TERM PROJECT PROPOSAL

TROY RAEN

1. Outline

It is well established that the light reaching our telescopes from distant galaxies is shifted toward the red end of the spectrum (relative to the frequency it was originally emitted at), and that this shift increases with the galaxy's distance from us. The combined measurements from many galaxies indicate that the universe itself (the space between galaxies) is expanding at a rate that increases with time. The precise calculation of this expansion rate is being pursued and it depends strongly on our ability to make accurate calculations/estimations of the amount by which the light from a distant galaxy is redshifted. (Actually, the calculations of all fundamental quantities in cosmology rely heavily on these redshifts.)

The calculation of the redshift from measurements of light intensity generally depends on being able to find known features in the intensity as a function of frequency. Poor frequency resolution then increases the error on an estimation of the redshift.

There are two ways in which telescopes can take measurements: spectroscopy and photometry. Spectroscopy records information about the amount of light coming in over a wide range of the frequency spectrum, at high resolution. Photometry essentially divides the spectrum into a small number of bins (on the order of 5) and records aggregated information for each bin. Thus photometry is much cheaper to do and so we have more data of this type. However, this low resolution translates into large errors on our estimates of redshift using this data. (A redshift calculated in this way is called a 'photo-z'.)

I have access, through my advisor, to a dataset consisting of simulated, photometric telescope data for $\sim 3\times10^6$ galaxies, with 3 quantities for each photometric bin (measurement of intensity of light, error on the measurement, one other that I do not yet understand). With 6 bins, this gives 24 features total. The dataset also includes the correct redshift for each galaxy.

The simulated data is intended to mimic the data anticipated from the upcoming Large Synoptic Survey Telescope (LSST). LSST will collect data from large volumes of the sky and at rates several orders of magnitude above any other telescope to date. The community is making large efforts towards dealing with data at this scale, and one of these efforts is toward quick and accurate photo-z calculations. Codes using machine learning algorithms are beginning (∼2000) to be used for these calculations.

My proposal is to study how the error in photo-z estimates scale with the size of the training set for different algorithms.

2. Learning Methods

I plan to start with 1) a neural network and 2) a regression decision tree and to quickly expand that to a random forest. Sánchez et al. (2014) review the codes in use at that time. Here I give an (incomplete) list of the types of methods used in those codes:

- artificial neural networks
- prediction trees and random forest
- boosted decision tree
- relevance vector machine
- normalized inner product nearest neighbor
- various Bayesian methods

I am particularly interested in the decision tree and Bayesian methods, but will look in to all of these options. With the current efforts, there may also be codes developed more recently that use other methods. I will do a more thorough search of the literature and also consult with my advisor about algorithms that may be particularly important for LSST.

3. Testing Solutions

Newman et al. (2019) have done photo-z error analysis on a random forest regression algorithm (see Figure 1), and found that the error scales as $1/N^{0.4}$ where N is the training sample size. The error was expected to go as $1/\sqrt{N}$, so this result was somewhat surprising. They use two estimates of the error, one is a normalized median absolute deviation and the other is a measure of the number of objects with errors greater than 10%. Photo-z estimates can be wrong by large margins, so this outlier detection is a useful statistic. I plan to use these two measures when comparing algorithms.

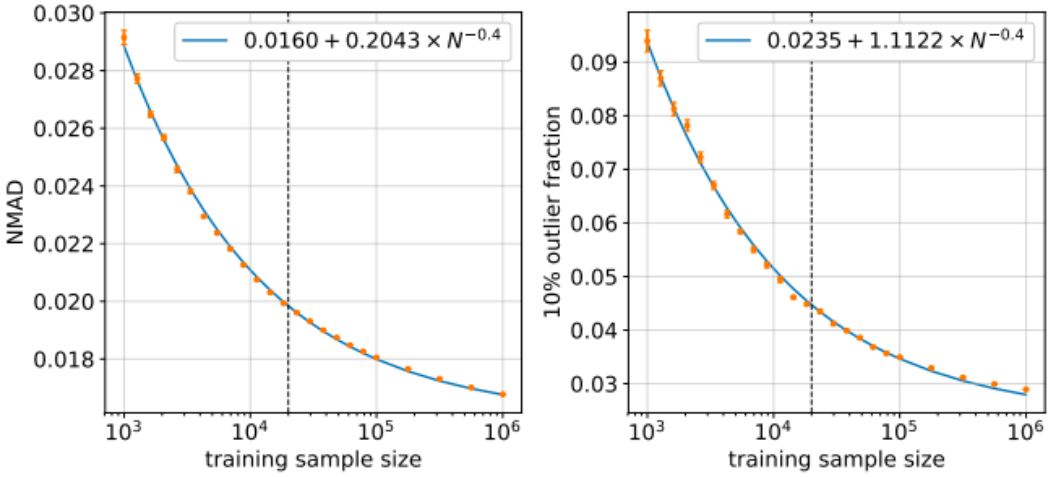

Figure 1: Orange points show photometric redshift errors and outlier rates versus the number of galaxies in the training set for galaxies with simulated LSST photometric errors. Photo-z's were calculated using a random forest regression algorithm. The left panel shows the photo-z error, quantified by the normalized median absolute deviation (NMAD) in $(z_{\rm phot} - z_{\rm spec})/(1+z_{\rm spec})$, as a function of training set size; similarly, the right panel shows the fraction of 10% outliers, i.e. objects with $|z_{\rm phot} - z_{\rm spec}|/(1+z_{\rm spec}) > 0.1$. A vertical dashed line shows the sample size for the baseline training survey from [5]. The blue curves represent simple fits to the measurements as a function of the training set size, N. This analysis uses a set of simulated galaxies from Ref. [11] that spans the redshift range of $0 < z < 4$, using a randomly-selected testing set of 10^5 galaxies for estimating errors and outlier rates.

FIGURE 1. Reproduced from Newman et al. (2019). Here, z_{spec} is the true redshift.

4. SCHEDULE

Tentative deadlines:

- March 26: neural network
- April 2: random forest
- April 9: method 3 (TBD)
- April 16: method 4 (TBD)
- April 25: write report

5. REFERENCES

Newman et al. 2019 (in prep. scheduled to be posted to the arXiv on March 21, 2019)
Sánchez et al., Monthly Notices of the Royal Astronomical Society, 1482?1506 (2014)