REPORT ON COSC 522 PROJECT 1

STUDENT NAME: JOSEPH OCHE AGADA
NET ID: JAGADA

TASK ONE

In line with the project requirements, synthetic datasets comprising 300 training and 150 test sets were generated for 2 classes. The distribution assumed is a bivariate Gaussian distribution. These datasets were generated for the covariance of 0.0, 0.5, and – 0.5. The mean vectors chosen for classes 1 and 2 are [2, 4.5] and [3.5, 2.8] respectively and these mean vectors are the same for all 3 cases. For each case (of covariance 0, 0.5, and -0.5), a scatter plot was implemented for training and test sets separately. Ultimately, all the datasets were merged into a single training set, and a single test set and scatter plots were created for the combined dataset. At a covariance value of 0, the scatter plot depicts no relationship between the 2 variables. The scatter plot took a circular shape which is indicative of no correlation. At a covariance of 0.5, the scatter plot depicted an obvious positive correlation between the 2 variables. This means that the points in the scatter plots started converging around an inexistent straight line. For covariance of -0.5, the strength of the relationship as depicted by the plots remains the same but the direction changed to a negative relationship, meaning that increase in values of one variable is followed by a decrease in the values of the other variable. In the end of the entire tasks, the datasets were merged across the various cases and saved in .tr and .te file extensions.

TASK TWO

In Task 2, I implemented KNN classification on the dataset (without using any machine learning framework) for k running from 1 to 50. Thereafter, I plotted k values against accuracy and realized that accuracy generally decreased with an increase in the value of k.

TASK THREE

In Task 3, I implemented the KNN Classifier, Quadratic Discriminant Classifier, Mahalanobis Distance Classifier, and Euclidean Distance Classifier. I also went ahead to calculate overall accuracy, class-wise accuracy, and run time for each of the models. Prior probabilities here are equal because the dataset used is perfectly balanced, meaning that both class 1 and class 2 have the same number of samples.

TASK FOUR

The four classification algorithms used, all performed well. However, KNN has the highest performance but at the expense of computational complexity, being that the runtime of 0.1175 sec was the highest among all the classifiers. In terms of the assumption of equality or non-equality of covariance, the reality is the covariance for the two classes is equal but not identical. This agrees with the assumption of the Mahalonobis distance classifier. However, the Mahalanobis distance classifier has the same performance as the Quadratic Discriminant Classifier and Euclidean Distance Classifier. Therefore, violation of the assumption of the difference covariance matrix for the Quadratic Discriminant Classifier and that of the equal and identical covariance matrix for the Euclidean Distance Classifier did not impact the performance of these classification algorithms. Therefore, it is safe to say that Quadratic Discriminant and Euclidean Distance Classifiers can always be used even if their assumptions about covariance matrices are violated.

TASK FIVE

The decision boundary is superimposed on the scatter plot for the various classification algorithms. As seen from the plot, the decision boundary for KNN and QDA are non-linear while those of Mahalanobis Distance and Euclidean Distance classifiers are linear. The behavior of the decision boundaries is in line with expectations based on the theoretical standpoint.

BONUS TASK ONE

As seen from the plot, prior probabilities have a significant impact on the performance of the minimum Euclidean Distance Classifier. The closer the probability of class 1 is to 0.5, the better the performance of the model. The model has its best performance at the probability of class 1 ranging from 0.44 to 0.65.