Ph 171 **June 6, 2019**

AstroML: Ph 171 Final Report

Vaishnavi Shrivastava

Github Link: https://github.com/vshrivas/AstroML

Applications of Machine Learning to Astrophysics

The amount of data being collected in the fields of astronomy and astrophysics is skyrocketing as the instrumentation used by researchers grows more and more advanced. As in most other fields, machine learning is becoming an increasingly popular strategy to truly understand the big picture created by big data. Through this independent study, I sought to explore the various areas in astrophysics in which machine learning is applied, and try to better understand how the science of astrophysics flows hand in hand with the algorithmic perception brought in with machine learning. Below I have described the highlights from some of the key research papers covering the intersection of astrophysics and machine learning. After doing a deep dive into the literature, I worked on two miniprojects to apply machine learning techniques to solving astrophysics problems, the results of which are also described below.

Exoplanet Discovery

One particularly interesting paper was *Identifying Exoplanets with Deep Learning: A Five-planet Resonant Chain around Kepler-80 and an Eighth Planet around Kepler-90*, the culmination of a collaboration between Google Brain, The University of Texas at Austin, and the Harvard-Smithsonian Center for Astrophysics. The paper focuses on utilizing data from the NASA's *Kepler* Space Telescope, which was created to "determine the frequency of Earth-sized planets orbiting Sun-like stars." (Shallue and Vanderburg) Unfortunately, many of these planets occur at distances far enough that the telescope's sensitivity is greatly undermined. Thus for several initial lists of potential planet candidates, scientists manually analyzed signals for true threshold crossing events, or TCEs, which would enable the detection of transiting planets. For such planets, the signal-to-noise ratio is incredibly low, and machine learning can be useful in overcoming this noise by helping discern which signals correspond to candidate planets, vs signals caused by a "false positive caused by astrophysical or instrumental phenomena." Previously, a different group had developed the Autovetter model, "a random forest model to classify TCEs based on features derived from Kepler pipeline statistics". (Shallue and Vanderburg) This paper used deep learning with fully connected architecture and CNNs on light curves, which consisted of "integrated flux measurements spaced at 29.4 minute intervals for up to four years" (Shallue and Vanderburg), from human-validated TCEs to predict the presence of exoplanets. The best models were able to achieve 98.8 percent accuracy, and discovered new planet candidates around Kepler-80 and Kepler-90, as likely exoearths.

The Problem of Noise in Astrophysics

Noise is a problem that plagues any field reliant on data collection, and astrophysics is no exception. Noise obscures the distinction between key findings and irrelevant detections, making almost all experimental astrophysics problems much more difficult to solve. Furthermore, since noise is a fundamental feature of the environment we are operating in, and the limitations of the instruments being used, there is no good way of eliminating it. Applying machine learning techniques to detecting noise within collected data is becoming increasingly important in experimental astrophysics, as the amount of data gathered skyrockets, making it

impossible for scientists to manually remove noisy data points. In the paper, *Application of machine learning algorithms to the study of noise artifacts in gravitational-wave data*, machine learning is used to detect non-Gaussian noise artifacts in LIGO data, by applying the algorithms to "data from auxiliary channels within the LIGO detectors that monitor degrees of freedom unaffected by astrophysical signals". (Biswas, Blackburn et. al) LIGO detectors suffer from two different types of noise, a "stationary component of colored Gaussian noise" and "short-duration non-Gaussian noise artifacts called 'glitches'". (Biswas, Blackburn et. al) The stationary noise is essentially background noise that is nearly always present, and perturbing observations in an expected, predictable way, and can more easily be eradicated from the collected observations. On the other hand, the non-Gaussian noise consists of erratic, transitory disturbances such as "temporary seismic, acoustic, or magnetic disturbances, power transients, scattered light, dust crossing the beam, instabilities in the interferometer, (and) channel saturations". (Biswas, Blackburn et. al)

To be aware of the state of the LIGO detectors, in relation to the various types of noise, the system constantly gathers feedback from numerous auxiliary channels to more accurately monitor the environment surrounding the device. As one may expect, non-Gaussian noise artifacts are particularly problematic because the artifacts confound the detection of true transient gravitational-wave signals, especially at higher signal-to-noise ratios. Consequently, researchers rely on multiple LIGO detectors to catch the same signal to overcome this problem, and provide more compelling evidence of a gravitational-wave detection. Despite having multiple detectors listening for signals, "accidental coincidence of noise transients across multiple detectors still dominates the search background"(Biswas, Blackburn et. al) Gravitational waves from binary neutron stars are easier to detect with higher confidence levels, because their waveforms have been more accurately modeled, making it possible to screen out signals that don't match the expected forms. However, even for binary neutron stars, "volume search sensitivity is 30 percent less than what it could be in the presence of only Gaussian noise". (Biswas, Blackburn et. al) The problems caused by noise are most exacerbated in the detection of generic gravitational-wave bursts and intermediate-mass binary black-hole coalescence, since researchers have less information about their wave structure, and the detectors can only sense the waves for very brief periods of time.

