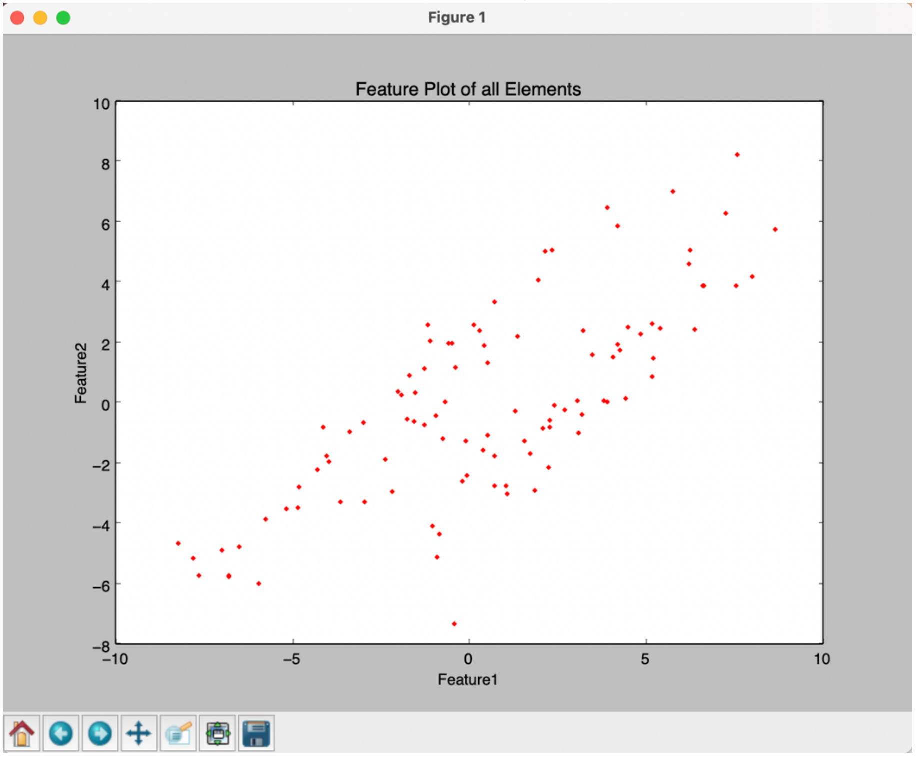
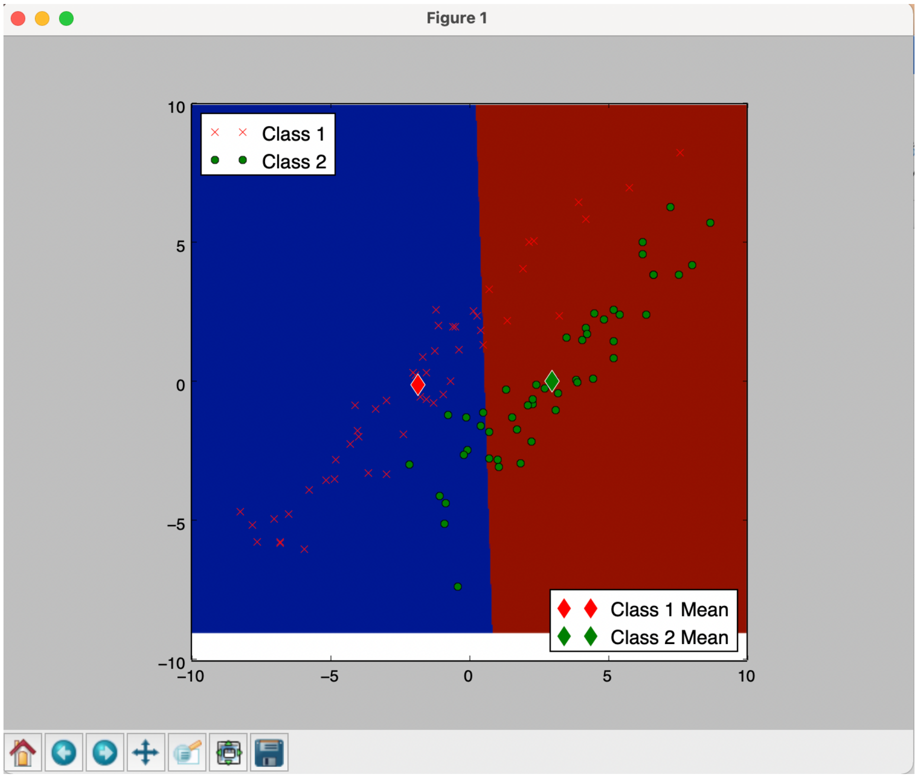
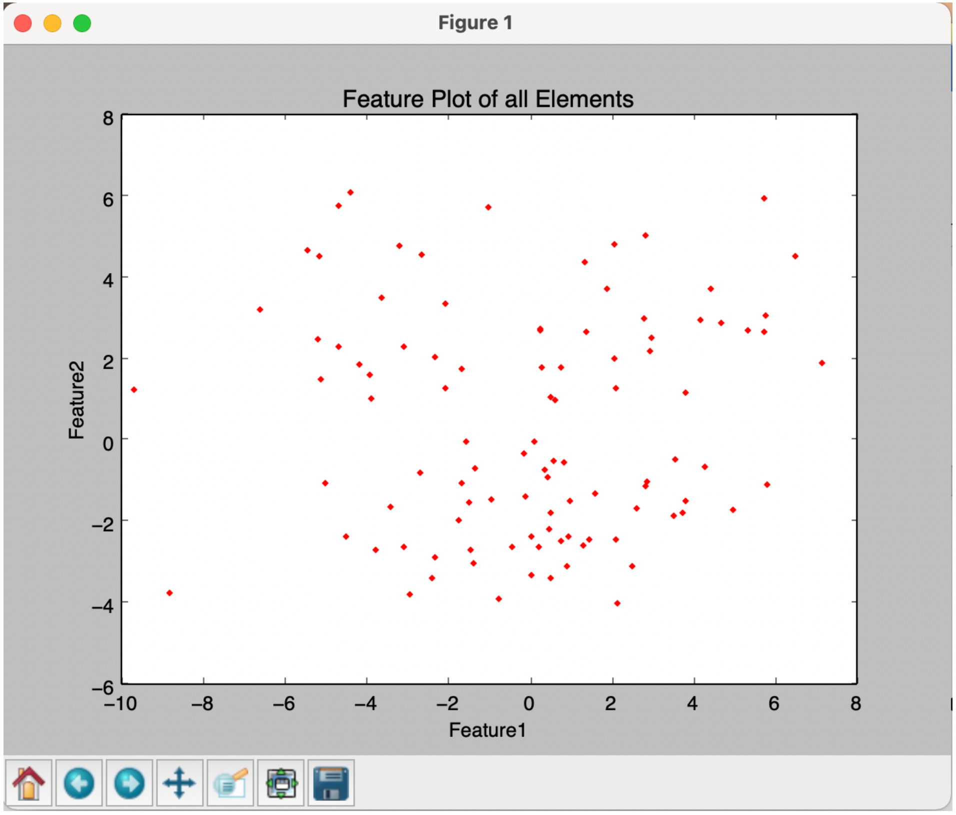
Haolun Cheng
1882563827
EE559 HW1

