Akshat Harit
GtId:903090915

Data-classification-prob.csv
RMS Error
Insample : 0.51881446463
Outsample : 0.515389956744
Correlation
OutSample : ', 0.77494317678056546)
InSample : ', 0.77213130680578901)

data-ripple-prob.csv
RMS Error
Insample : 0.705375081106
Outsample : 0.70409341685
Correlation
OutSample : 0.0162663457175
InSample : 0.040607011093

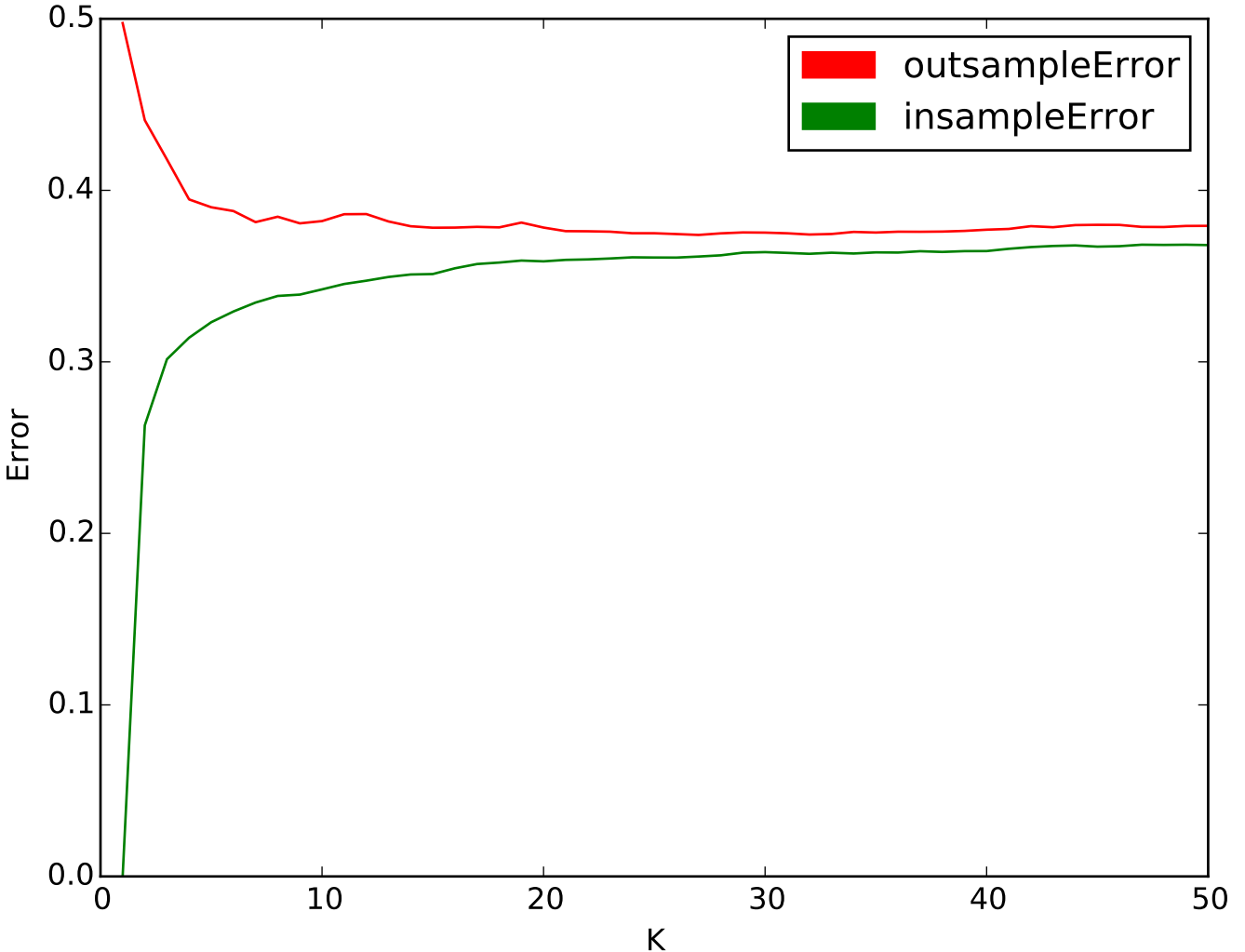
What is the "best" K for each dataset? Explain your reasoning. Note that there is not necessarily a single correct answer. I want to see your reasoning.
For data-ripple
K=3
For data-classification
K=27

