COMP ENG 4SL4: Machine learning

Assignment 2 - Nearest Neighbours and Cross-validation

Instructor: Dr. Sorina Dumitrescu

TA: Zewei Zhang

Richard Qiu – qiur12 – 400318681

K nearest neighbours simulation (k-NN)

To find the regression of k-NN, the average of the y values is used as the prediction points based on the range of k value, $1 \leqslant k \leqslant 60$. Below are my simulation graphs for every model.

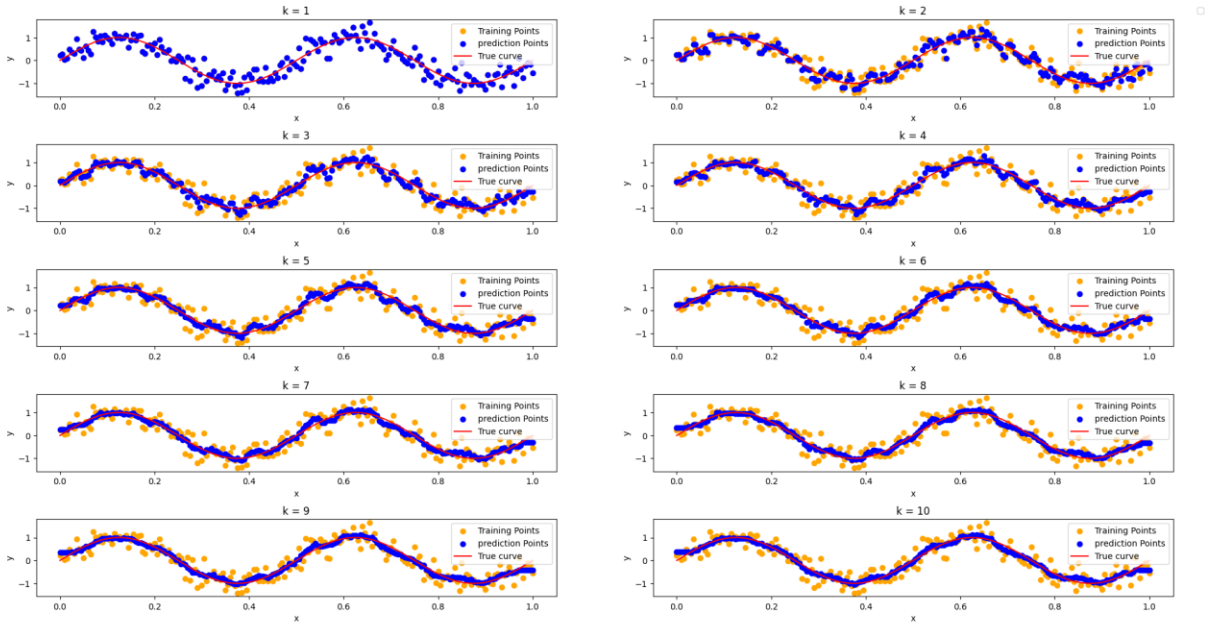

Figure 1 $1 \leqslant k \leqslant 10$

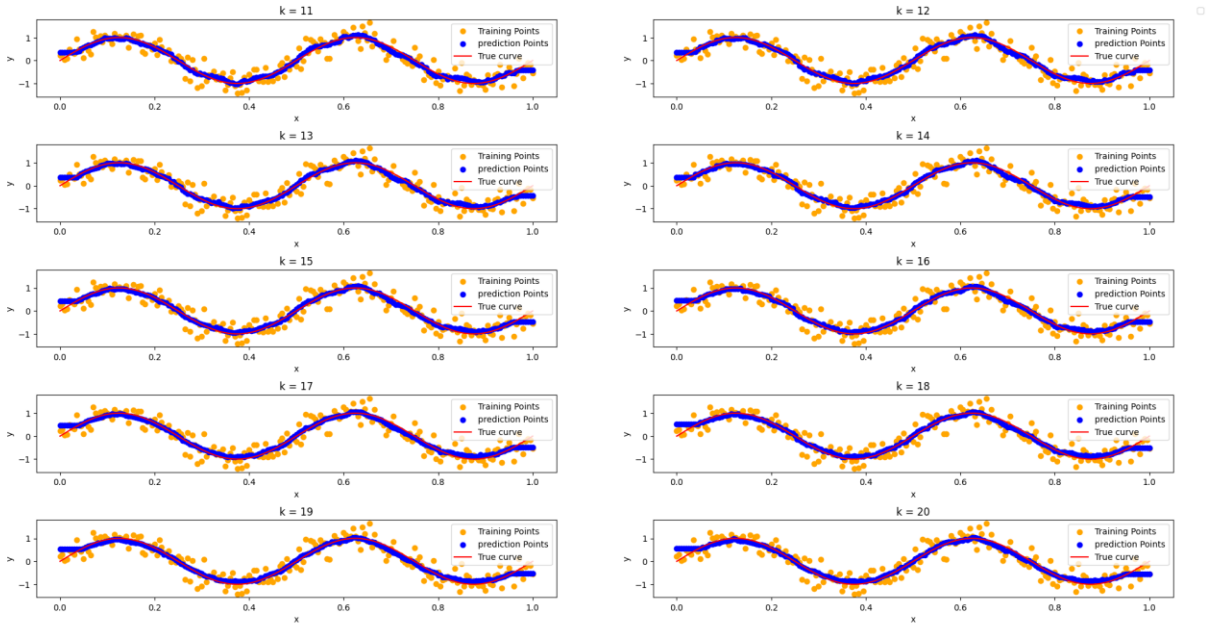

Figure 2 $11 \leqslant k \leqslant 20$

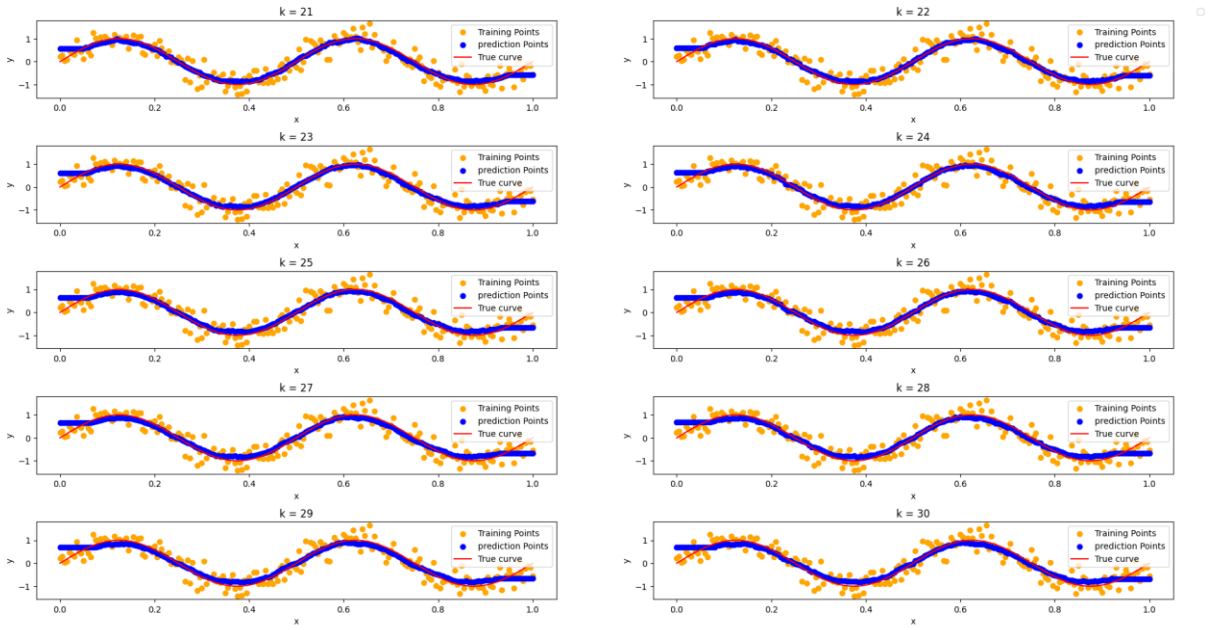