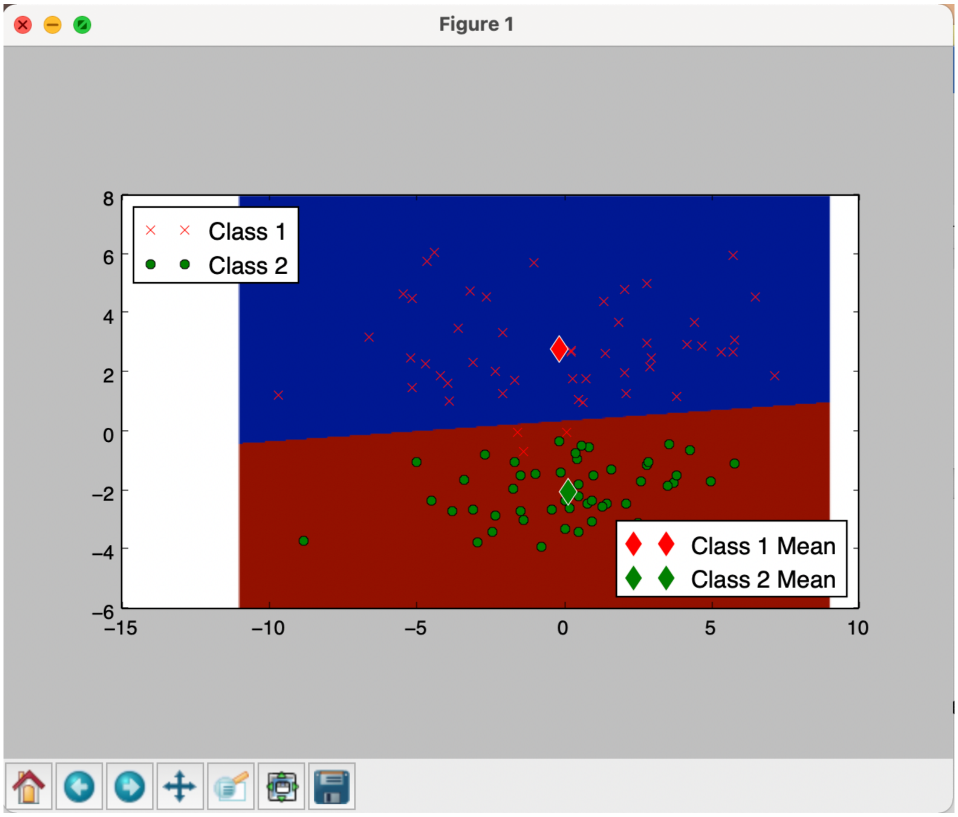
a) Synthetic dataset 1: (Error Rate: **0.21** (Training set), **0.24** (Test set))





