochastic gradient descent algorithm to find the effective weights and "color" vectors that minimize the number of misclassifications in the training set. If there is time, we will also train a set of weights using a "genetic" algorithm[3]. The weights will then be applied to the galaxies with unknown distance to classify them. We'll also discuss limitations of this method.

2.3 Specifics

The input data for the module is a set of ~10000 galaxies with 5 band magnitudes (and thus 10 separate different differences to look for cross-correlations), and about 100000 galaxies behind a galaxy cluster that we want to identify. We will train our simple network and run classification on two subsets of data–one with distribution of properties matching that of the training set, and one with a sample that differs significantly, to retrieve classifications and confidence intervals and demonstrate how the network behaves in the presence of irregular data.

2.4 Module Goal

At the end of the module, you will have gained experience with algorithms for training networks and have seen how they can provide both good and "bad" results depending on the match between training and target datasets.

3 Figures

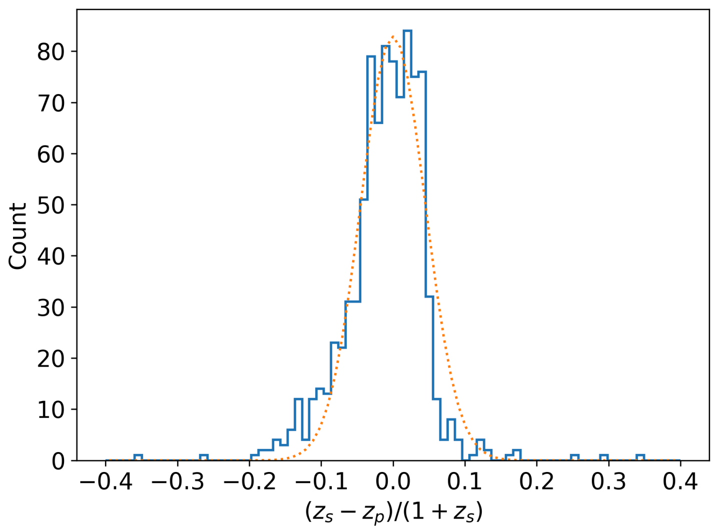

Figure 1. Difference between true and trained redshifts for a network trained on a set of clusters with spectroscopic data and applied to another cluster (data from Fu et al. 2021).

4 References

[1] Robert Hogan, Malcolm Fairbairn, and Navin Seeburn, "Gaz: a genetic algorithm for photometric determination",MNRAS,2015,**449**,2040, (https://academic.oup.com/mnras/article/449/2/2040/1079804)