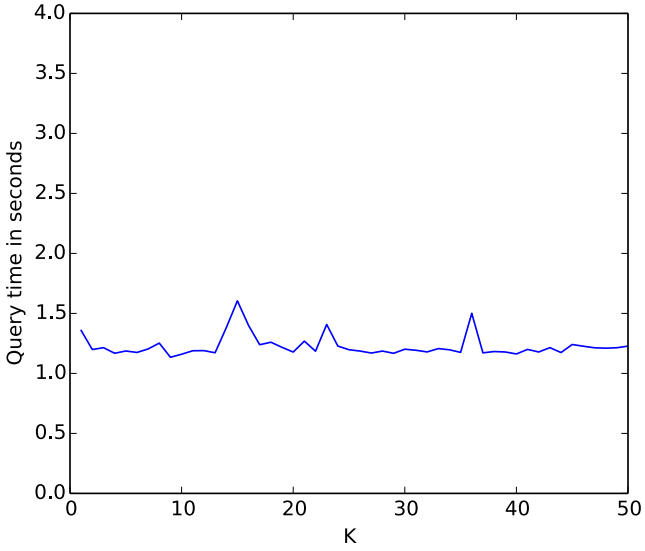
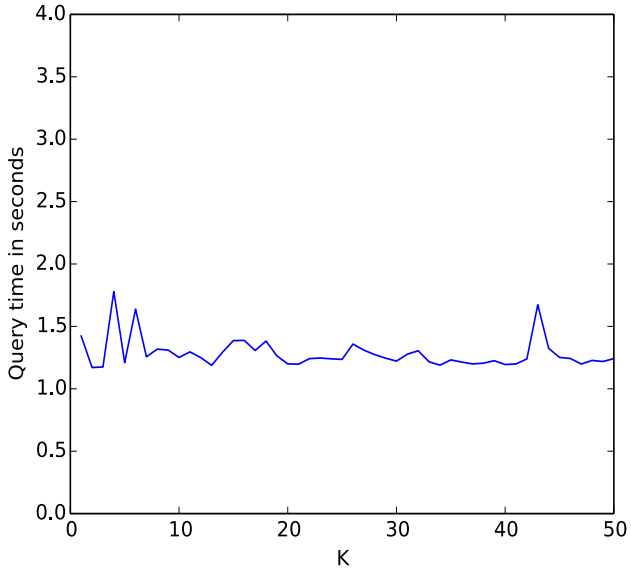
Machine Learning for Trading
CS 7646
Project 1 Report
Sriram Madapusi Vasudevan smv6@gatech.edu GT ID : 902916994

KNN Learner:

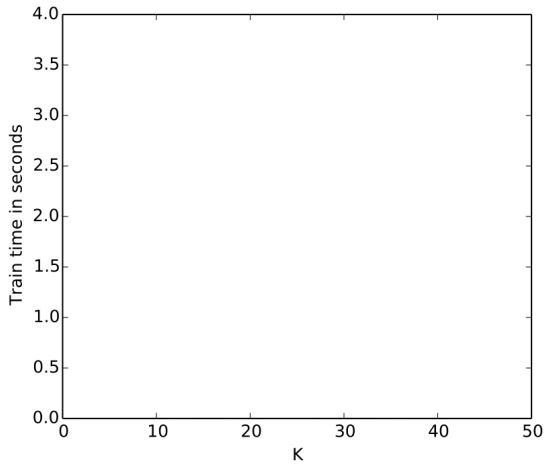
Query time: (Data classification (left) Daripple (right))

