Report

Yuliang Li
GTID# 903012703

The experiment results of KNN and Linear Regression are shown in Table 1.[1]

Table 1 Experiment results of KNN and Linear Regression

	Ave train time	**Ave query time**	**RMS Error**	**Correlation Coefficient**
KNN Classification (K=3)	0 ms	0.23 ms	0.4180	0.8622
KNN Classification (K=27, best)	0 ms	0.38 ms	0.3740	0.8895
LinReg Classification	0 ms	0 ms	0.5154	0.7749
KNN Ripple (K=3)	0 ms	0.28 ms	0.2078	0.9555
KNN Ripple (K=3, best)	1.6 ms	0.28 ms	0.2078	0.9555
LinReg Ripple	0 ms	0 ms	0.7041	0.0163

The average train time versus K of KNN for the two data sets is shown in Fig 1.

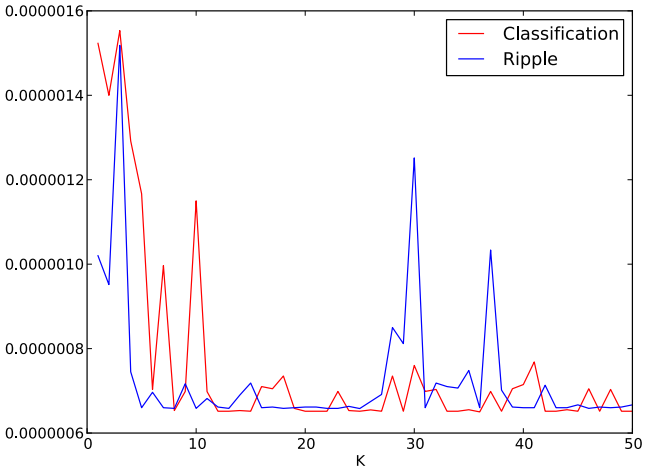

Fig 1 Train time versus K of KNN

[1] In the report, classification data is for data-classification-prob.csv, ripple data is for data-ripple-csv.

The average query time versus K of KNN for the two data sets is shown in Fig 2.

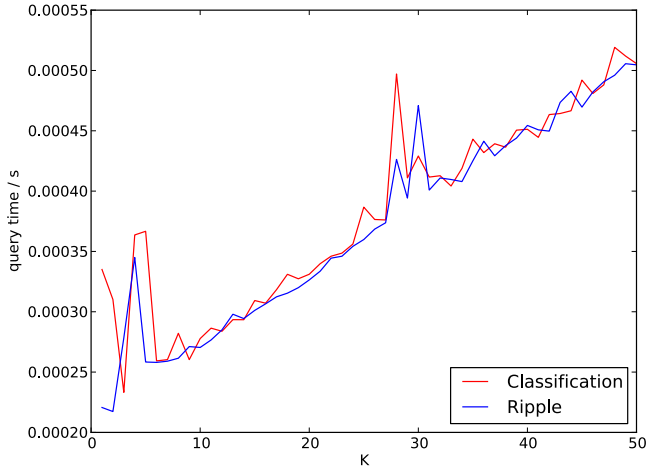

Fig 2 Query time versus K of KNN

The correlation coefficient of the response from the learner versus the correct response for two data sets is shown in Fig 3.

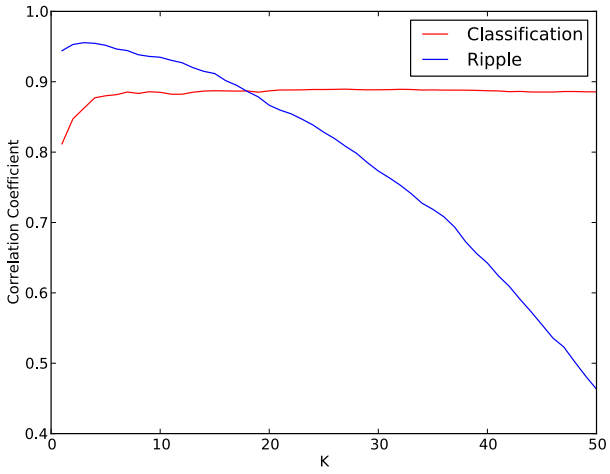

Fig 3 Correlation coefficient of the responses from learner and observation

Another indicator to judge the learner is RMS error. In the experiment, RMS error is used to measure the in-sample error and out-of sample error. In sample data is the train data, and the out-of sample data is the test data. The out-of sample error is the error of the learner when querying the test data. In sample error versus out-of sample error of two data sets are compared in Fig 4.