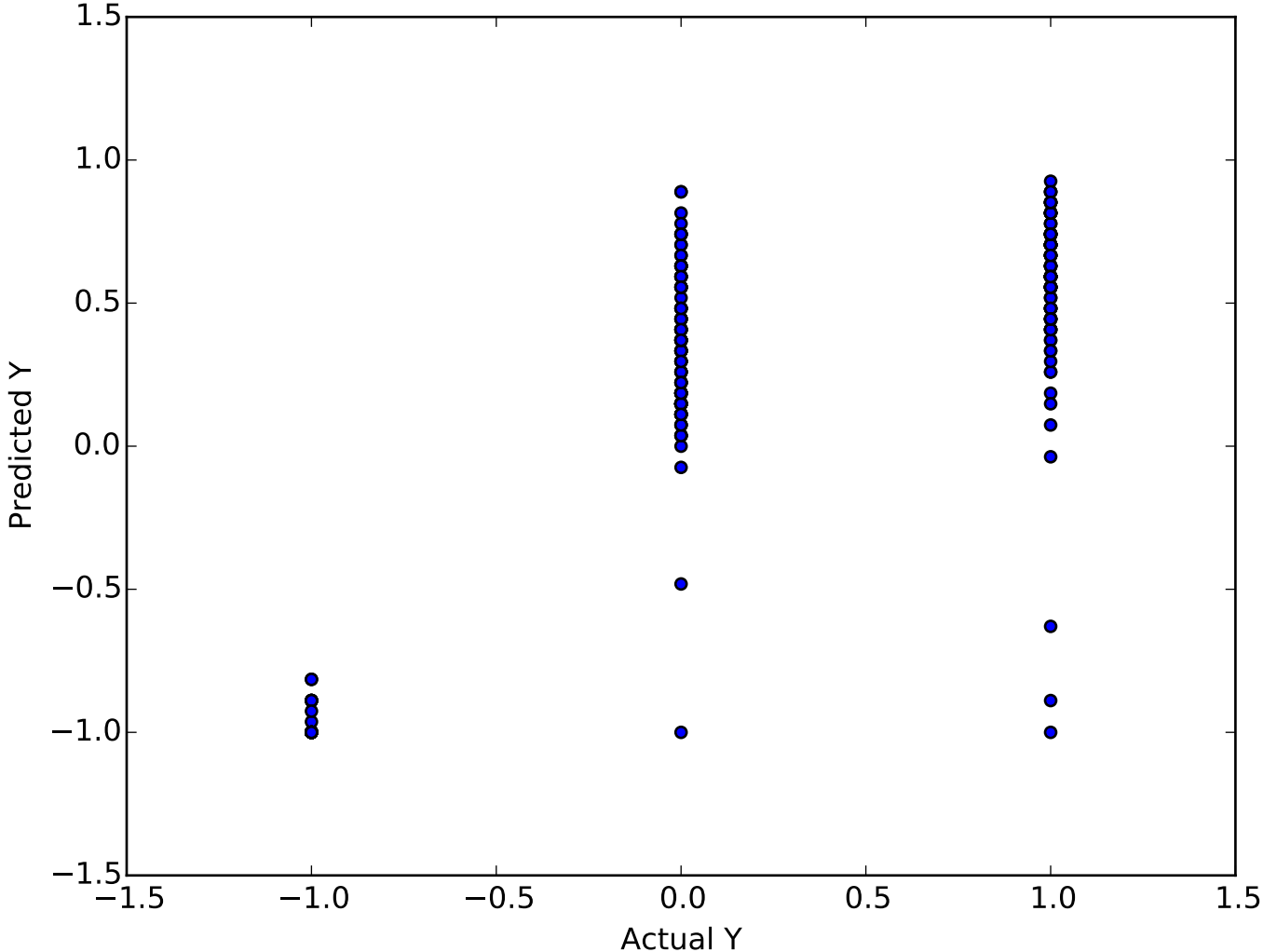
These K represent the lowest error(RMS) in the test data set. Another metric could have been for the highest correlation. We have considered the training set(out of sample) error only, as that is simulating the real world data in this case.