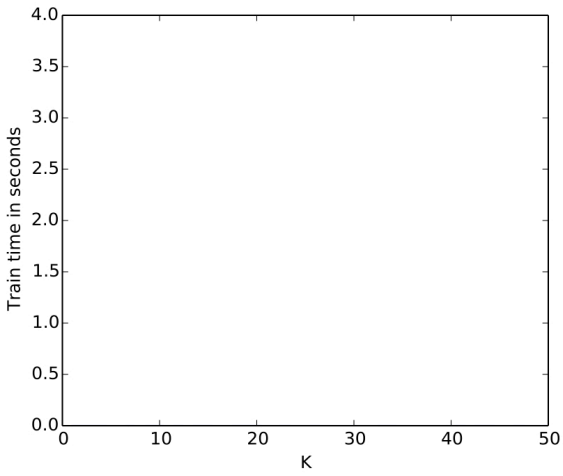




Query time is roughly constant for different K

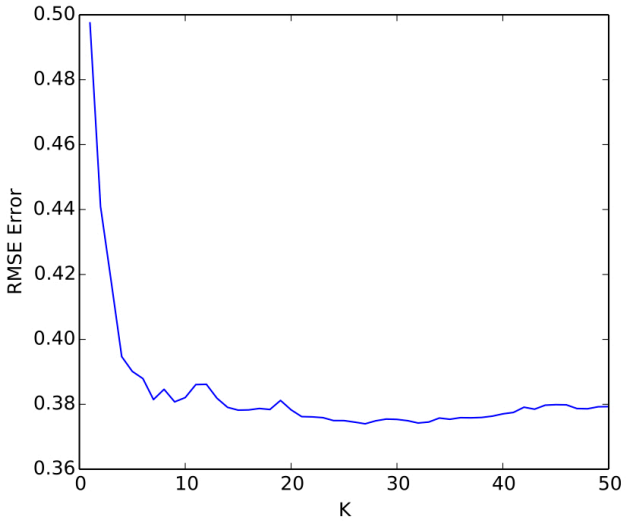
Train Time: (Data classification (left) Daripple (right))

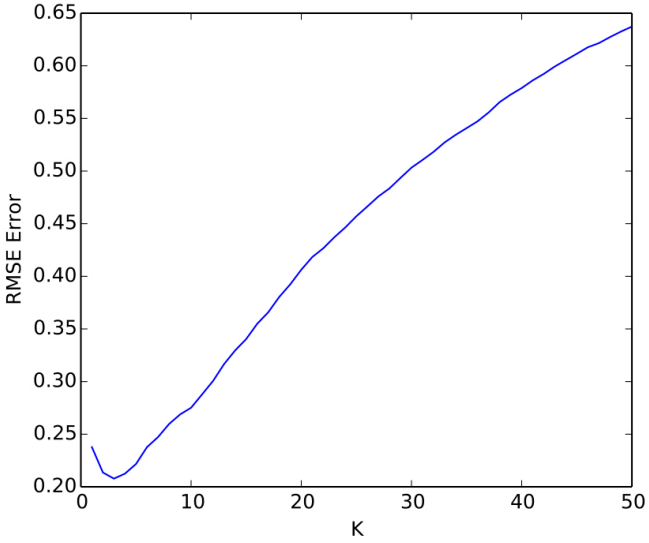




The training time is zero seconds for all K in both cases as we just loading the training dataset into the object of the KNN Learner class.

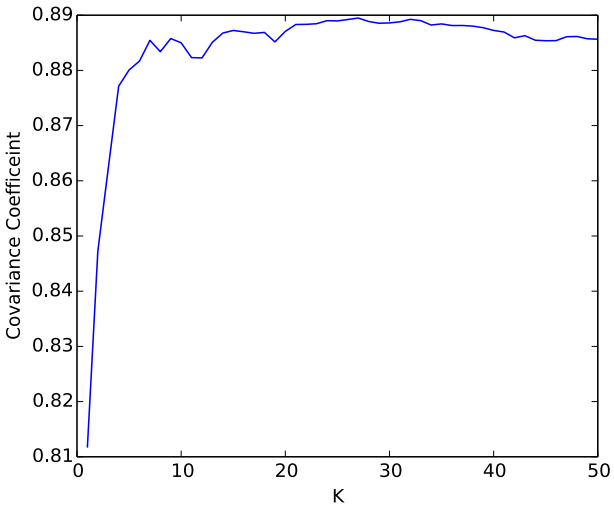
RMSE Error: (Data classification (left) Dataripple (right))

