Project Report

Hasan Raza | CSC 4980

1) Classification

I used Python to solve these 5 problems. I implemented a Naive Bayes Classifier in all of them. First, the program would store the train data and the train labels into arrays. The program would then calculate the mean for each feature, then loop through the train data replacing any missing values with the mean for that feature. I used this method because there were not many missing values, and so replacing them with the mean would not affect the distribution significantly. Next, the program would divide the dataset by label, and store the relevant data for each class in separate arrays. Then, the program calculates the mean and sample standard deviation for each feature, for each class. Using this information, the program then loops through the test data, and calculates the probability for each individual feature belonging to each class. It did this using the gaussian normal distribution equation. It multiplied each probability together to compute the sample's probability of being in each class. Then, the sample was labeled with the class with the highest probability.

2) Missing Value Estimation

I used Python to solve these 3 problems. I implemented a Kth Nearest Neighbor imputation algorithm to do so, with k being 1. The program would first store all the data into an array. Then, the program would loop through the array until it hit a missing value. When it did so, it

calculated the euclidean distance from that gene to all other genes. If another gene had a missing value, it's value was set as 0 during this calculation. Next, it sorted all the distances, and found the second shortest one (the shortest one would be 0, to itself). It tried to replace the missing value with the relevant value from the second shortest gene. However, if that value was also missing, the program would loop again, and find the next shortest distance. This continued until all missing values were replaced. This program showed the lazy nature of the KNN algorithm, as the program would do no calculations until it needed to find a missing value, and then, it would only perform the relevant calculations.