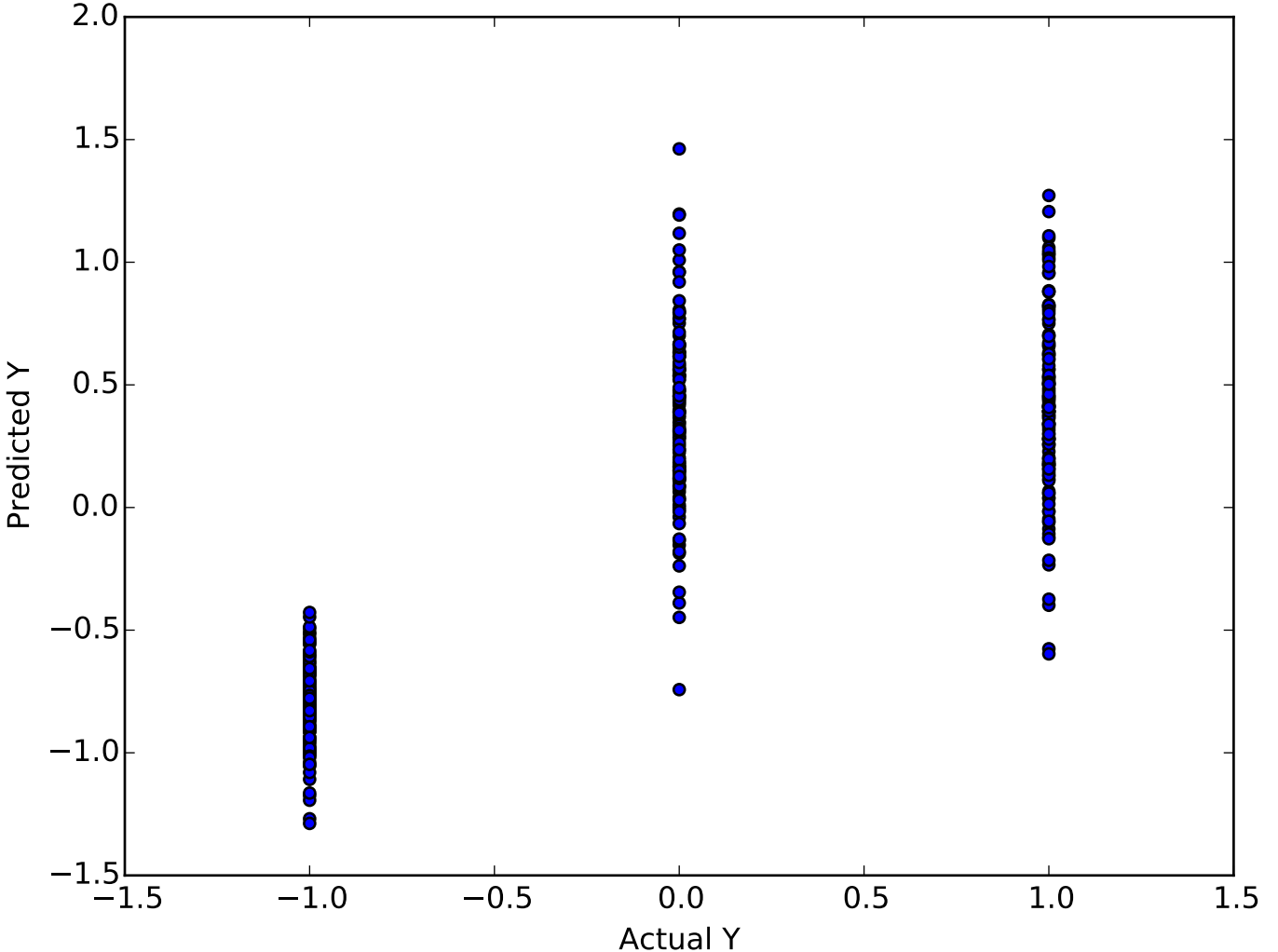
As K decreases, does overfitting occur for the datasets? At approximately which K does it start? Explain why you think this is occurring (or that it is not occurring).
Yes, over fitting does occur. It occurs for the Ks mentioned in the previous question. This is occurring because at lower K values, the learner is not generalized. An extreme example of this is the K=1 value. At that value, the training data is the value of the learner. This fits the training data perfectly, as the learner value is the the training data itself. But it doesn't generalize, as we can see with larger values of RMS errors initially. A larger value of K however over generalizes, taking into account many more neighboring values than necessary