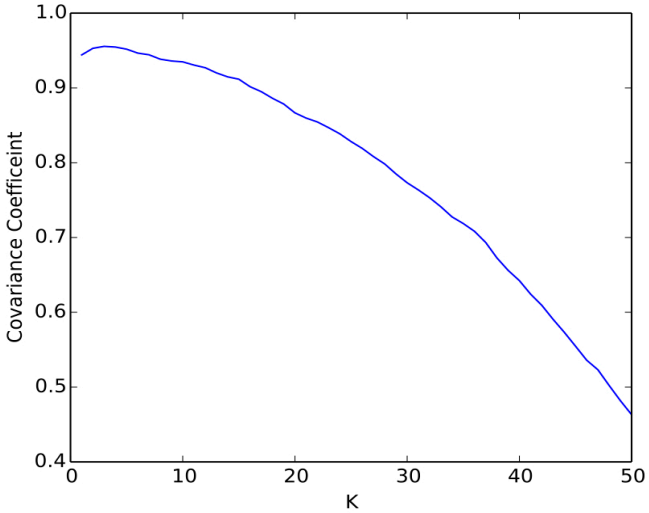




It be can be seen here that data classification data set lowers RMS error with increased k whereas data ripple data set has increased RMS for increase in k. It can perhaps be reasoned that the learner is better performing for the data classification data set than the data ripple data set.

Correlation : (Data classification (left) Dataripple (right))

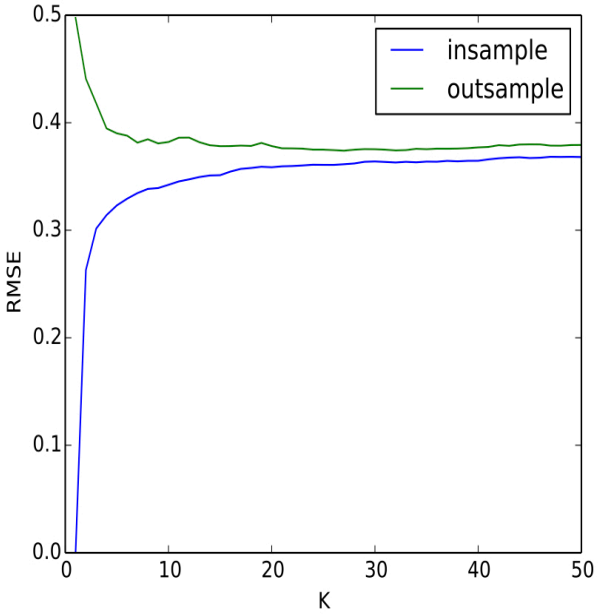


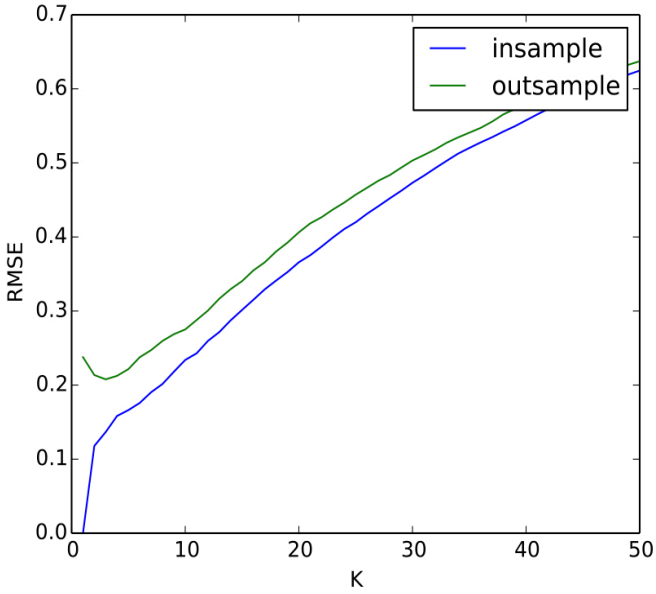


There is a higher level of correlation for the data classification dataset than the data ripple data set.

