DATASCI 3ML3 PROJECT PROPOSAL

Charles Leung

400049422

I've been fascinated by astronomy and the cosmic structures of the Universe for as long as I can remember. For example, galaxies are vast collections of stars, gas, and dust, bound together by gravity, and are the building blocks of our universe. Studying them can help us understand the formation and evolution of the universe and answer unsolved questions in physics (e.g. the nature of dark matter). One way to study galaxies is through their images, which are captured by telescopes and other instruments. These images contain a wealth of information about the properties of galaxies, such as their size, shape, color, and brightness.

Galaxies come in myriad shapes that are influenced by their composition and evolution. However, analyzing these images manually can be time-consuming and prone to human error. Because machine learning models can be trained to recognize patterns and features in galaxy images and classify them according to their properties, they can help astronomers and astrophysicists automate the process of analyzing galaxy images and extract useful information from them more efficiently. Machine learning can also help to identify new types of galaxies and explore the unknown regions of the universe. The objective of this project is to investigate the use of machine learning techniques for image classification of galaxies, using models such as k-means, decision trees and random forests and evaluate their effectiveness through testing.

Image classification is a fundamental problem in computer vision and has important applications in fields such as medical diagnosis, facial recognition, and self-driving cars. The application of machine learning models to perform image classification of galaxies has already shown promising results in

recent years. For example, a study conducted by the Galaxy Zoo team - a citizen science project - used decision trees and random forests to classify over a million galaxy images according to their morphological features, such as the presence of spiral arms or the shape of the galaxy. This study demonstrated that machine learning can be a powerful tool for analyzing large-scale astronomical data sets, and can complement traditional methods of analysis.

The problem of image classification of galaxies can be framed as a supervised learning problem, where a set of labeled galaxy images are used to a machine learning model to predict the labels of new, unlabeled images. The scope of the project will focus on binary classification, where each galaxy image can be classified as either a spiral galaxy or an elliptical galaxy. This is a common classification scheme used by astronomers, as spiral and elliptical galaxies have different properties and are believed to have different origins, as well as being the most common types of galaxies observed.

Training the machine learning model requires extracting features from the galaxy images that can distinguish between spiral and elliptical galaxies. There are various feature extraction techniques that can be used for this purpose, such as the pixel intensity distribution, the color distribution, the texture, and the shape of the galaxy. The combination of these features needed for classification will require further investigation, as different features may be more informative for different types of galaxies.

In this project, I'd like to focus on two machine learning models in particular: decision trees and random forests. Decision trees recursively splitting the data into subsets based on the features that best separate the classes. On the other hand, random forests are an ensemble of decision trees that combine multiple trees to improve the accuracy and robustness of the model. They work by randomly selecting a subset of features and a subset of data points for each tree and aggregating the results. Random forests can help correct for overfitting tendencies in decision trees, and it may be particularly insightful to compare the two in the context of this problem.

Implementation-wise, the scikit-learn library in Python can be used to implement the machine learning algorithms and perform the data analysis, while the Astropy library to read and manipulate the galaxy image data. A potential data set that can be used for the project is the Galaxy Zoo 2 data set, which contains over 300,000 galaxy images labeled by citizen scientists.

One potential limitation of this project is the quality of the galaxy image data. Galaxy images captured by telescopes can be affected by various factors such as noise, atmospheric conditions, and instrument calibration. These factors can introduce biases and uncertainties in the data, which can affect the performance of the machine learning algorithms. To address this issue, preprocessing may be needed.

Another potential issue is the imbalance of the data set, where one class (e.g. spiral galaxies) may have more images than the other class (e.g. elliptical galaxies). This can lead to biased classification results, where the model may be more accurate in predicting the majority class but less accurate in predicting the minority class.

Finally, the models will be evaluated and compared using metrics such as accuracy, precision and recall.