Mini Project 1

Jose Alfaro

January 23, 2019

Question 1: Consider the training and test data posted on eLearning in the files *1-tranining-data.csv* and *1-test-data.csv*, respectively, for a classification problem with two classes.

Part A: Fit KNN with $K = 1, 2, \ldots, 30, 35, \ldots, 100$.

Part B: Plot training and test error rates against K. Explain what you observe. Is it consistent with what you expect from the class?

When observing the error rates for K with values 1 through 100, it seemed that the testing error rates was consistently higher than the training error rates as K increase. However, when I was searching for the "U" shape pattern that is supposed to appear in the test error rates, I failed to find a trend that resembled the form of a minimum (Figure 1). To further investigate this issue, I ran the KNN algorithm for K with values 1 through 450 in hopes of finding a definite "U" shaped pattern in the data that. As a result, I was provided with a bigger picture as to what was happening with the test error rates as K increased and found a definite "U" shape in the data (Figure 2). I then reexamined the test error rates in greater detail and discovered that the highest error occurred when K = 1 (error = 0.228) and the lowest error occurred when K = 86 (error = 0.167). This implies that K = 86 is indeed the optimal K value for this data. After reading a series of articles relating to the KNN algorithm, I learned that a general rule of thumb for selecting an optimal K value is to take the square root of the number of observations in the training dataset. In this case, our training dataset consisted of 2,000 observations which lead us to conclude that the optimal K value for this dataset would be somewhere around K = 45. However, after closely following the trends in the data and searching for the K value associated with the minimum test error, I found that the optimal K value was actually K = 86. Therefore, my findings were not consistent with my initial intuition, but I suspect that my value for K will be about the same as the class.

Part C: What is the optimal value of K? What are the training and test error rates associated with the optimal K?

As discusses in Part B, the optimal value of K produced by the KNN algorithm was $K = 86$ and is associated with a test error rate of 0.167 as well as a train error rate of 0.162.

Part D: Make a plot of the training data that shows the decision boundary for the optimal K. Comment on what you observe. Does the decision boundary seem sensible?

The decision boundary plot can be seen in *Figure 3* and shows where the algorithm determines if an observation should be categorized as a "yes" response or as a "no" response. This seems relatively sensible since our dataset consisted of exactly 1,000 observations that had a response of "yes" and exactly 1,000 observations that had a response of "no". Therefore, this almost clear-cut behavior down the middle of the data does not seem out of the ordinary.

