Noble Kennamer 77163775

Ariel Kruger

Abhisaar Sharma

Dylan Cockerham

CS 273A Final Project

A Survey of Machine Learning methods in Astrophysics

What to talk about:

(a) Problem statement

(b)Citations

(c) who did what

(d) Experience coding it up

(e) results and experience getting those results

(f) conclusion

(g) Bibliography

**Problem Statement**

For our project we worked on two different problems. Both focused on applying and comparing different machine learning methods on astrophysical data. In observational astronomy there are two major types of data photometric and spectroscopic. Photometric data is centered on measuring the flux (you can think of this as the number of photons per area per second hitting a CCD camera) for very broad wavelengths. In the case of the SDSS photometric data was collected for each target (galaxy, star, quasar) with five filters u, g, r, i, z. Each of these filters corresponds to a different range in the electromagnetic spectrum and can be used to infer physical properties of the target. In the case of spectroscopic data you can think of it as photometry using infinitely many filters (in practice a diffraction grating is used). Hence with spectroscopy you have much greater resolution on the light (not necessarily visible) the target is emitting. An example of a spectrum of a typical star is shown in figure 1. There are several advantages and disadvantages to photometry and spectroscopy. The main difference is that photometry is much cheaper to collect while spectroscopy provides much greater information about the object. Because of this, large surveys will typically first take and use photometric data to plan and identify which targets that want to invest the time and resources to collect spectroscopy data on. It is also important to note that the amount of data is growing both in size and complexity. For example the SDSS surveyed over 14,000 square degrees of the sky (out of about 40,000 square degrees), collected data on over 900 million unique objects and over 4 million of these included the spectrums of the objects. In addition to the already large amounts of data more data will be collected from surveys like the Dark Energy Spectroscopic Instrument (DESI),the Large Synoptic Sky Telescope (LSST), and several others. The LSST will actually be doing only **Expierence Coding it up** (Everybody write about your expierence)

photometry on tens of billions of objects bringing us close to targeting every galaxy in the observable universe. Thus it is extremely important to develop techniques to process and analyze al this data. This is where machine learning comes.