There are a number of algorithms that are currently in use for glitch detection, all of which centrally rely on determining pairwise correlations between glitches in the gravitational-wave channel and those in auxiliary channels. Such algorithms seek to establish connections between the causes of noise being monitored on some auxiliary channel to noise artifacts produced in gravitational-wave detection. However, the major shortcoming of such algorithms is the built-in assumption of some specific relationship between glitches in the gravitational-wave channel and some auxiliary channel. Another fundamental concern is the inability of these algorithms to discern the underlying couplings between multiple auxiliary channels, which could have related responses to environmental perturbations. Due to the lack of scientific understanding of the relationships between responses of various auxiliary channels, and the sheer number of auxiliary channels in use at LIGO, non-Gaussian noise artifact detection becomes a problem particularly well-suited to machine learning, which is well-equipped to handle a problem of such high dimensionality. The paper demonstrates that each of the evaluated machine learning algorithms, artificial neural networks, support vector machines (SVMs), and random forests, were viable models able to detect a significant number of non-Gaussian noise artifacts. These algorithms also had comparable performance to the Ordered Veto List Algorithm, which performed an optimized version of pairwise correlation detection across auxiliary channels and the gravitational-wave channel.

Simulating Dark Matter with GANs

The cosmic web is a representation of the dark matter in the universe as it interacts with gravity, "to form a complex network of halos, filaments, sheets and voids". (Rodriguez, Kacprzak et.al) Unfortunately, studying the cosmic web requires being able to visualize these physical interactions, which is traditionally accomplished by building N-body simulations from classical physics. The N-body simulations would be able to emulate the distribution of matter over time, but consequently creating these simulations is a computationally intensive process, effectively bottlenecking the speed of research. Fortunately, Generative Adversarial Networks

