Language: python3

Problem 1

1. (a) For the "Synthetic1" dataset, the train and test error rates are **0.21** and **0.24**. However, on the "Synthetic2" dataset, they are **0.03** and **0.04**, respectively. Figures 1 and 2 illustrate the plots obtained on the two synthetic datasets, respectively.

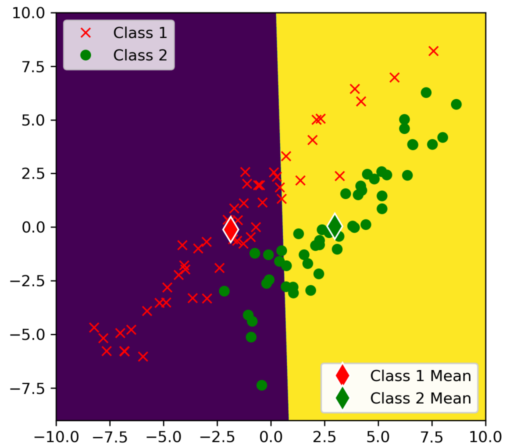

Figure 1: Synthetic1: Plot of training data, class means, decision boundaries, and regions.

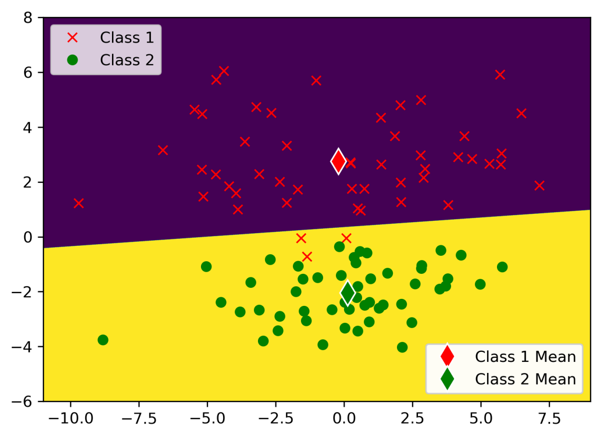

Figure 2: Synthetic2: Plot of training data, class means, decision boundaries, and regions.

	Synthetic1			Synthetic2		
Statistics	0	1	2	0	1	2
count	100	100	100	100	100	100
mean	0.55	-0.04	1.5	-0.03	0.35	1.5
std	3.99	3.27	0.5	3.4	2.76	0.5

Table 1: Summary of the statistics of the two synthetic datasets.

(b) Yes, there's a good amount of difference in the error rate between the two synthetic datasets. Table 1 gives a statistical summary of the two synthetic datasets after running *df.describe()*, where *df* is the dataframe of a particular dataset. The *describe()* function of Pandas generates the required statistics of a dataset. In Table 1, only *count*, *mean*, and *std* have been reported, for both the datasets. We can clearly observe that, for both the features (*0* and *1*), (*2* : labels), the standard deviation (*std*), is the least for "Synthetic2". This clearly translates to the fact that the variance is on a lower side and hence, the uncertainty of the spread around the mean, is less. This gives a stronger estimate of the confidence interval, and hence, the nearest-mean classifier gives a lower error rate for the "Synthetic2" data i.e **0.03** and **0.04** on the train and test, respectively.

(c) The nearest mean classifier was "trained" on the first two features (x_1 = alcohol content and x_2 = malic acid content) of the *wine* dataset by UCI. The error rates on the train and test set are **0.202** and **0.225**, respectively. Figure 3 is an illustration describing the training data points, class means, decision boundaries, and regions, for the first two features.

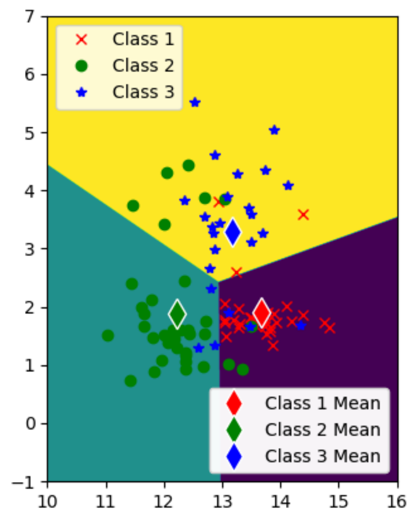

Figure 3: Wine: Plot of training data (x_1 = alcohol content and x_2 = malic acid content), class means, decision boundaries, and regions.

(d) The two best features, after a **brute-force approach**, are $\mathbf{x}_1$ and $\mathbf{x}_{12}$, giving an error rate of **0.079** on the training set. The following approach was implemented: At first, all possible combinations of pairs of features were taken. For each pair, the L2 norm was calculated between the data points of the two features and the means of the same features. The pair of features that gave the lowest error rate, were chosen as the best features. Figure 4 is an illustration of the data points of x_1 and x_{12}, class means, decision boundaries, and regions. The best features gave error rates of **0.079** and **0.124** on the train and test sets, respectively.

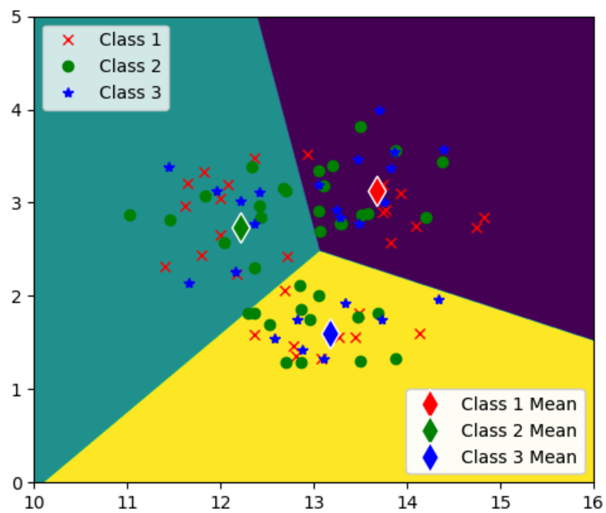

Figure 4: Wine (best features): Plot of training data (x_1 and x_{12}), class means, decision boundaries, and regions.

(e) No, there's not a *big* difference in the error rates between different pairs of features on the training set. For example, error rates between x_1 and x_{10} is 0.258, x_2 and x_8 is 0.326, x_7 and x_8 is 0.169, x_{12} and x_{13} is 0.247, and so on. Similarly, error rates on the test set are 0.225 for x_1 and x_{10}, 0.393 for x_2 and x_8, 0.247 for x_7 and x_8, and 0.303 for x_{12} and x_{13}. However, in machine learning, our objective would be to choose features which would give the least error rate, and hence, we choose x_1 and x_{12}, as the best set of features.