TERM PROJECT PROPOSAL

TROY RAEN

Abstract. In this work I study how the errors in photoz estimates scale with the size of the training set for various machine learning algorithms. A 'photoz' is an estimate of the redshift of a particular object (usually a star or a galaxy) that uses photometric data.

1. Introduction and Related Work

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, this is called a 'redshift' and is usually denoted by z), and that the magnitude of 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 understanding 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'.)

Various machine learning algorithms have been used to estimate photozs, with neural nets and random forest regressors showing the most success.

1.1. Dataset.

I use the dataset `Catalog_Graham+2018_10YearPhot` which consists of simulated, photometric telescope data for $\sim 3 \times 10^6$ galaxies. The dataset includes the correct redshift for each galaxy, so this is a supervised, regression problem.

Features: Previous algorithms have had more success by transforming the features. This transformation is motivated by physics, Motivated by the physics in play and the history of success in previous ML algorithms, it is customary to perform a feature transformation to galaxy 'colors' by subtracting the raw measurements/features pair-wise. I will not give the detail

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 ($\sim$2000) to be used for these calculations.

	Correlations with Redshift
redshift	1.0
tu	**0.4551**
tg	0.1438
tr	0.2717
ti	0.3933
tz	0.4234
ty	0.4136
u10	0.3932
uerr10	**0.518**
g10	0.1435
gerr10	0.06651
r10	0.2716
rerr10	0.1222
i10	0.3931
ierr10	0.3822
z10	0.423
zerr10	**0.4605**
y10	0.4109
yerr10	0.45

2. Methodology

10 runs with same sample size and pool abs(photz - specz)./(1+specz) for stats.
NMAD = 1.48* median(abs(photz - specz)./(1+specz)) out10 = sum(abs(photz - specz)./(1+specz) > 0.1)/N
Use python's scipy.optimize.curve_fit() to fit the function $a + b. * x.^c$

2.1. **Neural Nets.** I train multi-layer neural networks using two different architectures: 2x10 with 2 hidden layers, each with 10 units; and 3x15 with 3 hidden layers, each with 15 units. Both are motivated by approaches in [3] (see sections 4.1.1 DESDM and 4.1.2 ANNZ).

I use the Matlab `fitnet()` function with backpropagation optimized using the Levenberg-Marquardt method. After running a few tests I set the parameters `epochs = 500`, `max_fail = 50`, `min_grad = 1e-10`. Due to runtime constraints, I set `epochs = 200`, `max_fail = 100` for all training runs with Nsamples > 99999.

2.2. **Random Forest Regression.** I train random forest regression models using the Matlab `fitrensemble()` function. I did some preliminary runs with `OptimizeHyperparameters = 'auto'` and found the following "best" options:

Method	Bag
NumLearningCycles	495
MinLeafSize	1

Guided by these results I tested some settings and ultimately use `Method = 'Bag'` with `Learners = 'tree'`, `MaxNumSplits = Nsamples-1`, and `NumLearningCycles = BLANK`, `Crossval = BLANK`. This generates random forest models using bagged decision tree models.

EXPLAIN RF REGRESSION

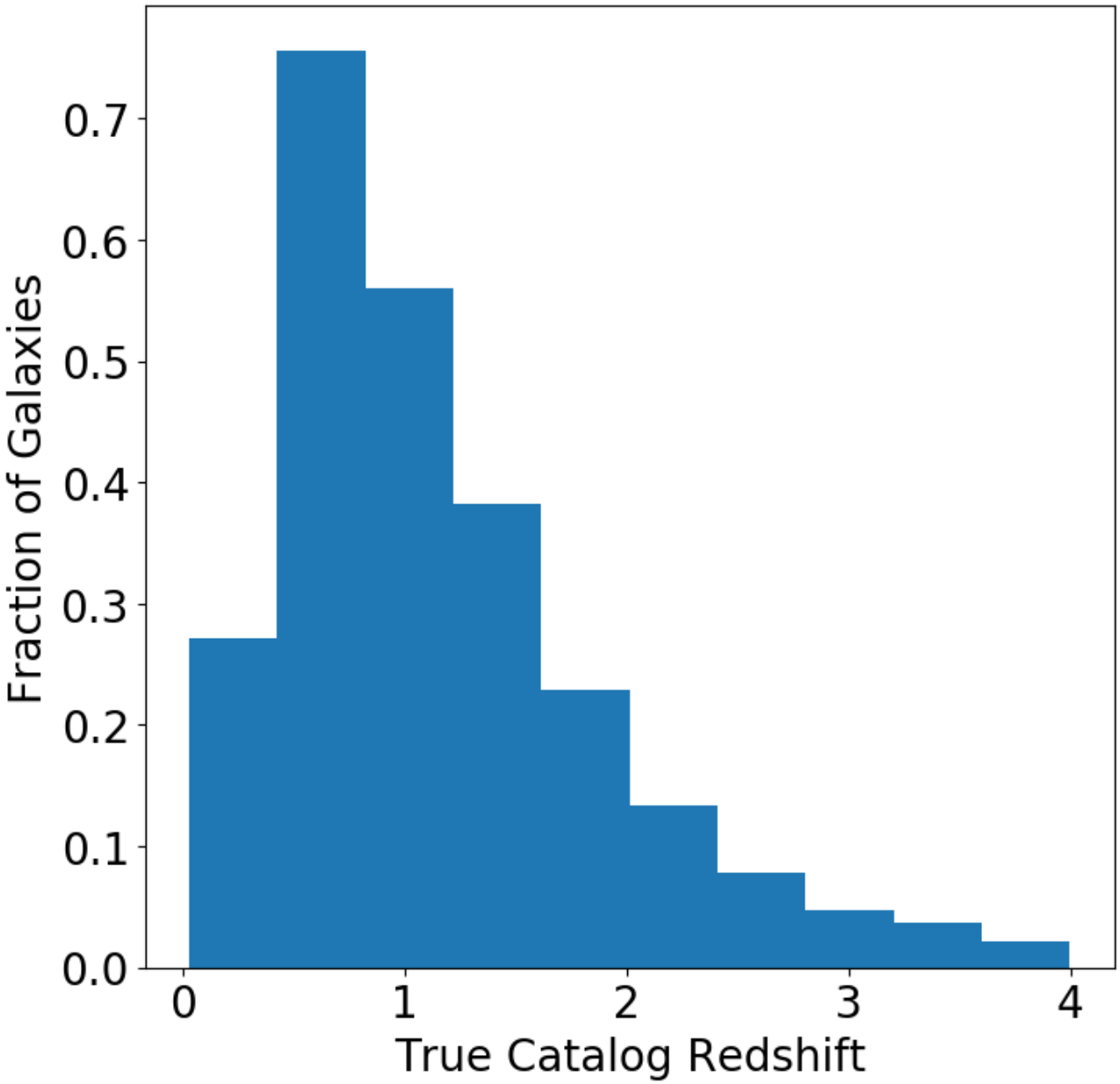
2.3. **Gaussian Process Regression.** Gaussian Processes (GPs)