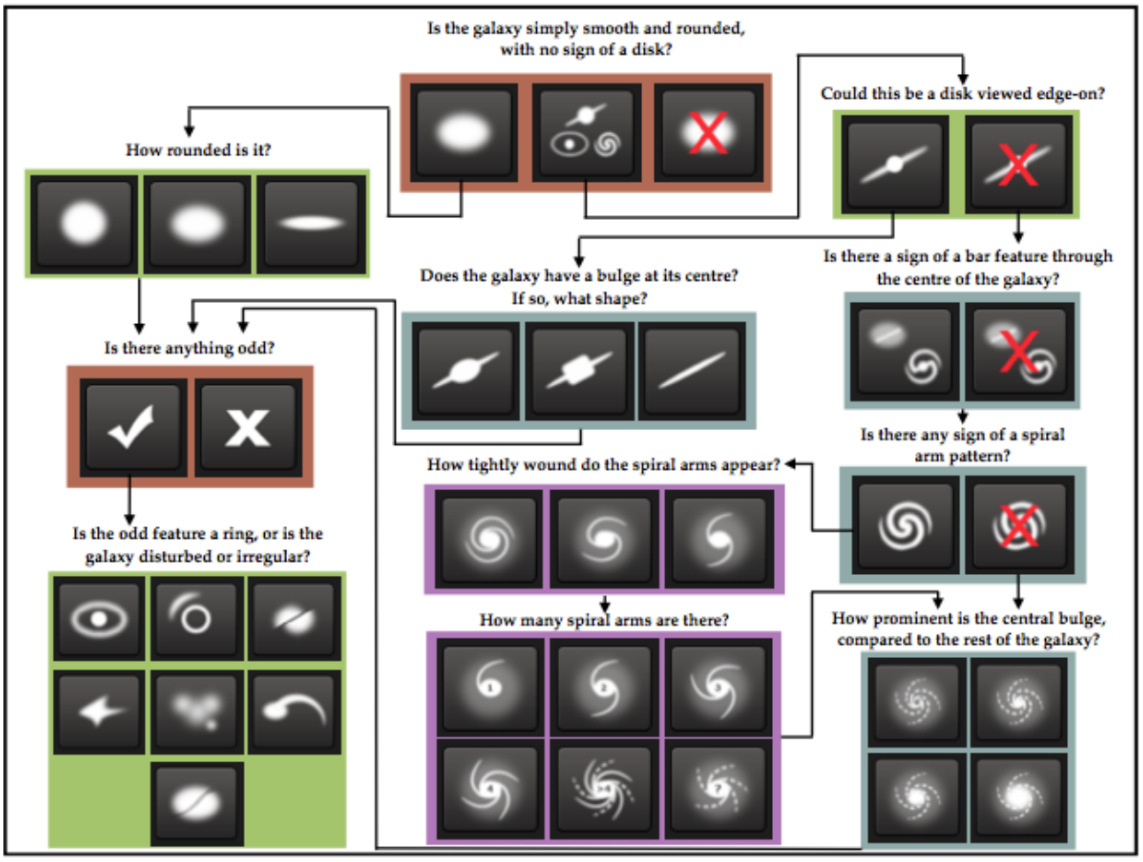

Figure 1: Flowchart showing decision process for classifying galaxy morphologically.

(GANs) can be used to learn the matter distributions and interactions from a few N-body simulations, and then "create new, random dark matter distributions that are uncorrelated to the training examples." (Rodriguez, Kacprzak et.al)
After analyzing the various avenues where machine learning can be applied to astrophysics, I chose two astrophysics problems that were of particular interest to me, and tried to solve them by applying machine learning algorithms.

Miniproject 1: Galaxy Zoo

Introduction

Galaxy Zoo was a crowd-sourced project that enlisted people to help morphologically classify galaxies based on their images from the Sloan Digital Sky Survey. Being able to classify galaxies morphologically is a critical problem in astrophysics, because knowing the shape and structure of a galaxy allows us to understand how it formed and evolved into its current shape. These questions formed a decision tree for those helping classify the galaxies, culminating in 37 different categories of galaxies. The responses of all individuals involved in classifying the galaxies, were aggregated to give the probabilities of each galaxy image receiving a particular classification, across individuals. Using machine learning to solve this problem would greatly increase the efficiency of classifying newfound galaxies, and potentially reduce confusion and time spent thinking over particularly difficult classifications, by allowing more objective, uniform patterns to be used.

Approach: Data Preprocessing and Network Architecture

The data for this miniproject was taken from a Kaggle contest for Galaxy Zoo. However, to simply the problem, the galaxy zoo decision problem was solved instead. The Kaggle dataset has 61,578 images for training on, and 79,975 images for testing on. However, due to limitations of time and computational resources, we chose to use a total of 10,000 images for training and testing combined, with a 20 percent validation split. The number of images used from the full dataset is a configurable parameter in the code. To preprocess the image data for training, the code traversed the image data directory, and matched each image up with a label based on the highest probability classification. The images were initially 424 x 424 color JPEG images. By observing the data we could see that only the center of each image contained the galaxy itself, so each image was first cropped to 212 x 212, and then downsampled to half resolution at 106 x 106 to reduce the number of parameters in the problem. The machine learning model we created was a convolutional neural network, the general workhorse for image classification, since relative spatial locations of

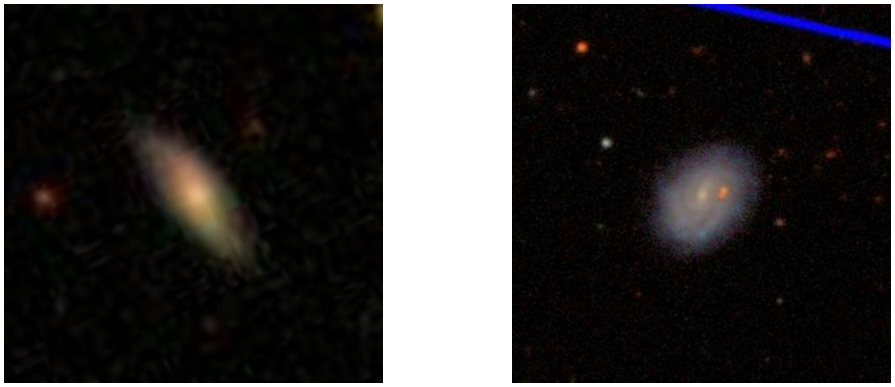

Figure 2: Example images of galaxy data.

pixels are critical in deciphering the complete image. After some experimentation, the final network architecture, included in the code, was chosen. The network was trained for 20 epochs, with a batch size of 50.

