Problem Set 1

Due Sunday, April 16th, at 11:59 PM

CS - 171 Spring 2017

Nicholas Willhite SID: 861239087

Problem 1. During your regular medical check-up, your physician orders a regular blood test (that is, a test she orders for everyone having an annual check-up) to check for foobarinosis, a disease that was only discovered since your last check-up. This test has a false-positive rate of 1% (that is, if you dont have the disease, there is a 1/100 chance that the test will come back positive) and a false-negative rate of 0.2% (that is, if you do have the disease, there is a 2/1000 chance that the test will come back negative). The disease is present in 1 out of 4,000 people.

Your blood test comes back positive. What is the probability you have this disease? (Show your calculations)

Let A = Test positive, A' = Test negative
Let B = Has the disease, B' = Does not have the disease

Probability of testing positive given not having the disease

$$P(A \mid B') = .01$$

$$P(A' \mid B') = 1 - P(A \mid B') = 1 - .01 = \boxed{.99}$$

Probability of testing not positive given having the disease

$$P(A' \mid B) = .002$$

$$P(A \mid B) = 1 - P(A' \mid B) = 1 - .002 = \boxed{.998}$$

Probability disease is present

$$P(B) = .00025$$

$$P(B') = 1 - P(B) = 1 - .00025 = \boxed{.99975}$$

Looking for: probability of having the disease given testing positive

$$P(B \mid A) = \frac{P(A \mid B)\,P(B)}{P(A)} = \frac{P(A \mid B)\,P(B)}{P(A \mid B)P(B) + P(A \mid B')P(B')}$$

$$P(B \mid A) = \frac{(.998)\,(.00025)}{(.998)(.00025) + (.01)(.99975)}$$

$$P(B \mid A) = \boxed{.024158}$$

Problem 3. Run your function plotdata on the housetrain.data. What does this plot tell you about the data and prediction in this dataset. Look at the file housing.names for information about the features to help your interpretation of the plots.

The subplots are each feature plotted with the Y-value (Median value of owner-occupied homes in $1000's). This shows the correlation that this particular feature has on housing prices. Within figure 2: you see each subplot then used with LLS analysis. With this analysis we are given more of an insight on which features are more effective on the housing price. For example feature 6 (average number of rooms per dwelling) has more of an effect on housing price than say feature 2 (proportion of residential land zoned for lots over 25,000 sq.ft.) based on how close they are to the LLS line that is plotted.