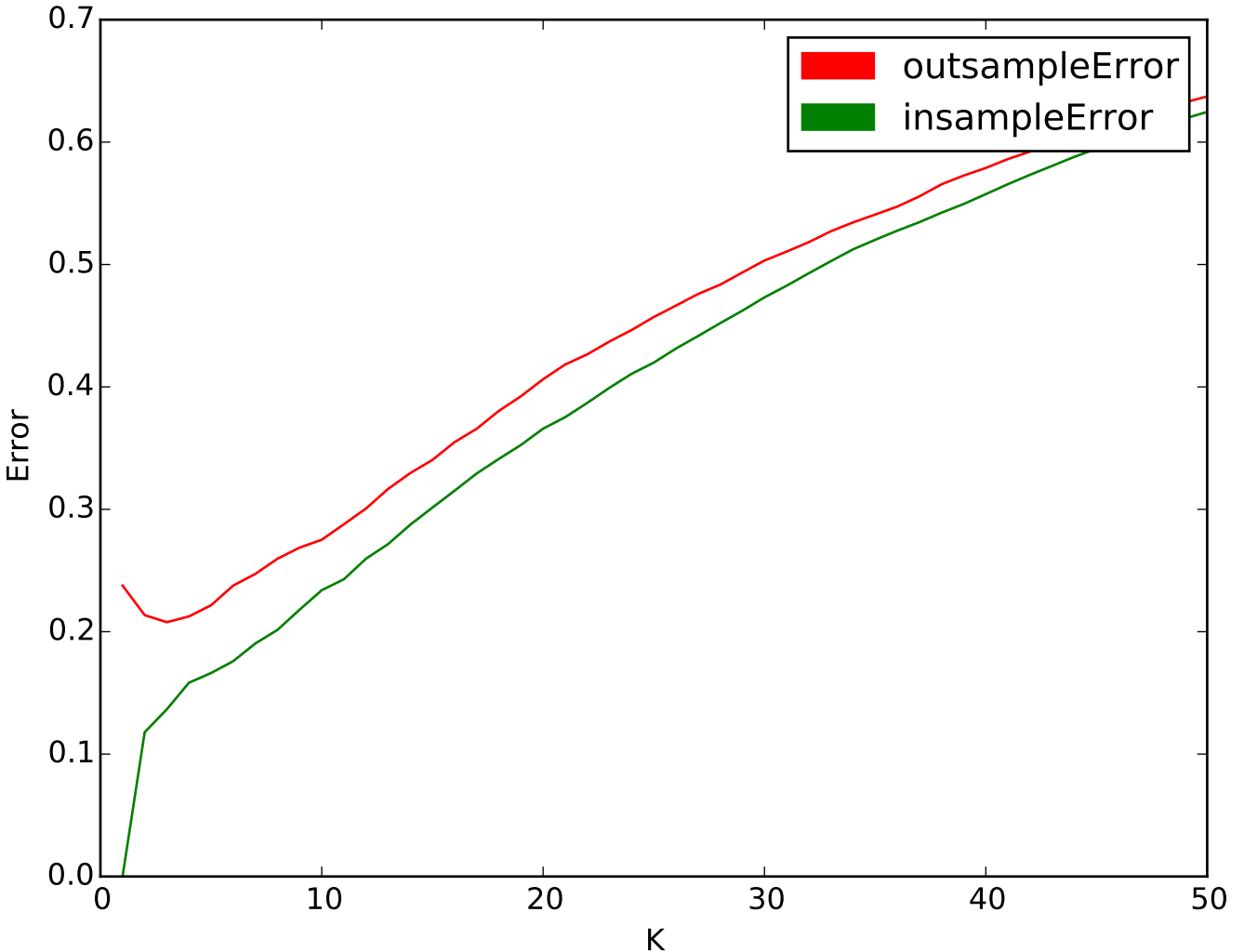
The figures are in the following order.