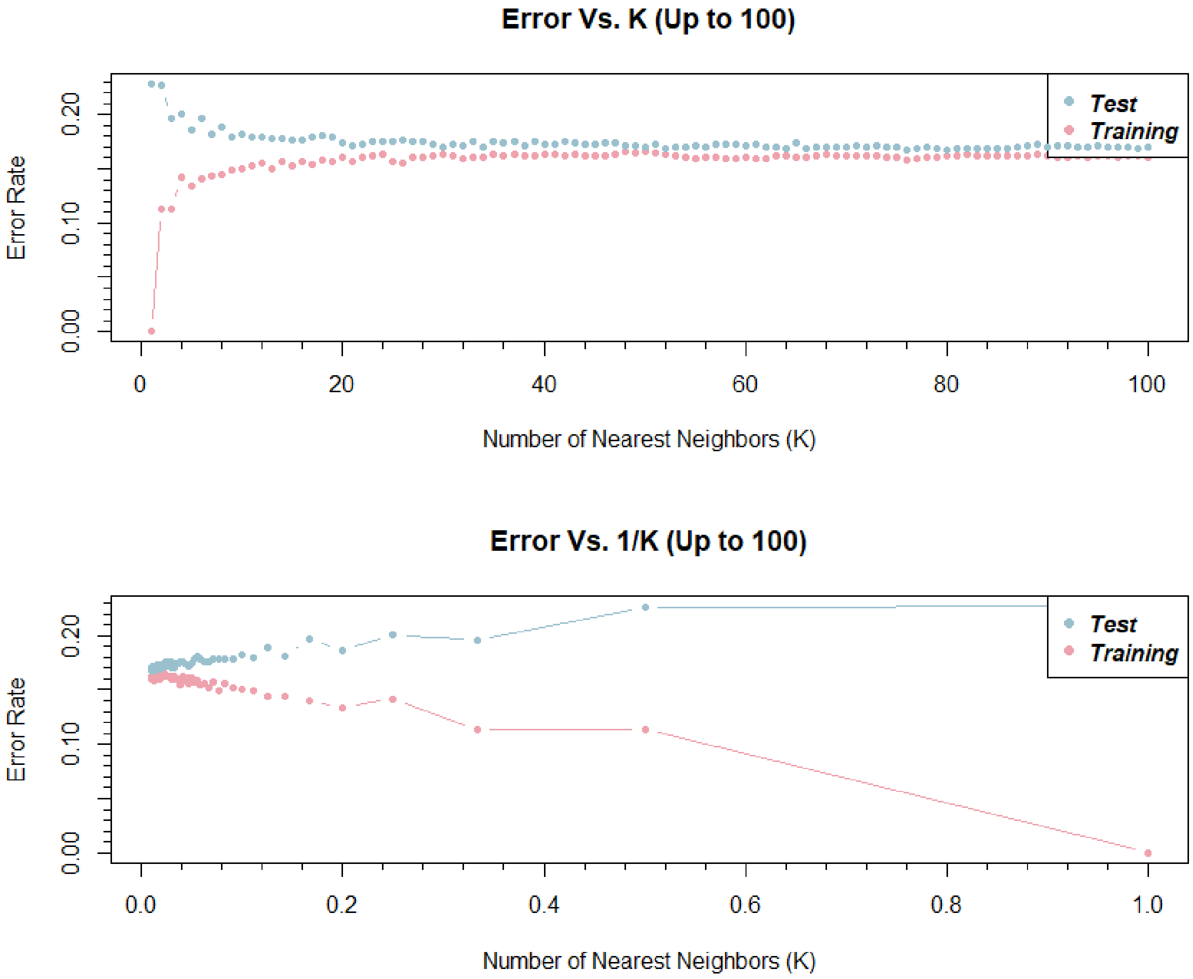

Figure 1: Error Vs. K (With K Between 1 and 100)