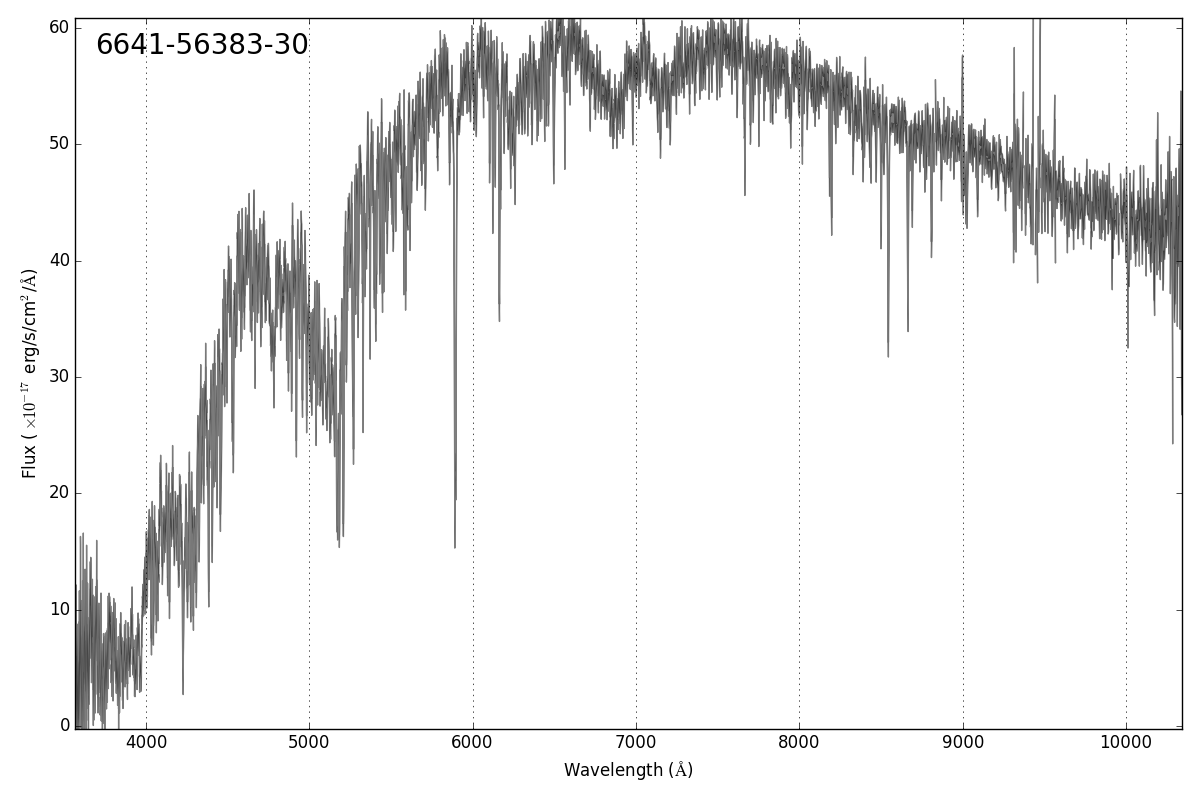


Figure 1: An example spectrum of a typical star. This spectrum comes from the Sloan Digital Sky Survey. The horizontal lines represent the amount of uncertainty of the flux value for a given wavelength.

For our project we worked on two different problems both of which used photometric data. One of the projects we worked was to classify RR Lyrae variable stars versus standard stars. This is a very important problem in astrophysics because RR Lyrae variable stars are used as standard to candles to measure distances (both galactic and extragalactic). Thus if we can develop reliable methods for discriminating between RR Lyrae stars from standard stars just using photometric data than observational astronomers can use this to select which stars to collect finer data (spectroscopy) on to determine distances (How the distances are measured is based on a relationship between the period of the variable star and its luminosity, which is interesting but not relevant to the project). The data, shown in figure 2, comes from the Sloan Digital Sky Survey (exactly how we obtained the data is discussed in a later section) and consists of four features called colors. These colors are u-g, g-r, r-i, and i-z and they are the difference of magnitudes between two different bands. For example the u-g colors is the magnitude from the g band (filter) subtracted from the u band. The dataset consists of about 93 thousand objects and only about 500 of which are RR Lyrae variable stars. This huge difference between the number of positive example (RR Lyrae stars) and negative examples makes this a vey challenging problem. Furthermore the detection of rare events is pervasive throughout observational astronomy so by identifying and developing methods that are good at addressing this issue might generalize to other problems. The detection of exoplanets, strong lensing, collisions, supernova, MACHOS(no one really cares about these anymore) and many more are all examples in astrophysics where detecting rare events is important.