Figure 3 $21 \leqslant k \leqslant 30$

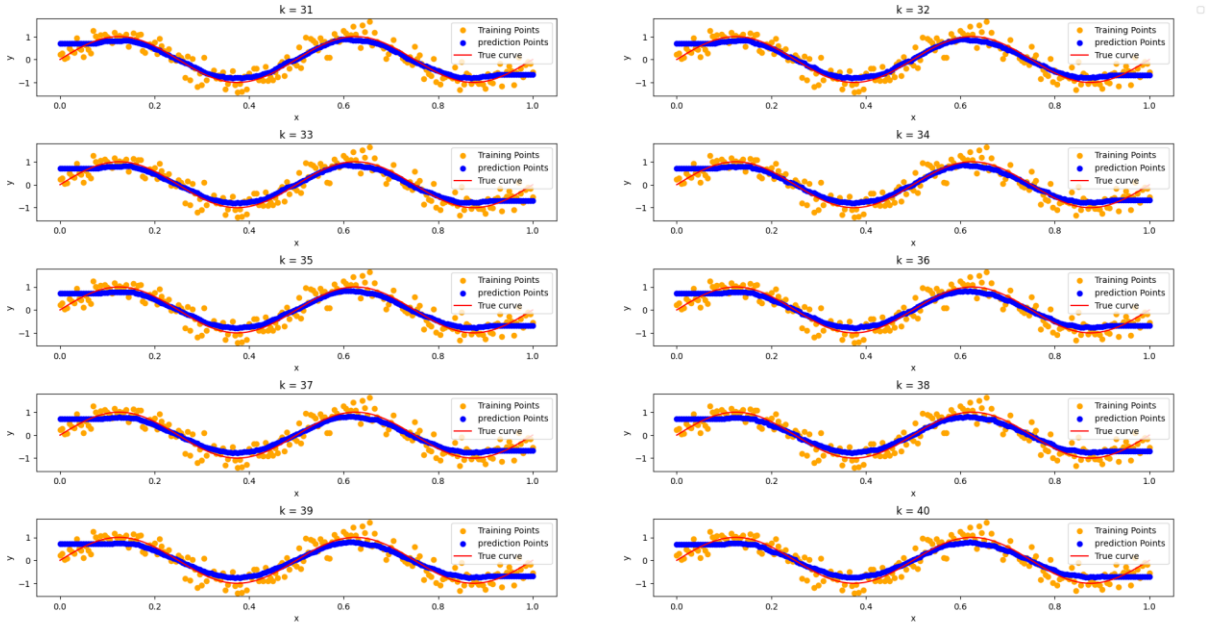

Figure 4 $31 \leqslant k \leqslant 40$

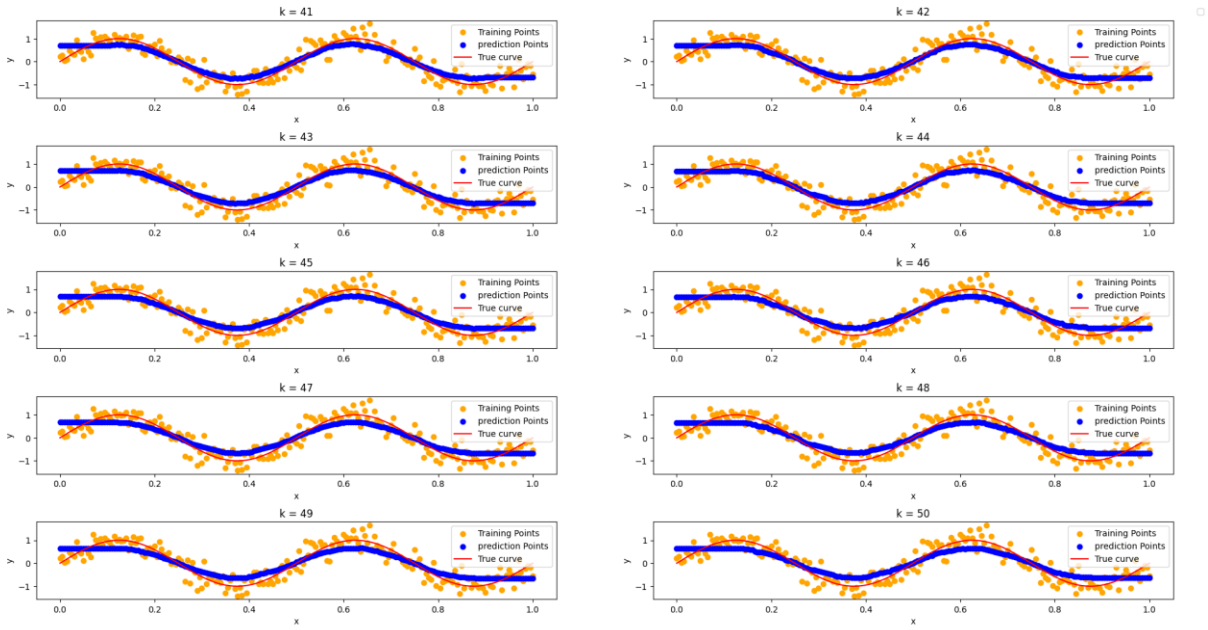

Figure 5 $41 \leqslant k \leqslant 50$

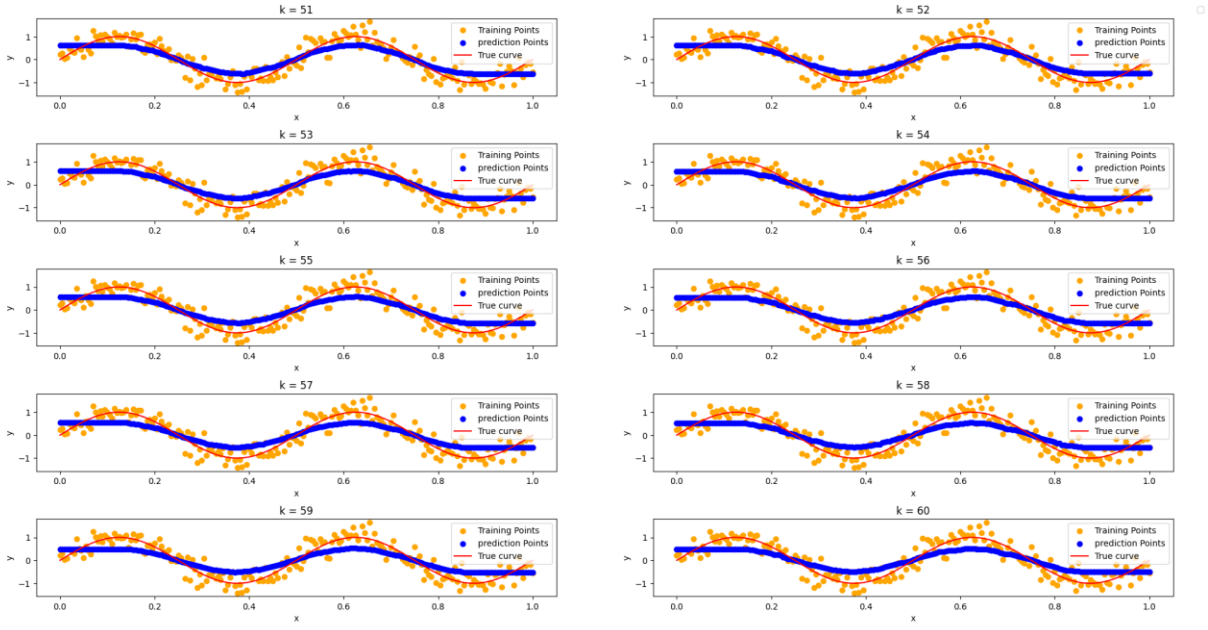

Figure 6 $51 \leqslant k \leqslant 60$

As we can see from figure 1 when k = 1, the predictions points are exactly the same as the training points which is why the graph only shows the blue points.

Training and Validation MSE Errors

Below is the graph after performing 5-fold cross validation on the training sets with k values range from 1 to 60.