Results and Potential Improvements

```
Epoch 15/20
8000/8000 [==============================] - 157s 20ms/step - loss: 0.0152 - acc: 0.6064 - val_loss: 0.0154 - val_acc: 0.5960
Epoch 16/20
8000/8000 [==============================] - 167s 21ms/step - loss: 0.0152 - acc: 0.6064 - val_loss: 0.0154 - val_acc: 0.5960
Epoch 17/20
8000/8000 [==============================] - 155s 19ms/step - loss: 0.0152 - acc: 0.6064 - val_loss: 0.0154 - val_acc: 0.5960
Epoch 18/20
8000/8000 [==============================] - 246s 31ms/step - loss: 0.0152 - acc: 0.6064 - val_loss: 0.0154 - val_acc: 0.5960
Epoch 19/20
8000/8000 [==============================] - 169s 21ms/step - loss: 0.0152 - acc: 0.6064 - val_loss: 0.0154 - val_acc: 0.5960
Epoch 20/20
8000/8000 [==============================] - 181s 23ms/step - loss: 0.0152 - acc: 0.6064 - val_loss: 0.0154 - val_acc: 0.5960
Vaishnavis-MacBook-Pro:GalaxyClassification vshrivas$
```

Figure 3: Screenshot of Galaxy Zoo Classification Training.

As we can see in figure 3, after training for 20 epochs, we have an accuracy of 0.6064 on the training set, and an accuracy of 0.5960 on the validation set. For these last epochs, we see the accuracy of both sets stagnating. It is also important to note that since this is a multi-class classification problem with 37 classes, achieving even this accuracy is indicative of reasonably good performance. Since the accuracy on both sets

seems relatively similar, we may need a deeper or more complex convolutional neural network to fully capture the complexity of the dataset, and achieve a higher accuracy. Training for more epochs on a larger number of images would also likely improve the accuracy. If training on more data or for longer periods of time resulted in overfitting, we could combat this with data augmentation, by modifying the existing data slightly, by performing some translations. Space has no real notion of up or down being distinct, so the images of the galaxies should be rotationally invariant, and performing such translations would help increase the amount of data, while capturing a wider range of image representations.

Miniproject 2: Exoplanet Detection

Introduction

Predicting the existence of exoplanets is a challenge on the one of the most exciting frontiers of astrophysics, the search for potential exoearths and extraterrestrial life. There are multiple strategies that can be used to search for exoplanets, but we will use data consisting of light intensity readings from the star the potential exoplanet is orbiting, over time. If there is an exoplanet orbitting a given star, then the light intensity, or flux observed from that star would be slightly reduced as the planet eclipsed parts of the star in our line of view as it continue its orbit. This sort of periodic dimming of flux, as shown in figure 4, could potentially suggest the existence of an exoplanet.

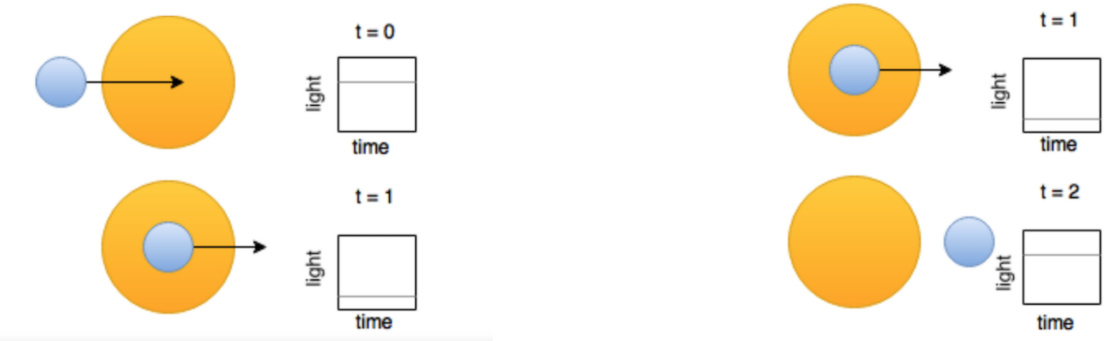

Figure 4: This diagram shows the periodic change in light intensity as a planet orbits around a star.

The dataset was obtained from a Kaggle competition. The training data consisted of 5087 observations, labelled as exoplanet or non-exoplanet stars, followed by 3198 columns or features, which were the flux values from the star over time. The test data had 570 observations, and the same number of flux features. The training set had 37 exoplanet stars, while the test set had 5 exoplanet stars.