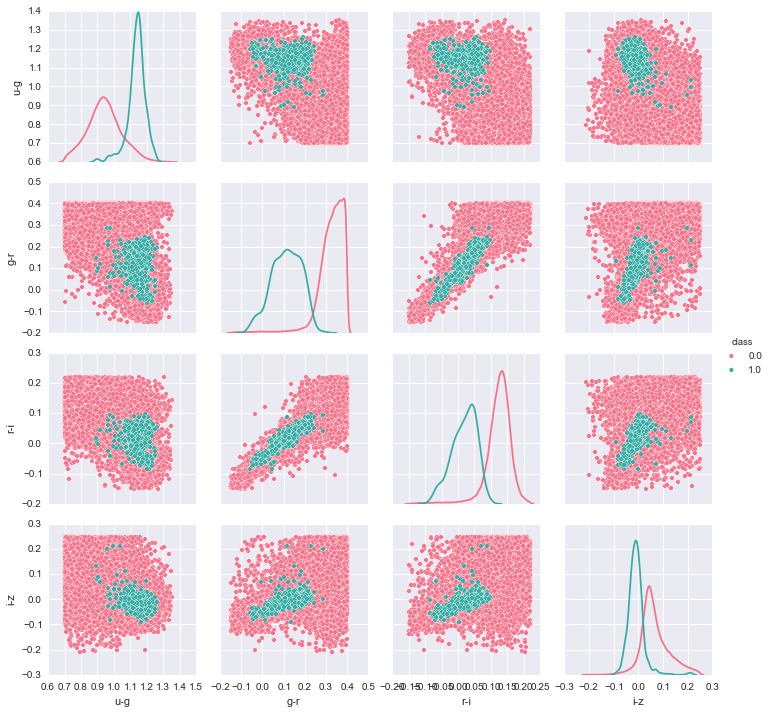


Figure 2: A pair plot of the RR Lyrae Variable stars(green) verse standard stars(pink). Our dataset consists of four features u – g, g – r, r – i, and i – z. The figure shows the relationships between all the features. For example the subfigure in the first row and first column shows u-g on the x-axis and u-g on the y-axis. The figure in the first and row and second column shows the g – r on x-axis and u-g on the y-axis. Note that the figure is symmetric about the main diagonal.

The other problem we worked on was a regression problem where we would like to assign a redshift to a galaxy using only photometric data. This dataset consists of over 60 thousand objects (all galaxies) where the features are the magnitudes of the 5 bands. Hence there are 5 features u, g, r, i, and z. The data is shown in figure 3.

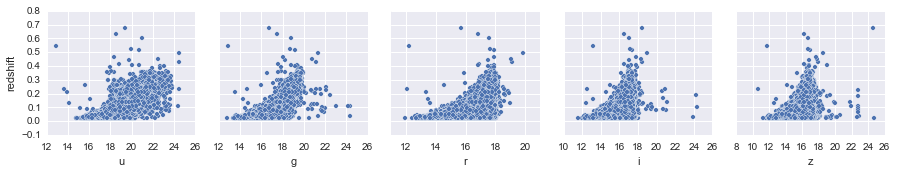


Figure 3: This figure consists of five subfigures. One for each feature of the dataset the feature is on the x-axis of each plot and the redshift is on the y-axis of every plot. Hence the middle plot shows the value of the r magnitude on the x-axis and the redshift on the y-axis with a dot for every galaxy in the dataset.

Figure 4 shows a histogram of the redshifts. As we can see the redshift ranges from about 0.001 to 0.4 with a peak around 0.1.

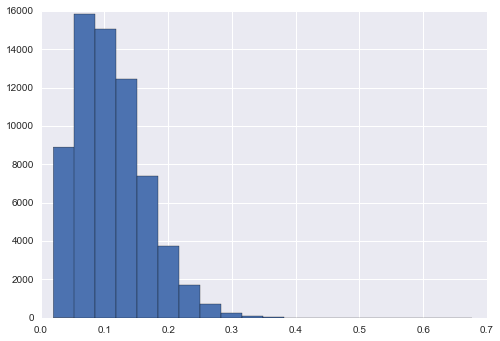


Figure 4: A histogram of the redshifts for every galaxy. As we can the redshift ranges from about 0.001 to about 0.4 with a peak around 0.1.

Redshift is very similar to a Doppler shift in that it is a phenomenon where the wavelength of the light emitted by an object gets distorted based on the motion of the object relative to the lab frame. Since all the object we are looking at are galaxies the redshift is actually from the cosmological expansion of the universe between us and the galaxy and the peculiar motions of the object play a very little role in causing the redshift. Since for our objects the redshift is caused by the expansion of the spacetime there is a direct relationship between the objects distance from us and its measurement. By determining the redshift of as many objects as we can scientists can better map the large scale structures of the universe and compute n-point correlation statistics and compare these with predictions from theoretical models of the universe. For our dataset the ground truth comes from determining the redshift from spectrums of the objects since it is a much easier problem when the spectrum is known. But as stated earlier we do not always have the spectrum and future surveys like the LSST will have no spectrums. Thus it is important to develop models that can determine redshift values from only photometry.