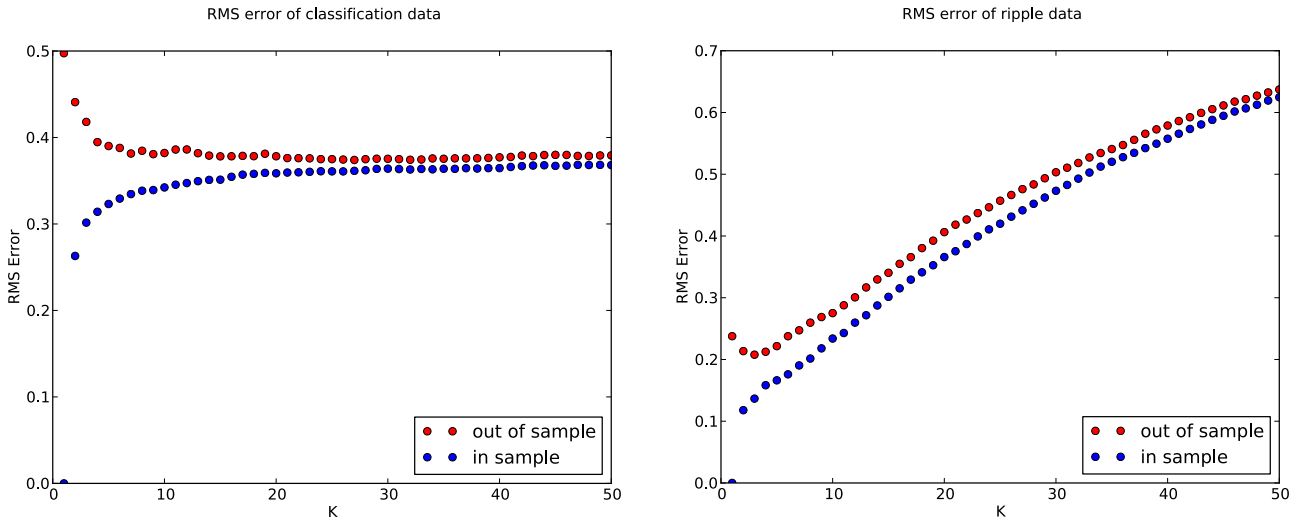

Fig 4 The RMS error of in sample and out-of sample

In Fig 4, as K decreases, both figures show that out-of sample error increases and in-sample decreases. That means the overfitting occurs for both data sets. The reason that overfitting occurs is, when K is so small that we don't need the data more than query data themselves to get the response. When K is small, the number of neighbors we can use to get the average is small, and it will cause more error than using the query data themselves.

Hence, to choose the best K for the learner, we should consider this error. For classification data, overfitting occurs when K is below 16, and for ripple data the minimum appropriate K is 3.

In KNN experiments, the best K is that gets the maximum correlation coefficient of the response from the learner versus the correct response (the 40% out of sample data). The larger correlation coefficient means the results from learner are more approaching to the actual observed value, and it indicates the learner is more successful. Although we need to consider the time costs of query and training, that time costs are small enough in this experiment because computers are much faster than before. Thus, we choose the correlation coefficient as the main indicator to decide the best K. With the consideration on overfitting, the best K to test classification data is 27, while 3 for ripple data.

The predicted Y versus actual Y for the best K are plotted in Fig 5. Concluded from the figure, learner for querying the ripple data is better than the classification data for the classification data consists of only 1, 0 and -1.

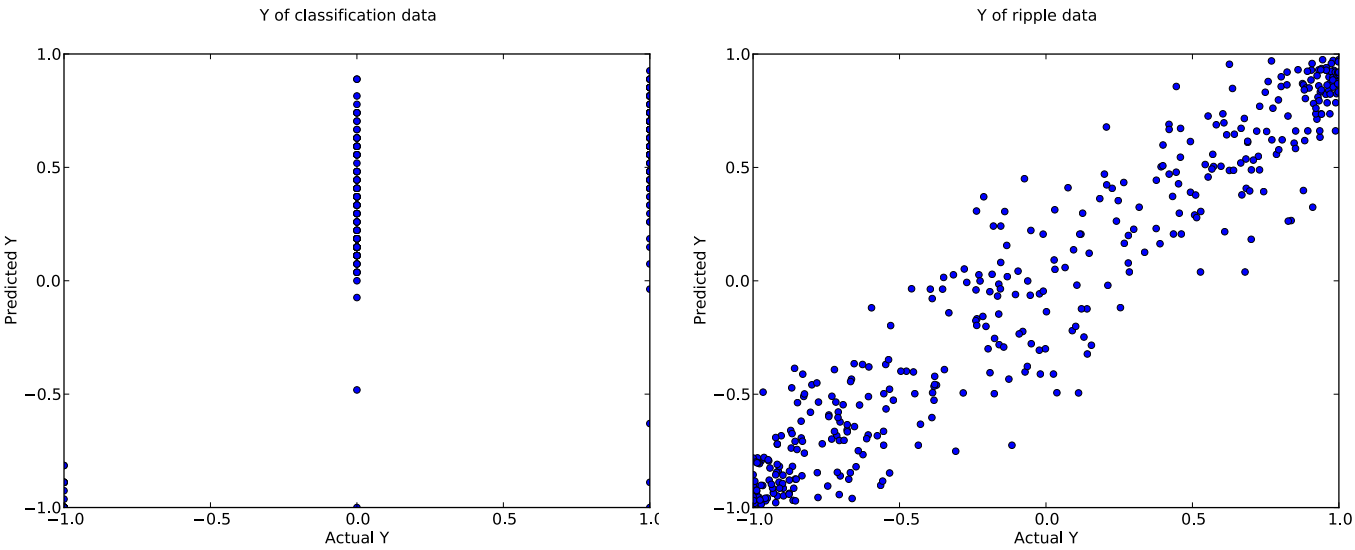

Fig 5 Predicted Y versus actual Y when K is best

The linear regression learner get the fit line for train data, and the equation of the regression line is Eq (1) and Eq(2). The former one is for the classification data, and the latter one is for the ripple data.

$$y = 1.142x_1 - 1.260x_2 + 0.393 \qquad \text{Eq(1)}$$

$$y = -0.043x_1 - 0.026x_2 - 0.090 \qquad \text{Eq(2)}$$

The RMS error and correlation coefficient for two data sets have been shown in Table 1.