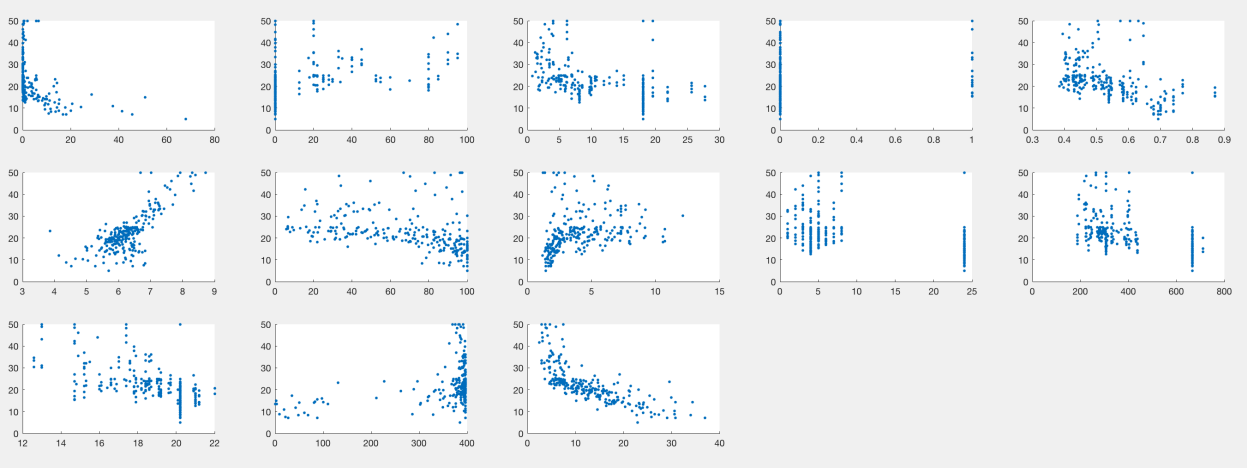

Figure 1: Subplots from housetrain.data

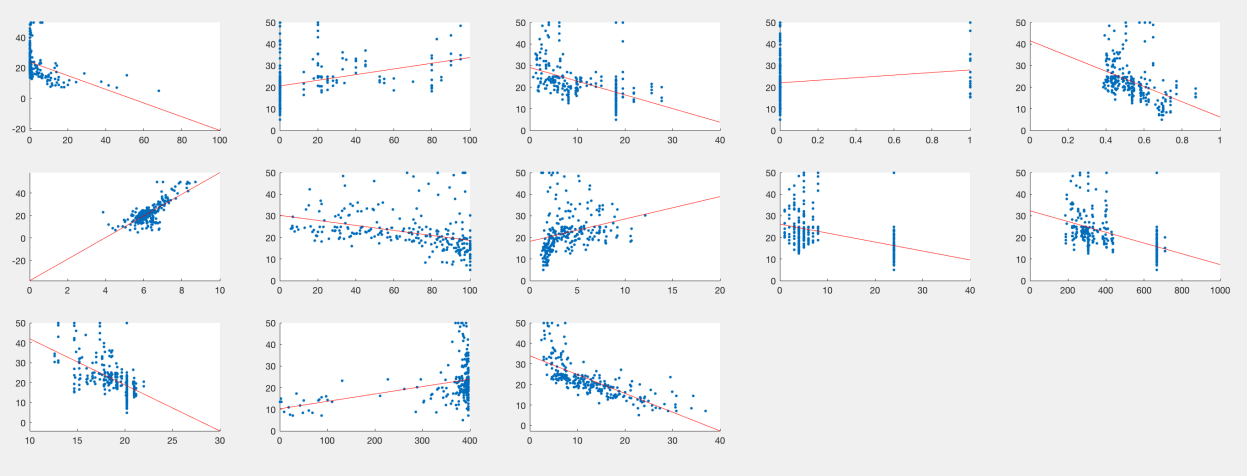

Figure 2: Subplots with LLS line from housetrain.data

Problem 4c. Run the supplied function plotacc that uses the ridgells and llserr functions you wrote above. What does this plot tell you about ridge regression? Read the code to understand what it is plotting. Think carefully about how you would use ridge regression and what the two curves represent.

The graphs for running plotacc shows us how accurate we are when predicting the housing prices given the two data sets. With the training set giving us minimum error compared to the testing set shows the we are not over fitting the data and still have a rather big gap of error in our predictions. If we were to introduce cross validation we may be able to shrink this gap and allow for a more accurate plot.

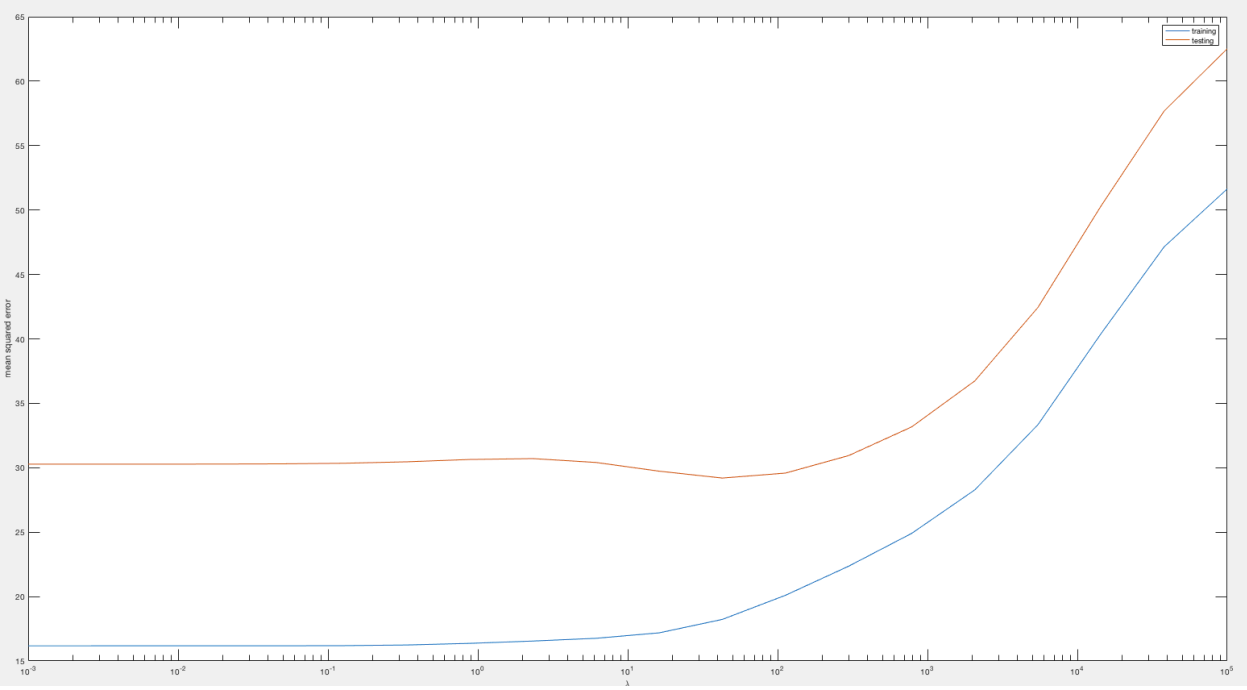

Figure 3: Ride Regression Graph