**Who Did What** (Everybody write what you did)

Looking at the commit history of the group GitHub can give a quick picture as to who did what. The Github can be found at the following link:

https://github.com/NobleKennamer/astro\_porject

**Noble Kennamer –**

* I organized and collected the two data sets.
* I wrote the code for the pipeline which included making the visualizations for the data, training the models, compute metrics, and displaying the results for the models in dataframes and figures. This pipeline allowed anyone in the group to write a model (conforming to a specified a.p.i) and quickly plug the model into the pipeline and immediately visualize the results in comparison with other models. Most of the figures in this report were generated from this pipeline.
* I implemented a Neural Network for the RR Lyrae classification problem. I used TensorFlow to define the graph structure of the network.
* I implemented Naïve Bayes for the RR Lyrae classification problem.
* I implemented Linear Regression (gradient descent based implementation) for the galaxy redshift problem
* I optimized several models from the sci-kit learn python library such including KNN, Decision Trees, Random Forest, and SVM for the RR Lyrae classification problem. And Decision Tree, Random Forest, Boosted Decision Tree, KNN, and Ridge regression for the galaxy redshift problem.
* Wrote the code to add more RR Lyrae stars to the training set using a KDE approach (using sci-kit learns implementations) and simply selecting a random sample from the original data to duplicate in the original training set.
* Helped organize the poster.
* Wrote the introduction, results, conclusion of the paper.

**Ariel Kruger –**

**Abhisaar Sharma -**

**Dylan Cockerham –**

In addition to our individual contributions our group was in constant contact either meeting in person, communicating through GitHub or through our facebook group. This led to many fruitful discussions on what would be interesting things to try, advice on why something might not be working or how it could work better and also kept everyone in the loop on what others were doing. Ultimately these discussions not only enhanced learning, but we feel led to a stronger project.

**What did we use** (Everybody write about the libraries and other sources you used to do your work)

**Noble Kennamer**

* Throughout the entire project I wrote all my code in python using numpy (matrix library).
* To build the pipeline I used Matplolib and Seaborn to create the plots. And functions in the sci-kit learn library to split the data into training and test sets and to compute various metrics for both problems.
* To Implement the Neural Network I used TensorFlow to define the graph structure of the network and compute the gradients.
* I implemented Naïve Bayes and Linear Regression (using gradient descent) from scratch.
* Any other model that I worked on optimizing for the respective problems I used an implementation fro sci-kit learn.

**Ariel Kruger –**

**Abhisaar Sharma -**

**Dylan Cockerham –**

**Experience** (Everybody write about your experience doing this project)

**Noble Kennamer**

I had a great experience working on this project. I learned a lot about applying machine learning methods on real world problems. Not only did I enhance my theoretical knowledge of different machine learning algorithms and methods to evaluate them I also gained practical skills in working with data.

Specifically I learned a lot about how to effectively visualize data and display results in a meaningful way that helped me to make decisions about what I should try next or why a particular model is or is not working. On the visualization side I learned a lot about the matplotlib libtary, the seaborn library and how to take advantage of pandas dataframes to quickly visualize data.

I learned a lot about TensorFlow (Google’s new Deep Learning library) by going through many of the tutorials and ultimately using it to implement a Neural Network(with dropout) for classification. I’m already applying these newfound skills towards my research by using tensor flow to implement a convolutional network for analyzing spectrum data.

I also increased my knowledge of the sci-kit learn library including what kind of models it offers and how to train them along with what code it provides for evaluating them. I can also use this knowledge in my research by using the sci-kit learn library to quickly try out many basic models and get baselines.

In addition to these practical skills I also learned a lot about many different machine learning models such as decision trees, random forests, boosted decisions trees, nearest neighbor methods, density estimation methods, various enhancements to linear regression, support vector machines and generative methods. Either by implementing them on my own or by using a sci-kit implementation I was forced to read in more detail about these methods to get an idea if they would be effective for the problems faced in this project and how to tune them to get the best performance possible out of them.

I think the greatest experience I gained was in comparing these different models and also the different methods for adding more data in my training set. The comparison gave me much greater insights into the strengths and weaknesses of the different approaches and is probably one of the most important lessons I am taking away from this project. The comparison between the different approaches will be elaborated on in the results and conclusion section.

I also really liked working in a group. This project is closely related to my Ph.D. research and by having the opportunity to share my knowledge of astrophysics and my related research. I was able to get valuable insights from members of the group of what might be interesting approaches to try. I think I gained a lot more by working in a group that I would have otherwise.

**Ariel Kruger –**

**Abhisaar Sharma -**

**Dylan Cockerham –**

**Results**

**Conclusion**

**Bibliography**