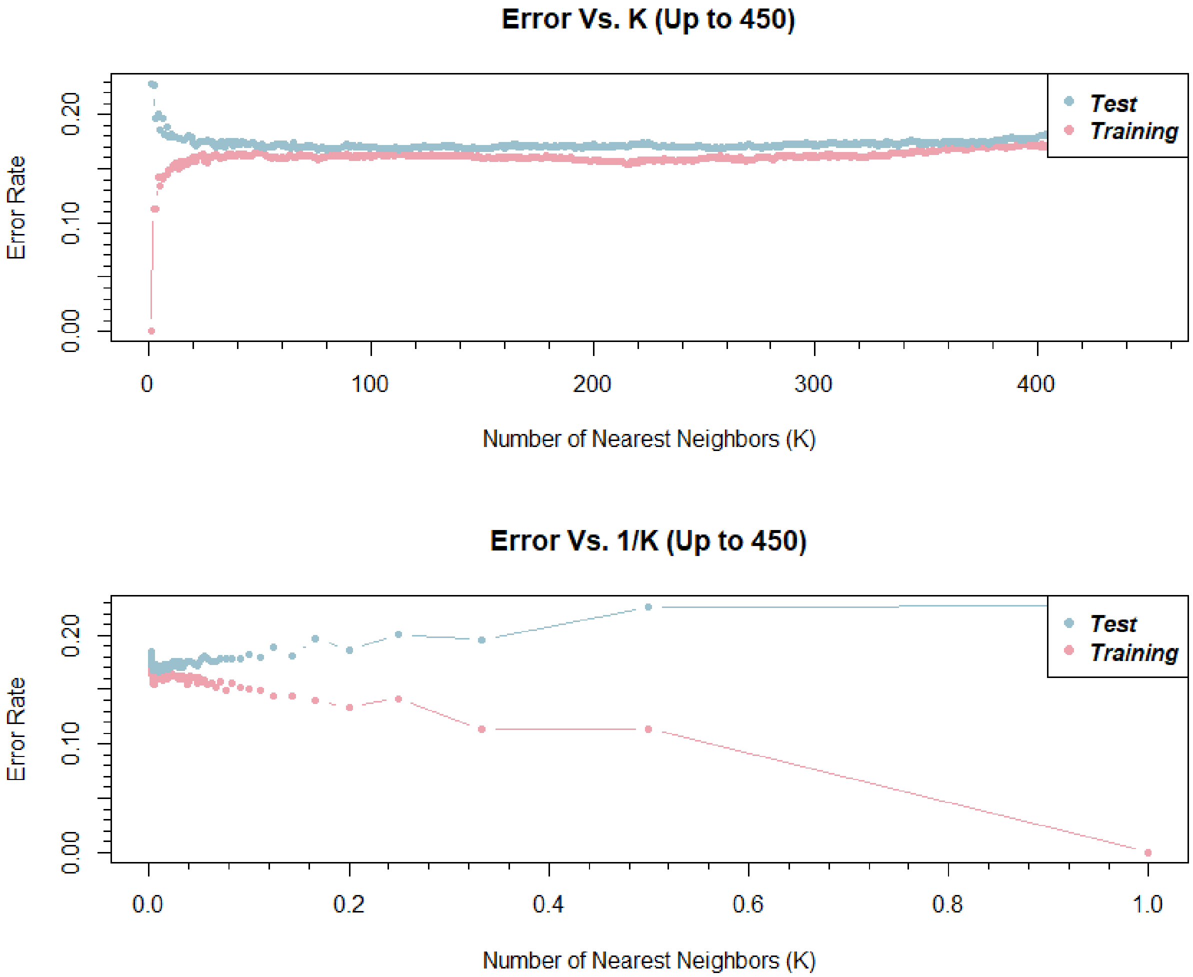

Figure 2: Error Vs. K (With K Between 1 and 450)

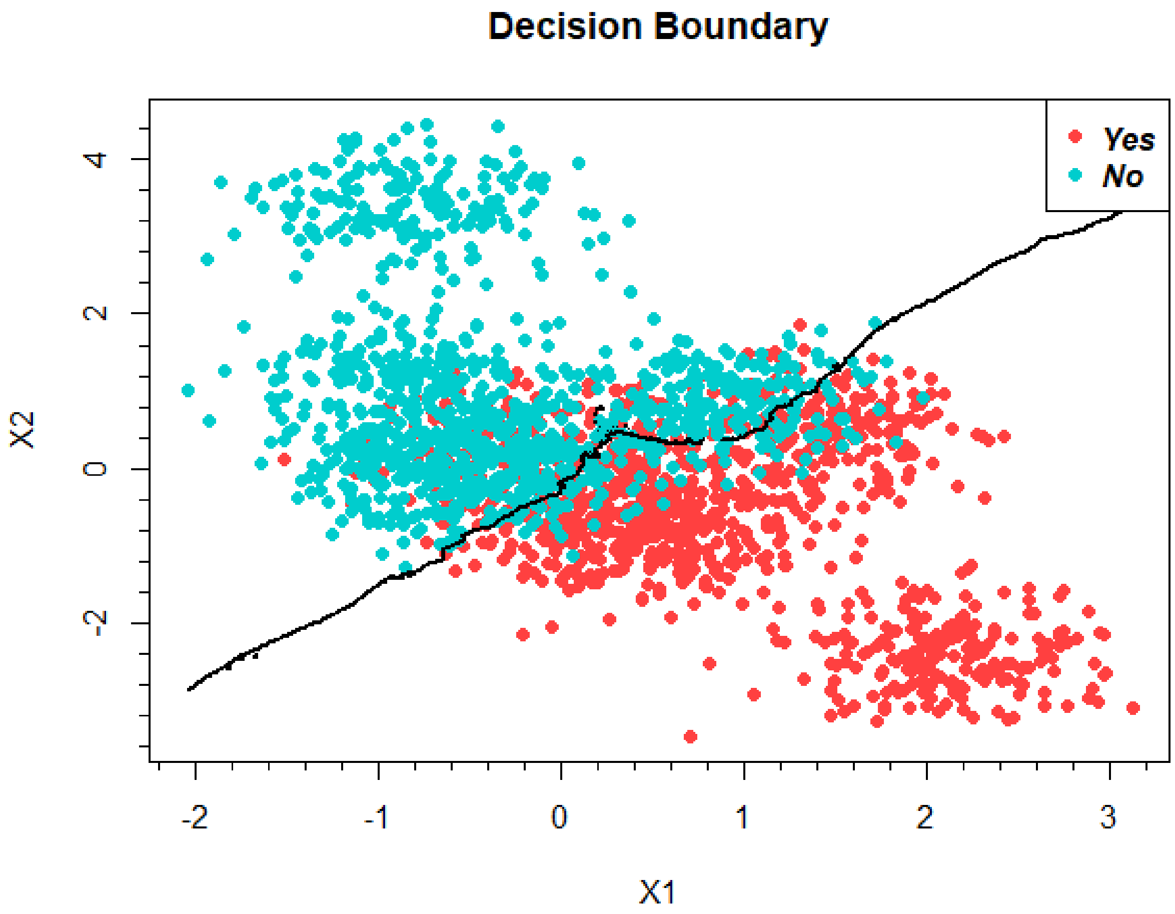

Figure 3: Decision Boundary for K = 86