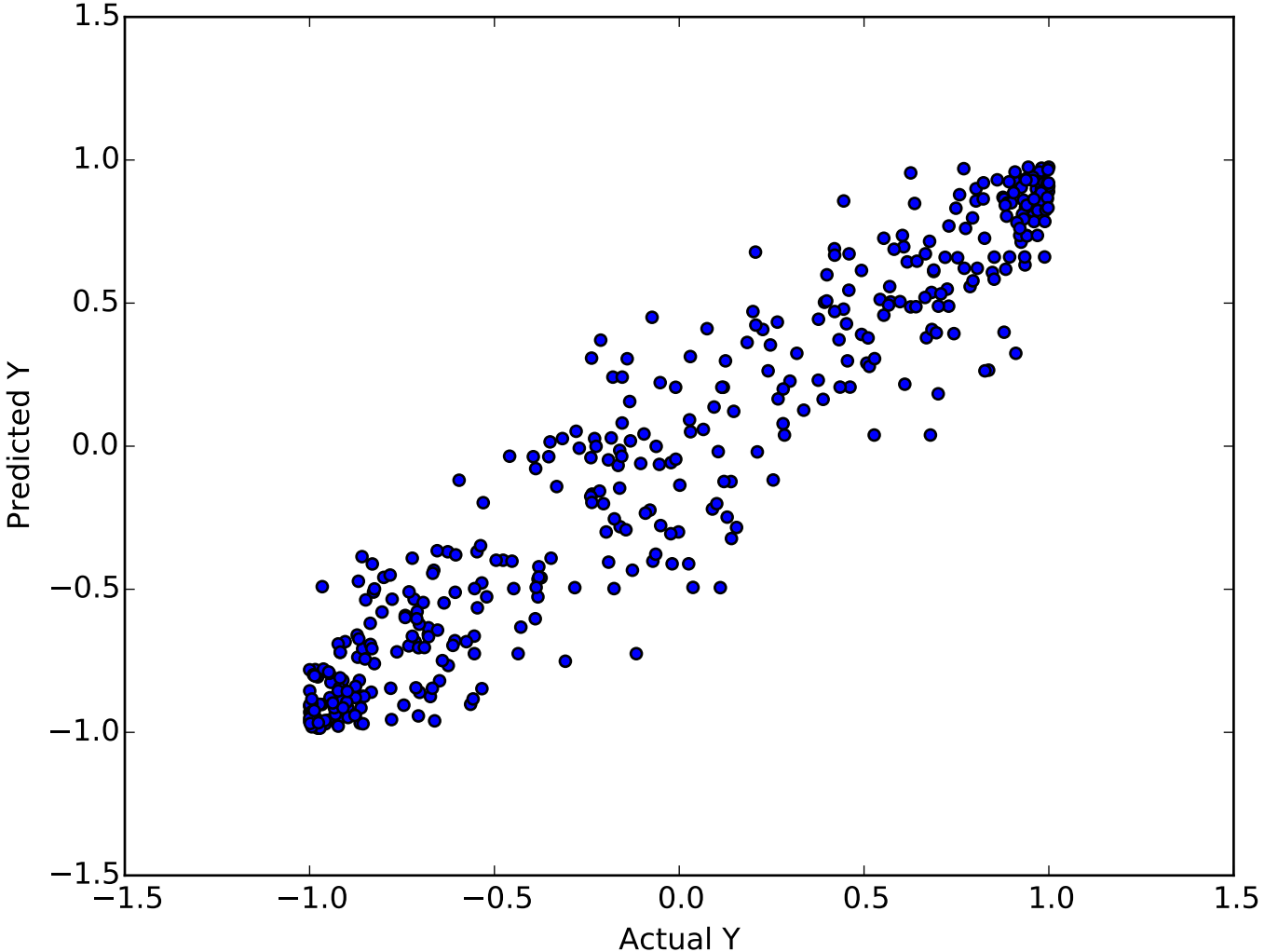
1. For data-classification-prob
    1. Error vs KNN
    2. Scatter plot for KNN optimal value(27)
    3. Scatter plot for LinRegLearner
2. For data-ripple-prob
    1. Error vs KNN
    2. Scatter plot for KNN optimal value(3)
    3. Scatter plot for LinRegLearner

