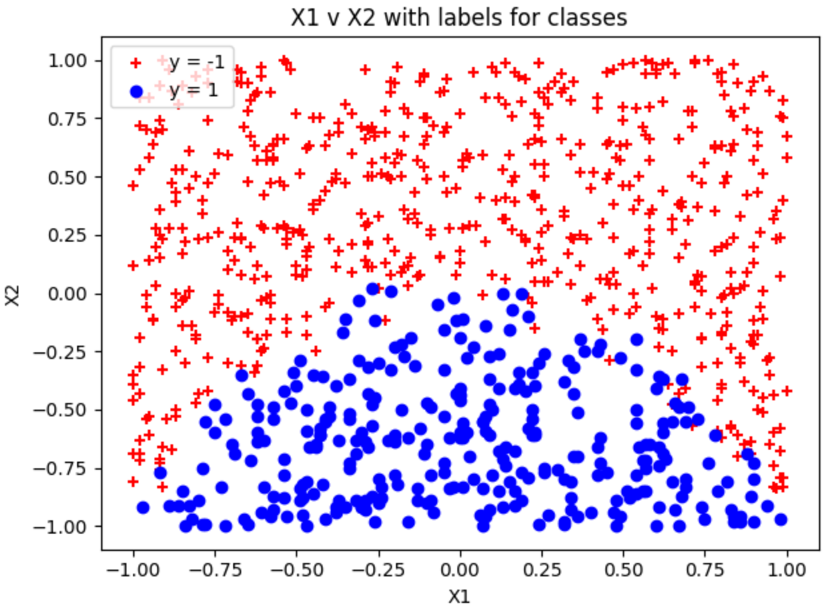
Week 4 Assignment

Efeosa Eguavoen - 17324649

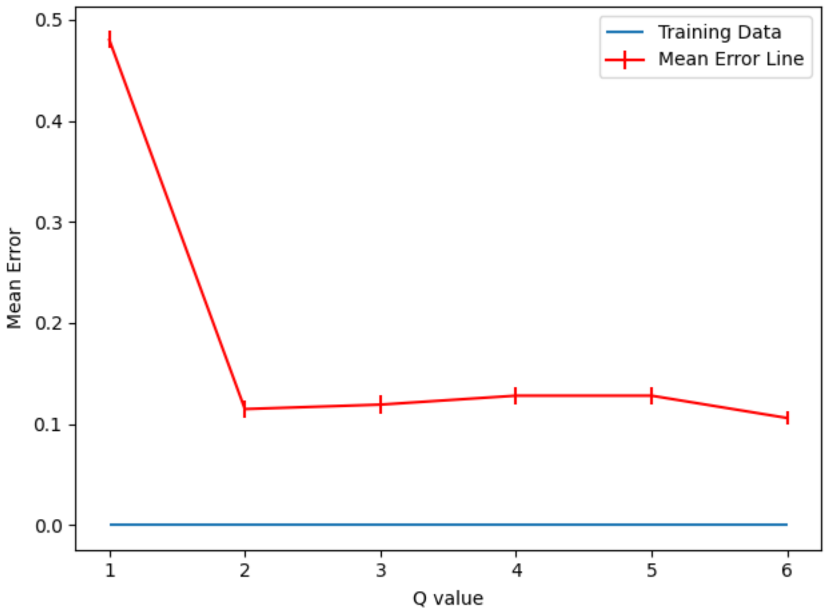
November 1, 2020

1 (i)