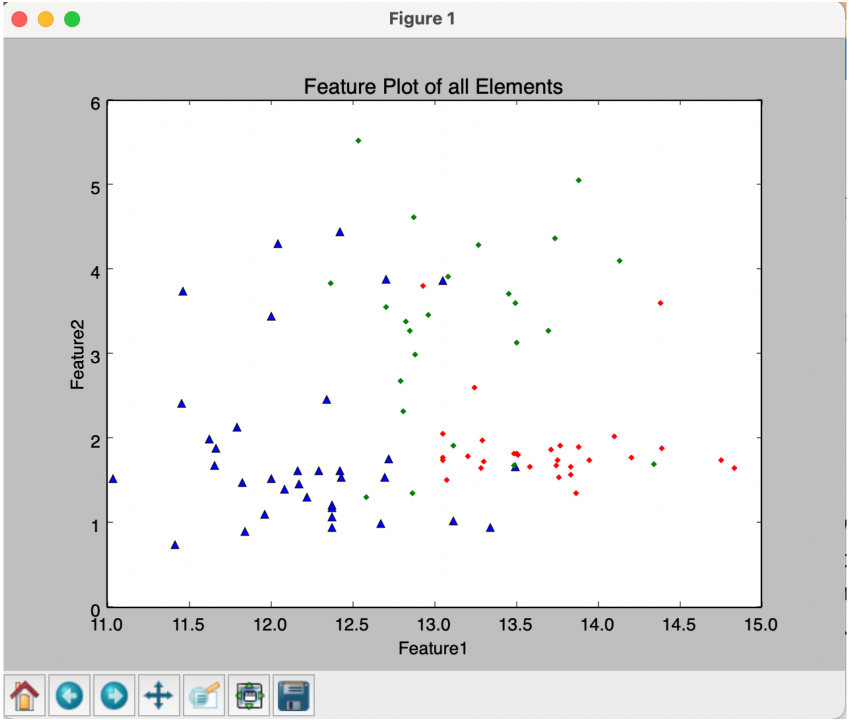
Synthetic dataset 2: (Error Rate: **0.03** (Training set), **0.04** (Test set))

