Cao, Marvin
SID: 861214117
4/13/17
CS-171: PS1

Cs-171 PS1

Problem 1:

Given: (1)False-Positive rate = 1% (if not diseased, 1/100 chance tests positive)
(2)False-Negative rate = 0.2% (if has disease, 2/1000 chance tests negative)
(3)Disease in 1/4000 people
(4)My test is positive.

Question: Probability I have this disease (foobarinosis?

Let us define:

D = has disease(foobarinosis)
Po = tests positive

(3) => 1. $P(D) = 1/4000$

(1) => 2. $P(Po|!D) = 1/100$

(2) => 3. $P(!Po|D) = 2/1000$

We are looking for the chances of having the disease given my circumstances P(D | Po).
Using the equations for marginal probabilities and conditional probabilities, we can get

$$P(Po) = P(Po, D) + P(Po, !D) = P(Po|D) * P(D) + P(Po|!D) * P(!D)$$

$$P(Po) = [1 - P(!Po|D)] * P(D) + P(Po|!D) * [1 - P(D)]$$

$$P(Po) = (998/1000) * (1/4000) + (1/100) * (3999/4000)$$

$$P(Po) = 998/4,000,000 + 3,999/400,000 = (998 + 39,990) / 4,000,000$$

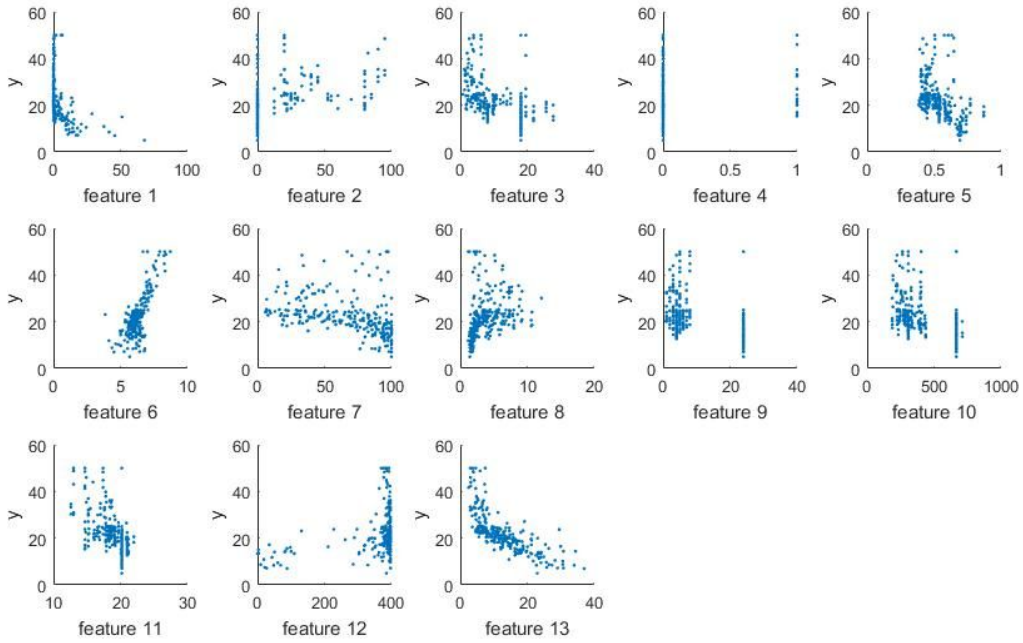
$$P(Po) = 40,988 / 4,000,000$$

Using the Bayes' Rule, we can find

$$P(D|Po) = P(Po|D) * P(D) / P(Po) = [1 - P(!Po|D)] * P(D) / P(Po)$$
$$= (998/1000) * (1/4000) / (40,988/4,000,000)$$
$$= 998 / 40,988 = 0.02435$$ or **2.43%** chance that I have the disease

Problem 3:

Cao, Marvin
SID: 861214117
4/13/17
CS-171: PS1



The strength of a feature correlation, positive or inverse, would improve our predictive power over this dataset, so we will examine each of the features in relation to the y value and determine their correlation by their scatter plots.

Feature 1: The first feature is the per capita crime rate. The crime rate seems to have a slight inverse correlation in the bottom half of the graph and no correlation in the top half of the graph. As the per capita crime rate increases, the value of y decreases.

Feature 2: The proportion of residential land used for lots of over 25,000 square-feet does not seem to have much of a correlation that could help in our predictions in this dataset.

Feature 3: The proportion of non-retail business acres per town has a slight inverse correlation made by the majority of points if we were to consider the three points in the top right as outlier.

Feature 4: Binary Charles dummy variable has no correlation to the value of y since it is binary and lines both the right and left edges of the graph giving us no helpful predictions to draw from.

Feature 5: The nitric oxide concentrations seems to have a slight inverse correlation on the bottom half of the graph whereas, the top half has no correlation.

Feature 6: The average number of rooms per dwelling seems to have a relatively high positive correlation with the value of y that will likely contribute the most to the prediction of this dataset.