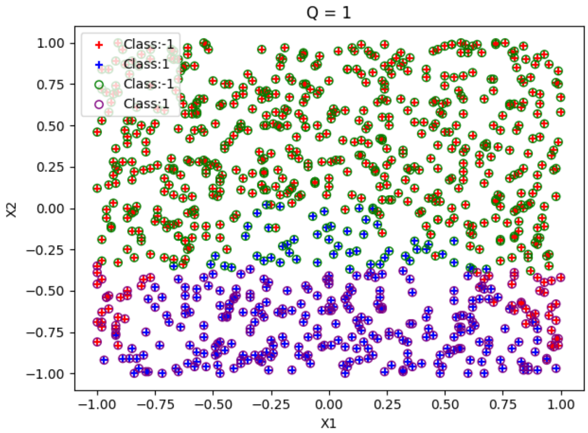
1.1 A



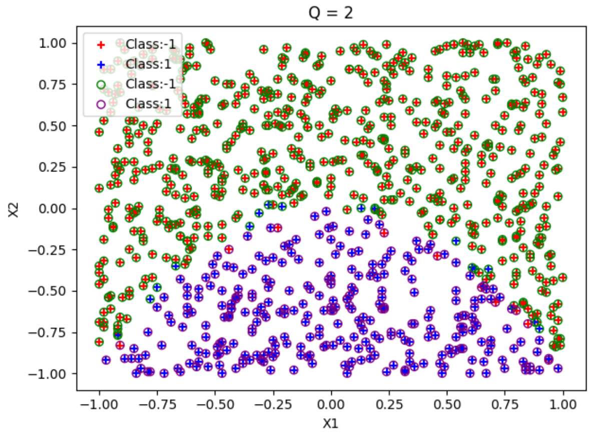
From the above plot, the data is not linealy seperable so some feature engineering will be required to get the correct decision boundary. The decision boundary I would plot based off the above plot would be some sort of quadratic line.



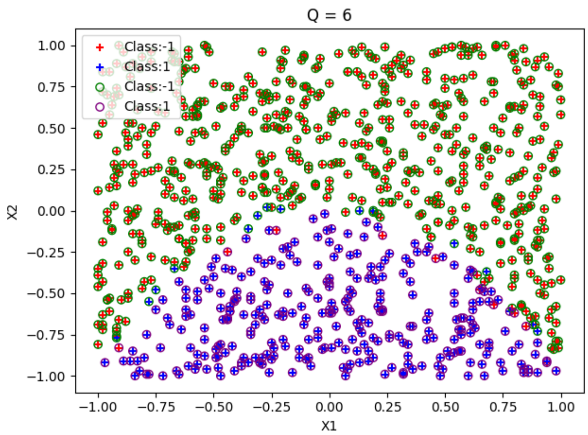
To select the correct order polynomial to use for my model, I started by looking at my graph to get an estiamate of the order of polynomial that would be required to get the best decision boundary. From the first graph, I