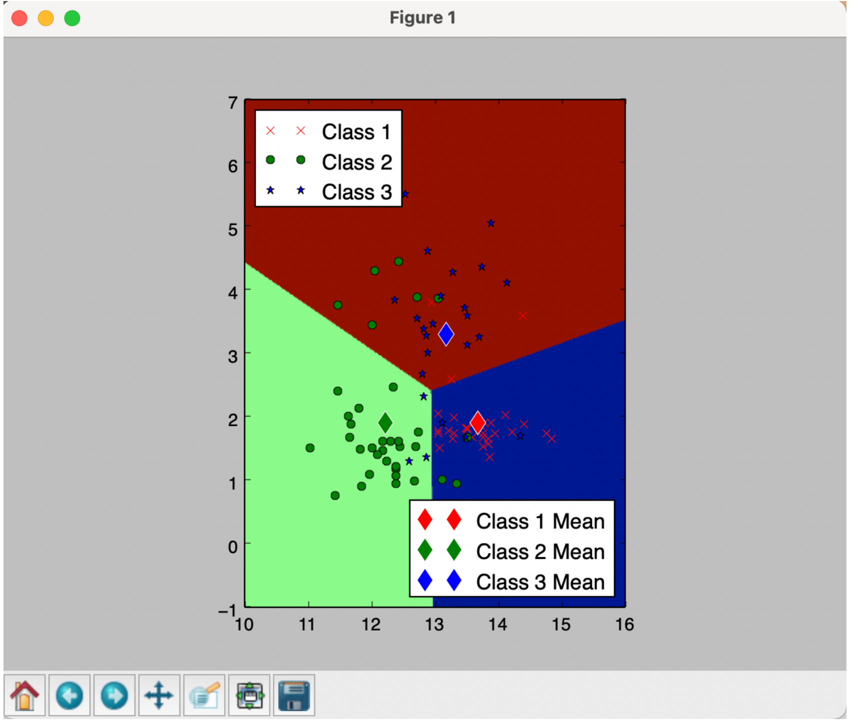




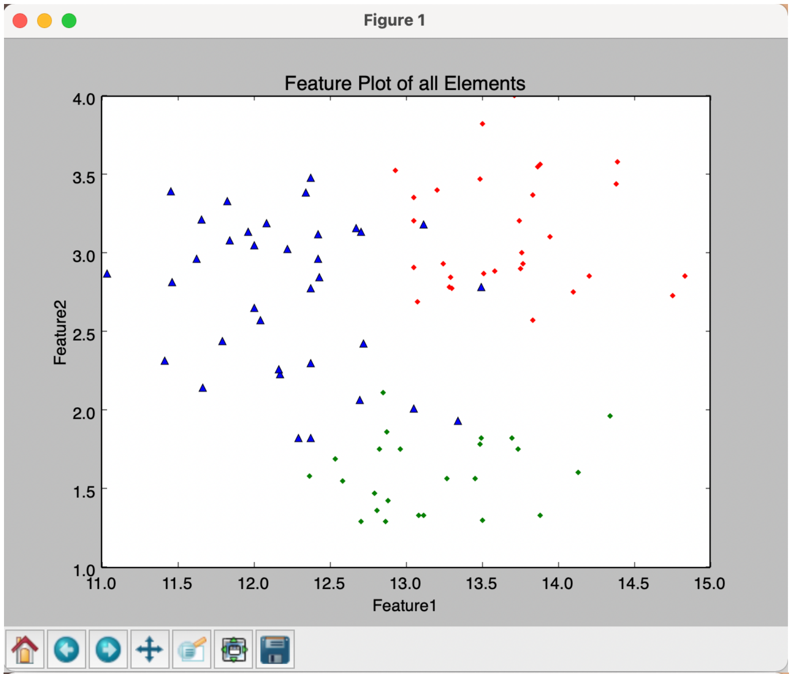
b) Yes, there is a huge difference in error rate between the two synthetic datasets. Synthetic 1 dataset has a higher error rate than that of synthetic 2 dataset. From the two plots of the decision boundaries, we can see that synthetic 2 dataset uses a much better decision boundary than synthetic 1 dataset uses. Thus, the error rate of the synthetic 2 dataset is lower, and better, than the error rate of the synthetic 1 dataset.

