COMPENG 4SL4

Assignment 3

Instructor: Dr. Dumitrescu

Hritheekka Chinnakonda – chinnakh – 400292782 – C01 – L03

As a future member of the engineering profession, the student is responsible for performing the required work in an honest manner, without plagiarism and cheating. Submitting this work with my name and student number is a statement and understanding that this work is my own and adheres to the Academic Integrity Policy of McMaster University and the Code of Conduct of the Professional Engineers of Ontario. Submitted by [Hritheekka Chinnakonda, chinnakh, 400292782]

Specify the model that you deem to be the best and indicate the test error. Justify your choice.

The first plot displays cross-validation and training error from k = 1 to 405, the entire dataset. It was observed that the best model was when k = 2, which yielded a test error of 33.95007352941176. I then limited k = 100, which provides a clearer and more limited graph. Overall, this model had a good bias-variance trade-off and the lowest test error. The model overfits after approximately k = 125, as cross-validation and training error continue to increase as k increases. When looking at the plot on the right (seen below) the curves plateau around k = 75 to k = 125, the model underfits here.

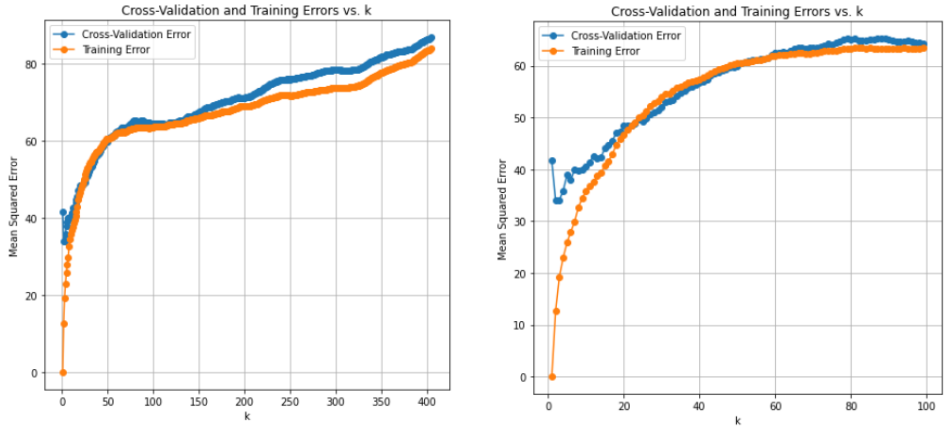

Figure 1: Cross validation and training errors vs k., k = 405 (left), k = 100 (right)

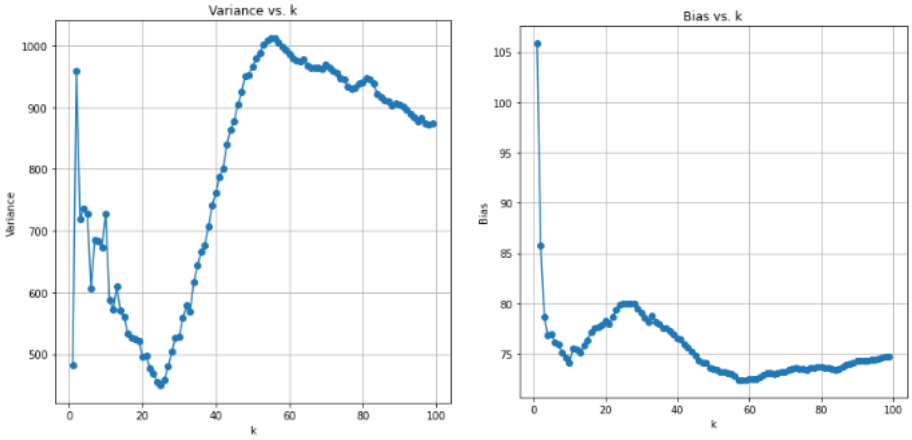

Figure 2: Variance and Bias vs k., k = 100

When looking at the bias and variance plots, the variance starts rapidly increasing between approximately k = 25 to k = 55, while bias decreases at a slower rate between the same range. This indicates a point where the model transitions from a state of underfitting to overfitting, the bias-variance trade-off. The model is becoming more capable of fitting the training data (decrease in bias), and the model is becoming overly sensitive to small variations in the training data, fitting the noise, resulting in a higher-cross validation error. This is seen in the cross-validation and training error plot as the curves rapidly increase.

Further testing was done by removing the NaN values, and standardizing the data in the dataset to see how it would affect the plots and see if the best k-NN model would change. The resulting plot is seen below, and the best model was k = 3 with a test error of 9.337746623093683. The curves are constantly increasing, after a short plateau around k = 40.

Figure 3: Cross validation and training errors vs k., removed NaN, k = 203 (left), k = 100 (right)