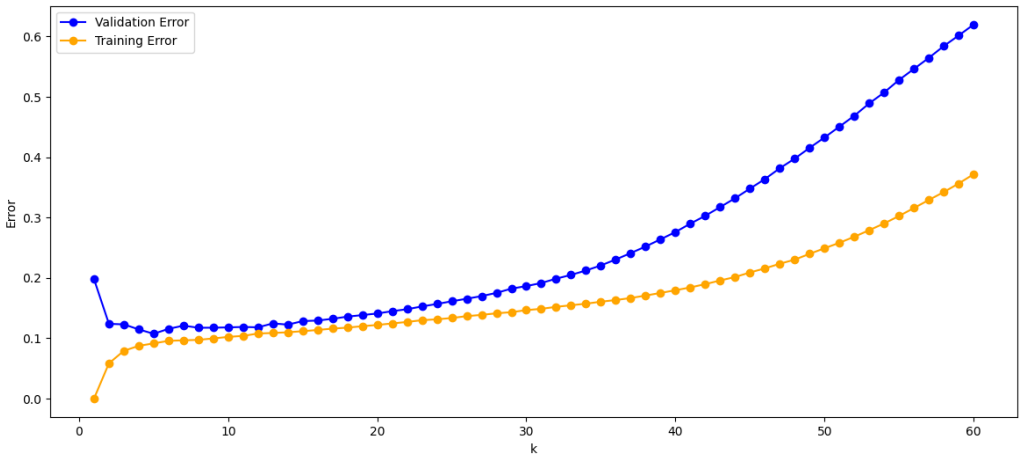

Figure 7 5-fold cross validation, Training vs Validation Error

To analyze this graph.

Overfitting (0 – 5): As when K is small, the KNN prediction points will likely overfit due to it's too sensitive to the training data and fails to generalize to new data. And it can be proved that the validation errors don't perform well when K is between this region.

Optimal Capacity (5 – 15): This is the region where the difference between training error and validation error are small as we can seen from the graph. The model also maintains a good generalization capability.

Underfitting (20 – 60): When K is greater than 20, the model becomes over simple and fails to capture the necessary relationship in the data. From the graph, we can see both validation and training errors get increased, and their difference gets increased as well.

Best model testing

Based on figure 7, we can observe that when K = 12, the difference between training and validation errors are the lowest, hence this value will be chosen for the best k-NN model.

Below is the graph when K = 12.

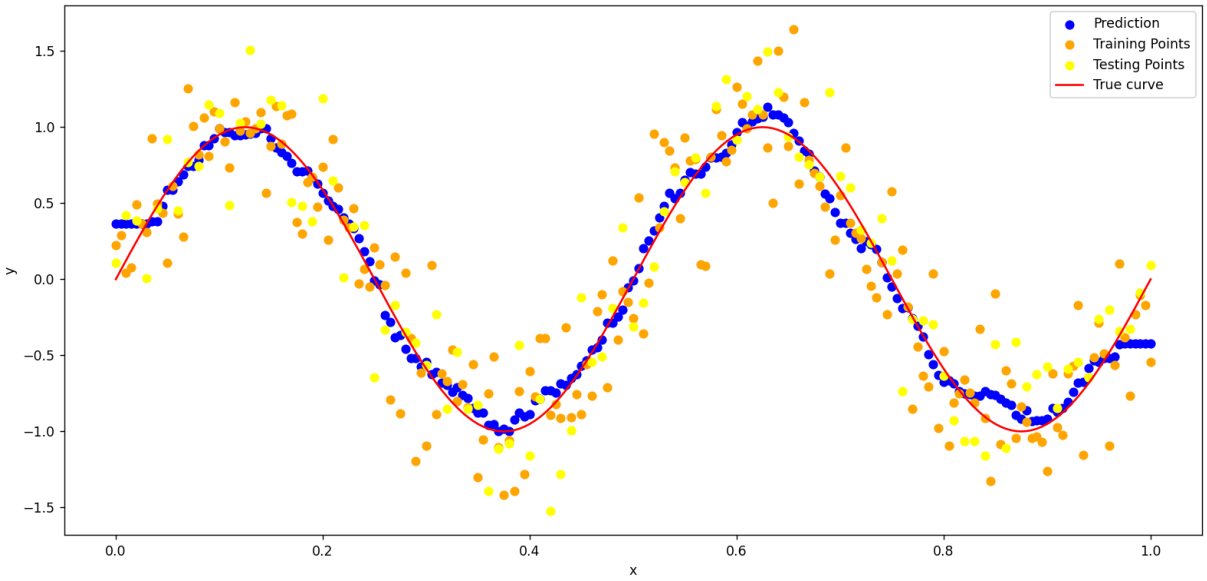

Figure 8 K = 12, prediction

Performance compare

After train the model with the full training sets and calculate the testing error on the test sets, the MSE is 0.0876 and RMSE is 0.295941871259781 when K = 12.

Compare with the RMSE we obtain from assignment 1, the best model is when the degree of M = 7 with 10 training points and the RMSE is equal to 0.5115211774171243.

In summary, the k-NN algorithm seems like perform better by comparing the RMSE, but it is also important to notice that the in assignment1, we only use 10 data points for training sets and we use 200 data points for training sets in this assignment. And also k-NN algorithm is computational expensive as it may not be suitable for larger data sets.