non-parametric, probabilistic model that assumes output is predicted by some function of the input plus Gaussian noise. Uses kernel basis functions phi (specifically, the radial basis function) to model the density around a given data point. Does Bayseian analysis with likelihood = p(y—w) = N(mean=phi*w, sigma propto uncertainty due to density of training points in the region (=¿ uncertainty in mean prediction) + noise or error in the nearby input datapoints (heteroscedastic incorporated here)), prior = ?. Gives a posterior conditional , from which I take the point estimate.

heteroscedastic noise =¿ errors in datapoints are correlated, in other words noise is variable and input-dependent

In [2] they assert that using the sum of squares as the minimization objective function biases the metric

TRY VARYING m, THE NUMBER OF BASIS FUNCTIONS [2] pg 4, but plot 5 shows delta z does not vary much with training size for a given number of basis fncs =¿ this could improve the predictions but should not change shape of scaling with N? Could bring down out10?



3. Experimental Results

4. Analysis of Results and Discussion

5. Conclusion

In [1] they use principal component analysis to further pre-process the features, speeds up optimization.

References

[1] I. A. Almosallam, M. J. Jarvis, and S. J. Roberts. GPZ: Non-stationary sparse Gaussian processes for heteroscedastic uncertainty estimation in photometric redshifts. *Monthly Notices of the Royal Astronomical Society*, 462(1):726–739, 2016.

[2] I. A. Almosallam, S. N. Lindsay, M. J. Jarvis, and S. J. Roberts. A sparse Gaussian process framework for photometric redshift estimation. *Monthly Notices of the Royal Astronomical Society*, 455(3):2387–2401, 2016.

[3] C. Lidman, H. Lin, D. Burke, C. Cunha, N. Greisel, J. Zuntz, A. Fausti, K. Glazebrook, G. Bernstein, J. Gschwend, D. Gerdes, M. A. G. Maia, C. Sánchez, S. S. Allam, A. Dey, E. Sánchez, H. T. Diehl, D. Finley, M. Lima, D. Capozzi, R. Scalzo, A. Sypniewski, S. Jouvel, E. Fernández, I. Sadeh, S. Seitz, J. L. Marshall, J. de Vicente, J. Frieman, A. Walker, P. Pellegrini, I. Sevilla-Noarbe, A. Roodman, M. Carrasco Kind, M. J. Childress, B. X. Santiago, R. C. Nichol, F. J. Castander, A. Carnero, R. Miquel, P. Doel, F. Ostrovski, T. Davis, A. Kim, A. Evrard, F. Valdés, S. Desai, J. Estrada, O. Lahav, C. Bonnett, G. Tarle, F. B. Abdalla, D. L. DePoy, E. Gaztanaga, J. P. Bernstein, R. L. C. Ogando, D. L. Tucker, M. E. C. Swanson, W. Hartley, A. Amara, F. Yuan, K. Honscheid, S. A. Uddin, N. Kuropatkin, T. Abbot, D. Thomas, B. Flaugher, M. M. Rau, R. Brunner, E. Buckley-Geer, L. A. N. da Costa, M. Sako, K. Kuehn, M. Banerji, D. Atlee, P. Martí, M. Makler, and R. C. Smith. Photometric redshift analysis in the Dark Energy Survey Science Verification data. *Monthly Notices of the Royal Astronomical Society*, 445(2):1482–1506, 2014.