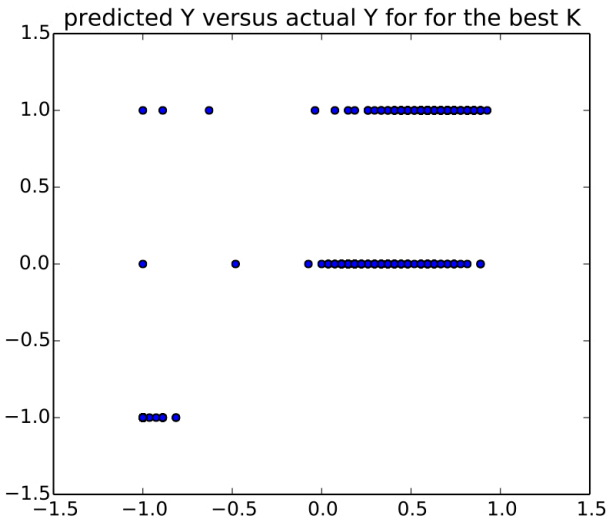
In-sample error vs out-sample error:

(Data classification (left) Dataripple (right))

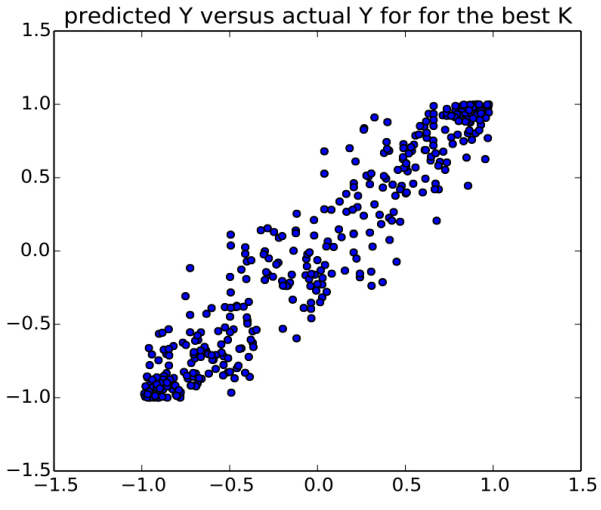




Scatter plots: (Data classification (left) Dataripple (right))

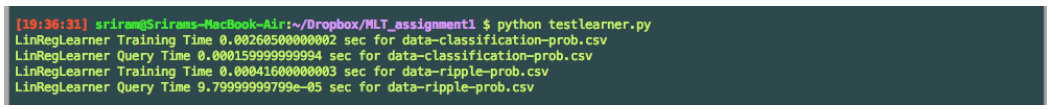


k=27



k=3

Linear Regression :



Linear Regression Learner Training time was 0.002 sec on data classification dataset
Linear Regression Learner Training time was 0.0004 sec on data ripple dataset

The Query time was close to zero on both data sets.

What is the best k for each dataset?

The best k for each data set was determined by finding at which k the rmse error was lowest. Conversely where the correlation was the highest would have also lead to the best K . The method I had chosen was to use the least root mean squared error.

The best k for data classification data set was 27 and was 3 for the data ripple data set.

As K decreases, does overfitting occur for the datasets? At approximately which K does it start? Explain why you think this is occurring?

Data classification data set:

Overfitting occurs for the data classification dataset when K is low, i.e it does not output the best results (high RMSE Error), instead showcasing the noise/random error in the dataset. But some point after k>10 RMSE error drops lower and the overfitting stops.
This is further verified by looking at the insample vs outsample plot for the data classification data set.
So there is probably overfitting between 1<k<10

Data ripple data set:

For the data ripple data set, overfitting is not as high as when k is low, but is present somewhere in the range till k<=3.
But the insample vs outsample plot for the data ripple set reveals that testing using either test data or training data doesn't really provide a good solution as both tend to have increased RMSE error as k increases. This might just be the nature of the dataset.

Why does overfitting occur?

Overfitting generally occurs because we are training our model on a training data set, and then testing its performance on a different test data set. Usually, the test data set consists of some unseen data.

References:

Stackoverflow (ideas on using numpy)
Piazza (1 liner of code for RMSE, ideas for using numpy libraries)