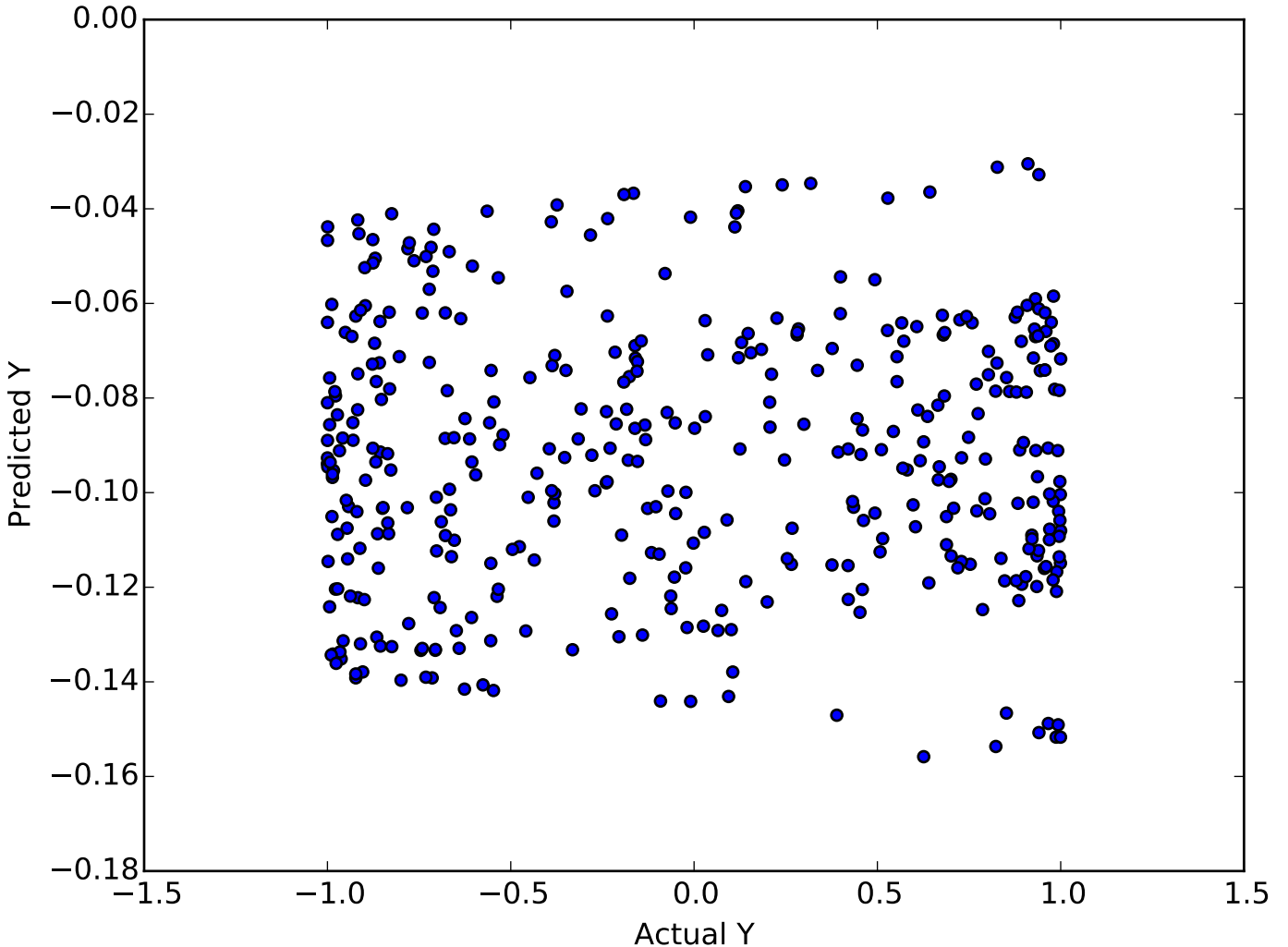

outsampleError
insampleError
Error
K
0.0
0.1
0.2
0.3
0.4
0.5
0
10
20
30
40
50


1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
Predicted Y
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Actual Y


2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
Predicted Y
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Actual Y


outsampleError
insampleError
Error
K
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
10
20
30
40
50


1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
Predicted Y
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Actual Y


0.00
−0.02
−0.04
−0.06
−0.08
−0.10
−0.12
−0.14
−0.16
−0.18
Predicted Y
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Actual Y