Feature 7: The proportion of owner-occupied units built prior to 1940 has a inverse correlation that is weak considering the large amount of data that lies outside that trend.

Feature 8: The weighted distance to five Boston employment centers has a slight positive correlation that is weaker for higher y-values.

Feature 9: There seems to be no predictive trend as the scatter plot consists of two separate columns of data for high and low indexes of accessibility to radial highways.

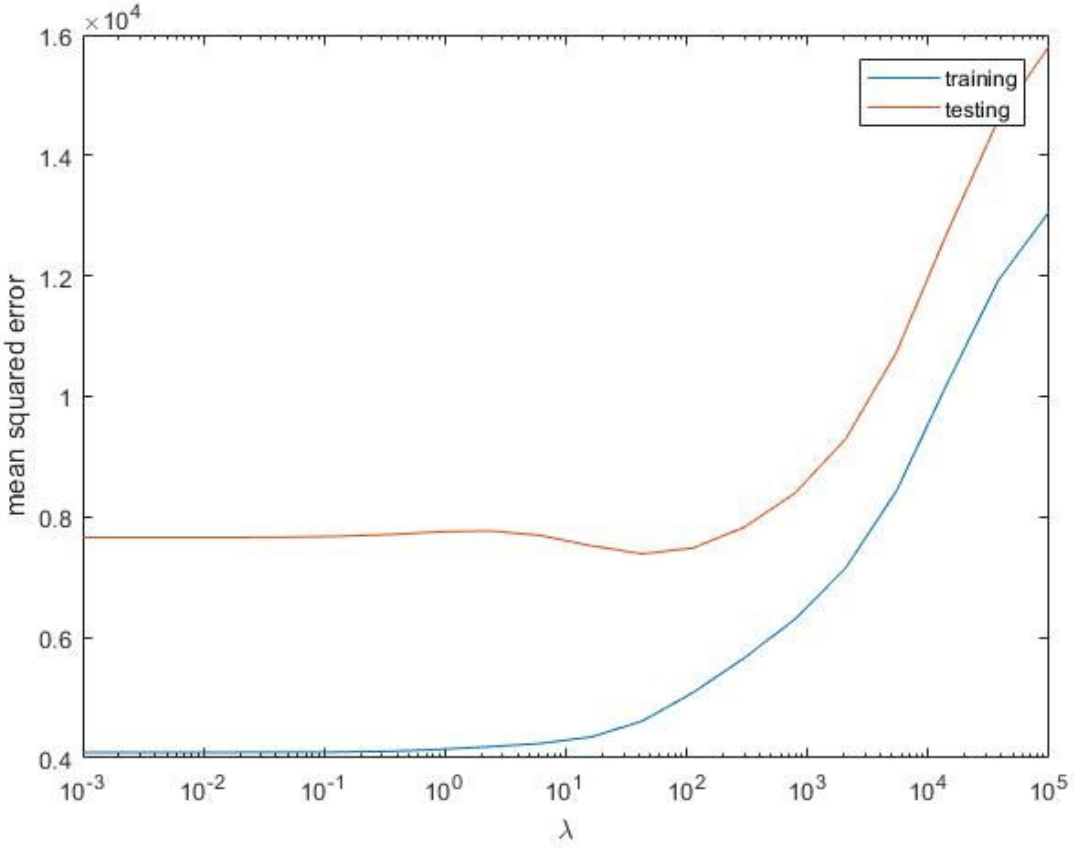
Feature 10: The full-value property tax rate per $10,000 seems to have a similar problem as feature 9 where there are two columns of data that do not provide trends to predict the value of y.

Feature 11: The pupil-teacher ratio has a moderate inverse correlation with the value of y and may be helpful for prediction.

Feature 12: This feature an equation involving the proportion of blacks by town which has a fairly weak correlation considering how only a few features are along that correlation and how there are many instances that are along the right edge of the graph.

Feature 13: The percent lower status of the population has a moderate inverse correlation to the value of y.

Problem 4c:



The graph display the increase in error as we increase the value of lambda along a logarithmic scale. At some point, the training and testing mean squared error increases significantly. Before then, we have a fairly stable mean squared error for lower lambda's in which the testing data has an expected larger mean squared error. This is expected since the values of the weights are built from the training data, thus resulting in a smaller error. A slight drop in error for the testing dataset occurs around the point where the mean squared error of

the training dataset begins to increase. The lowest mean squared error of the testing data, the drop in error mentioned, is the lambda I would use to predict the value of an unknown house since it is best lambda for untrained data which is what the unknown house would fall under.

The curve of the training data is how well the weights would predict the data it was created from which creates an expected upward curve as the value of lambda increases limiting larger weights that could be more specific to that training data and not actual real world data. The curve of the testing data is how well the weights predict a dataset that is not used for training but is still part of real world data. Thus this curve better displays the predictive power of ridge regression with different values for lambda.