Approach

Precision and recall were used as metrics, instead of accuracy, due to the strong imbalance in data favoring negative examples. Data Augmentation was used to create new examples by using the given dataset and making slight modifications using the Synthetic Minority Oversampling Technique (SMOTE). Since the signal of light intensity from the star is actually a combination of different pure frequencies put together, an approach worth considering is decomposing this signal into its set of pure frequencies. This would allow us to have more features to train with, and enable us to explore if the pure frequencies of signals from exoplanet stars different from those of nonexoplanet stars. The model used for a simple vanilla neural network, the specifications of which can be seen in the code. The model was trained for 50 epochs with a batch size of 32.

Results and Potential Improvements

```
Epoch 45/50
10100/10100 [==============================] - 0s 42us/step - loss: 0.0225 - acc: 0.9998
Epoch 46/50
10100/10100 [==============================] - 0s 40us/step - loss: 0.0396 - acc: 0.9932
Epoch 47/50
10100/10100 [==============================] - 0s 41us/step - loss: 0.1945 - acc: 0.9451
Epoch 48/50
10100/10100 [==============================] - 0s 40us/step - loss: 0.6037 - acc: 0.7987
Epoch 49/50
10100/10100 [==============================] - 0s 41us/step - loss: 0.2524 - acc: 0.9162
Epoch 50/50
10100/10100 [==============================] - 0s 40us/step - loss: 0.0421 - acc: 0.9993
train set error 0.0011794770984863145
dev set error 0.012280701754386003
------------
precision_train 0.8604651162790697
precision_dev 0.4166666666666667
------------
recall_train 1.0
recall_dev 1.0
------------
------------
Train Set Positive Predictions 43
Dev Set Positive Predictions 12
------------
```

Figure 5: Screenshot of Exoplanet Detection Result.

As we can see in figure 5, the recall is 1.0 for both the training and test sets. The precision is 0.8605 for the training set and 0.4167 for the test set. This result could potentially be improved by trying other ways of data augmentation, attempting techniques to balance the dataset for both labels, and further combatting overfitting.

References

Banerji, Manda, et al. "Galaxy Zoo: Reproducing Galaxy Morphologies via Machine Learning."Monthly Notices of the Royal Astronomical Society, vol. 406, no. 1, 2010, pp. 342353., doi:10.1111/j.1365-2966.2010.16713.x.

Biswas, Rahul, et al. "Application of Machine Learning Algorithms to the Study of Noise Artifacts in Gravitational-Wave Data." Physical Review D, vol. 88, no. 6, 2013, doi:10.1103/physrevd.88.062003.

Gonzalez, C. A. Gomez, et al. "Low-Rank plus Sparse Decomposition for Exoplanet Detection in Direct-Imaging ADI Sequences." Astronomy amp; Astrophysics, vol. 589, 2016, doi:10.1051/0004-6361/201527387.

Khalifa, Nour Eldeen, et al. "Deep Galaxy: Classification of Galaxies Based on Deep Convolutional Neural Networks." 2018 International Conference on Computing Sciences and Engineering (ICCSE), 2018, doi:10.1109
/iccse1.2018.8374210.

Lahav, O., et al. "Neural Computation as a Tool for Galaxy Classification: Methods and Examples." Monthly Notices of the Royal Astronomical Society, vol. 283, no. 1, 1996, pp. 207221., doi:10.1093/mnras/283.1.207.

Rodrguez, Andres C., et al. “Fast Cosmic Web Simulations with Generative Adversarial Networks.” Computational Astrophysics and Cosmology, vol. 5, no. 1, 2018, doi:10.1186/s40668-018-0026-4.

Shallue, Christopher J., and Andrew Vanderburg. “Identifying Exoplanets with Deep Learning: A Five-Planet Resonant Chain around Kepler-80 and an Eighth Planet around Kepler-90.” The Astronomical Journal, vol. 155, no. 2, 2018, p. 94., doi:10.3847/1538-3881/aa9e09.

Vanderplas, Jacob, et al. “Introduction to AstroML: Machine Learning for Astrophysics.” 2012 Conference on Intelligent Data Understanding, 2012, doi:10.1109/cidu.2012.6382200.

Zingales, Tiziano, and Ingo P. Waldmann. “ExoGAN: Retrieving Exoplanetary Atmospheres Using Deep Convolutional Generative Adversarial Networks.” The Astronomical Journal, vol. 156, no. 6, 2018, p. 268., doi:10.3847/1538-3881/aae77c.