Jahanvi Patel

Professor Cai

CSC 6850

26 April 2020

Machine Learning Estimation and Classification

Methods and algorithms are a fundamental aspect in understanding and applying machine learning. In this project, the goal was to program the concepts to predict and classify the given data sets. To successfully predict the missing values, I had to sit down to first understand how to go about the problem. After months of trying to understand the data set, I figured out that the rows with missing values were repeated throughout the data set with the same pattern along with the value of the missing value. For example, in TrainData3, the first line contains a missing value, shown by the following:

2 1 5 4 5 5 3 3 0 1 1 7
1.00000000000000e+99

As the value 1.00000000000000e+99 represents the missing value, so to find the pattern, you can see that the following values:

2 1 5 4 5 5 3 3 0 1 1 7

can be found throughout the data set as a complete row:

2 1 5 4 5 5 3 3 0 1 1 7
1

This data can be found in lines 44, 47, 5352, and 6179. It can be then concluded that the missing value for the first row of TrainData3.txt is 1. Using this thought process, I found it difficult to come up with working Python code, as I had never worked with Python. I had never worked with such large data sets, nor Python, so setting it up and implementing my thought process was very difficult. I used this thought process to find the missing values in the first half of Question 1 and in Question 2. The data types were difficult to set up between the float values and the integer values since some were floats and some were integers.

Classification can be done in many different methods. Some include: Naïve Bayes Classifier, Support Vector Machines, K-Nearest Neighbor (KNN), Decision Trees, and Random Forests. Unfortunately, the fact that I had never utilized Python and it was time-consuming to implement my missing values algorithm, I could not implement any of the classification algorithms. My future plan in continuing this project would be to implement the different classification algorithms and use the TrainingData and TrainingLabels to calculate the accuracy rate, to apply to the TestData. For the Multi-Label classification, I was planning to implement the K-Nearest Neighbors, Decision Trees, and Neural Networks in an algorithm.

All in all, due to time constraints and the fact of never utilizing these technologies, I found it difficult to complete.