Blake Lopez
ECE 532 - Project Proposal
10/22/2020

Hazardous Asteroid Classification

A data set from http://neo.jpl.nasa.gov/ was generated using the Near Earth Object Web Service (NeoWs). The data set has several features for 4687 asteroids orbiting Earth labeled as either hazardous or not. About 84% of the asterous are non hazardous and 16% are hazardous. The features include estimated diameter (min and max), relative velocity, orbit uncertainty, eccentricity, mean motion, and many more. Some of the many features are repeats for different units for diameter and velocity and some of the features are ways of identifying the asteroid, dates of importance, and some features are the same for every asteroid. The full dataset can be found at https://www.kaggle.com/shrutimehta/nasa-asteroids-classification.

Due to the high number of features, feature selection and pre-processing of the data will be crucial to completing a chosen algorithm. For example, due to the many different types of features, eliminating redundant features and features that are constant must first be performed. Additionally the feature could be normalized in order to better understand the distribution of each feature when selecting relevant parameters. K-means will be used as well to understand feature relevance as well in the data processing phase.

Additionally, the dataset does not break the data into training and test data so this will have to be performed as well. Final testing data consisting of about 10% of the dataset will be held for final testing and comparison of the different models generated.

The algorithms to be used on the data will be least squares, LASSO, and a neural network. The least squares will first be implemented to understand which features appear to the most important. For this algorithm, k-fold cross validation will be used to compare different models based on different features.

Least absolute shrinkage and selection operator (LASSO) will be used to perform both feature selection and regularization. For this algorithm, the regularization parameter will be an additional parameter that can be varied. The regularization and the more automated parameters selection should lead to a more accurate and efficiently produced classifier than the least squares. Much like with the least squares, k-fold cross validation will be completed.

Finally, a neural network will be used to classify the data. A neural network depends on many features such as number of input nodes, different combinations of input nodes, recurrent nodes, and more.

Once all of the classifiers produced, they will all be evaluated on the holdout training data (the 10% of the original dataset) to determine the generalization and accuracy of each of the models. Based on this analysis, a alogitimh will be chosen and a classifier will be produced from the entire data set.

To complete the classification of this dataset, a proposed timeline is below. The Github page for following the progress of this project is https://github.com/lopezbl/ECE532_Project.git

Project Timeline		
Date	**Tasks**	**Submission**
11/01/2020	Complete pre-processing of the data	None
11/08/2020	Complete the least squares algorithm for classification	None
11/15/2020	Complete the LASSO algorithm for classification	None
11/17/2020	Update the Github page with progress	Update 1
11/27/2020	Complete the neural network algorithm for classification	None
12/01/2020	Update Github page with progress and initial comparison of the algorithms	Update 2
12/03/2020	Complete final comparison and recommended classification	None
12/08/2020	Complete first draft of the final report	None
12/12/2020	Submit final draft of the final report	Final Report
12/17/2020	Evaluate two other projects	Peer Evaluations