(a) Q =1

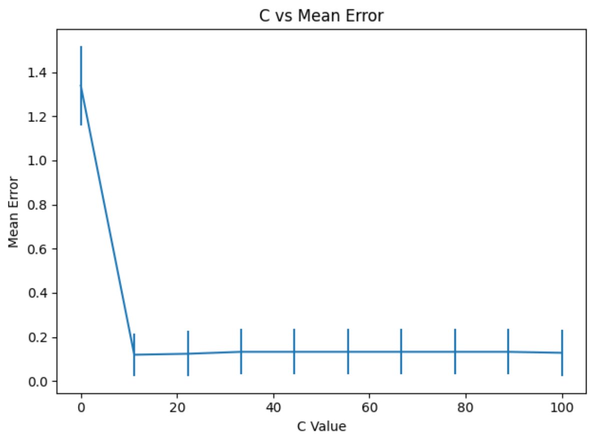


(b) Q = 2

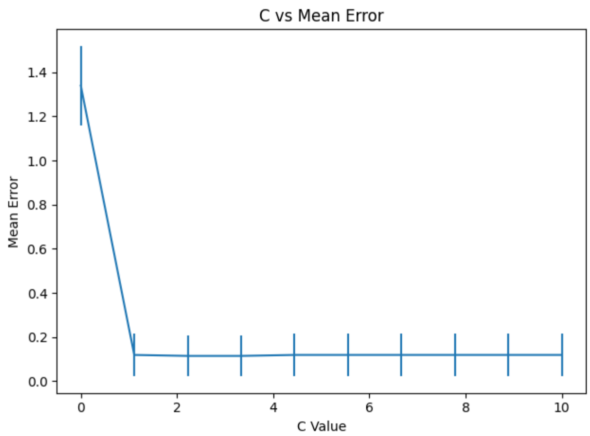


(c) Q = 6

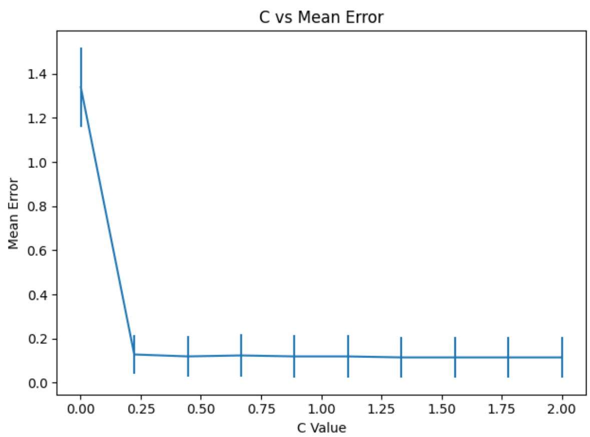
knew that it would at least need to be quadratic to I scanned a range of values between 1 and 6 as having too high an order of polynomials would lead to having too many features that would be unnecessary. From the above graph, we can see that when Q = 2 and when Q =6 the mean error is lowest. I went with Q=2 as it's better to keep the model as simple as possible. Above I've also plotted the prediction vs training for different Q values. We can clearly see when Q =1 the data on the bottom right and bottom left corner of the graph has been misclassified. In comparison, when Q =2 and when Q=6 we've gotten much better predictions and much less data has been misclassified. The difference between Q=2 and Q=6 is very little based off the graphs.



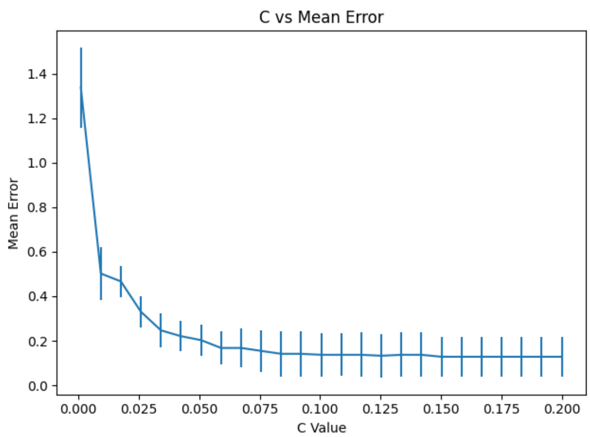
(a)



(b)



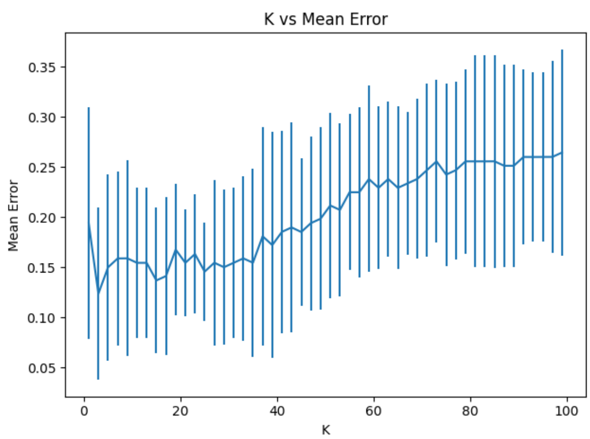
(c)



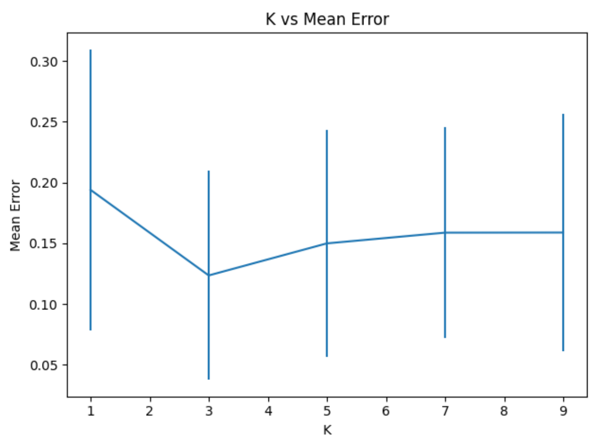
(d)

To select the correct value of C, I used a range of values of C between 0 and 100 initially to get wide enough spread to select the optimal value of C. From there I reduced the range of values further and further to get the best value of C versus the mean error. From the above graph, C = 0.15 is the optimal value to use for the model as it reduces the error down the most. Higher values of C seem to keep the error around roughly the same value so I just went with the smaller value for simplicity.

1.2 B



(a) K- Large Range



(b) K - Small Range

To select a value for K, I first did some research online about guidelines for calculating K. From what I read, I saw k=sqrt(n). I used this as a baseline for the range of values to search over, so I searched between 1 and 100 initially and from there reduced the range down further. I used cross validation for each value of K, with a k value of k=10. Based on the graphs, above, K = 3 is the optimum value of K to use for the given dataset. The mean error when K = 3 is lowest other than when K = 15 which also has a low error rate. Choosing K = 3 is best as it keeps the model simpler.