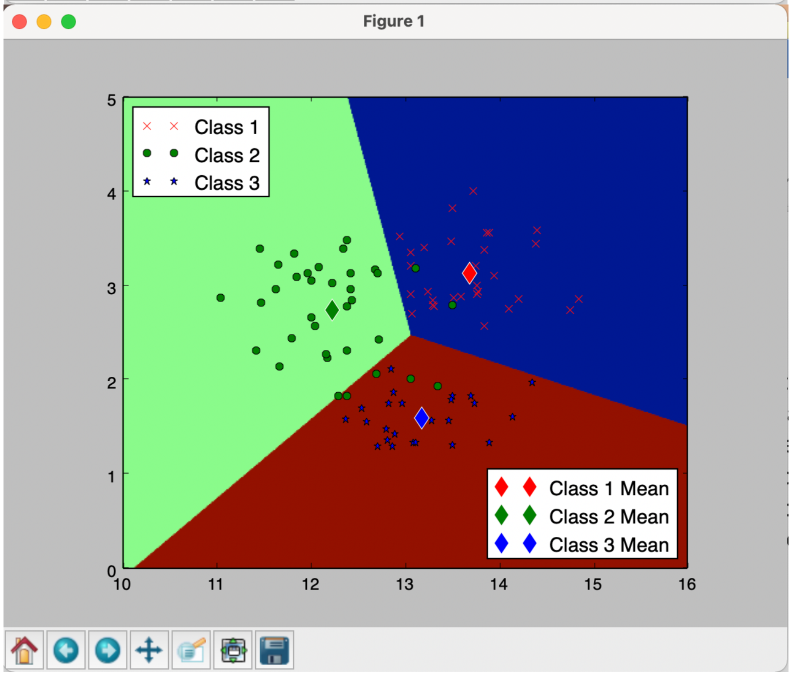
c) Wine datasets: (Error Rate: **0.202247191011** (Training set), **0.224719101124** (Test set))





d) Method for choosing the best pair of features: I used nested for loops in finding the two optimal features among the 13 provided features in the dataset. My approach is to iterate through each pair and calculate the error rate for each pair of features. When iterate through each feature pair, I used a local variable to keep recording the minimum error value such that the local variable always stores the best feature pair after the nested for loops. Finally, I used the selected feature pair for further predictions on both the Training and the Test datasets.





Features Chosen: **feature 1 Alcohol and feature 12 OD280/OD315 of diluted wines**

Error rate for the training set: **0.0786516853933**
Error rate for the test set: **0.123595505618**

e) There are huge differences in **training-set** error rate for different pairs of features for the wine dataset. From the algorithm that I wrote, the lowest error rate is around 0.07865 and the highest error rate is around 0.573. The reason behind the difference gap is the decision boundary. For some feature pairs, the decision boundary is not suitable for classifying the data points which can cause the data points to be classified into wrong class. For other feature pairs which have the relative low error rate, the decision boundary is a perfect fit during the classification process.

The reasoning also holds true for the **test-set** error rate differences for different pairs of features for the wine dataset. The lowest error rate for the test-set is around 0.12 and highest is around 0.51. The way the decision boundary is drawn largely decides the error rate for a certain pair of features.