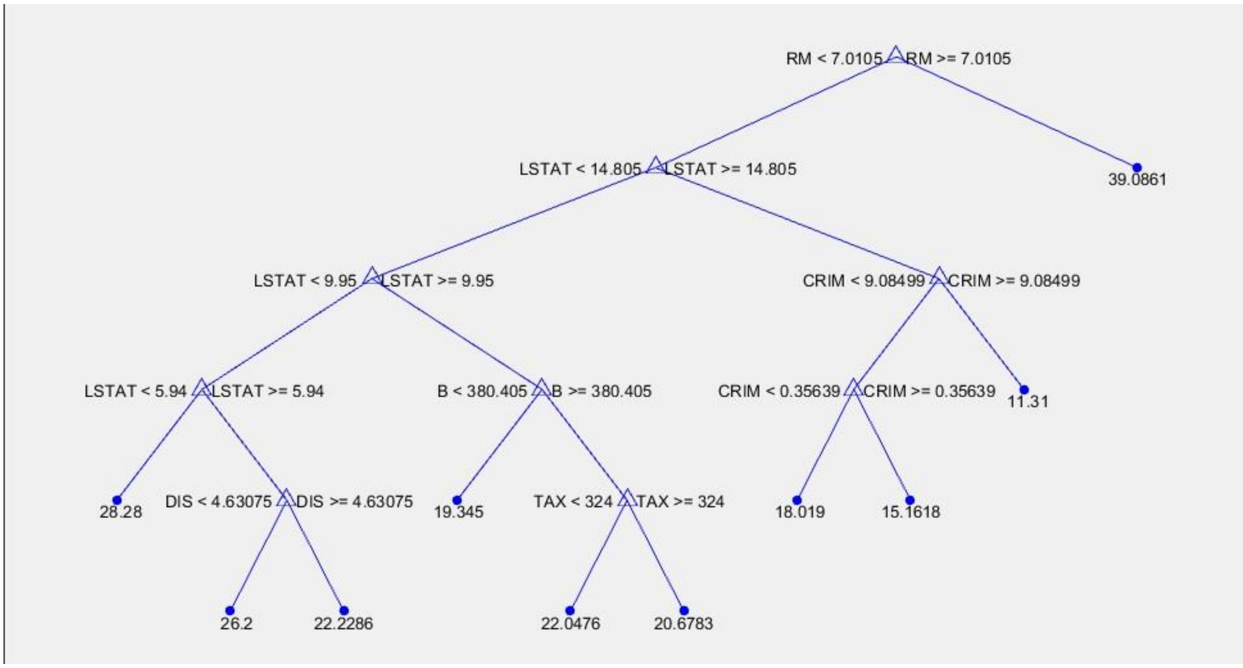
EC503: Learning From Data
Matlab Assignment 3
April 14, 2016

Athar Roshandelpoor
U46981822

Problem1

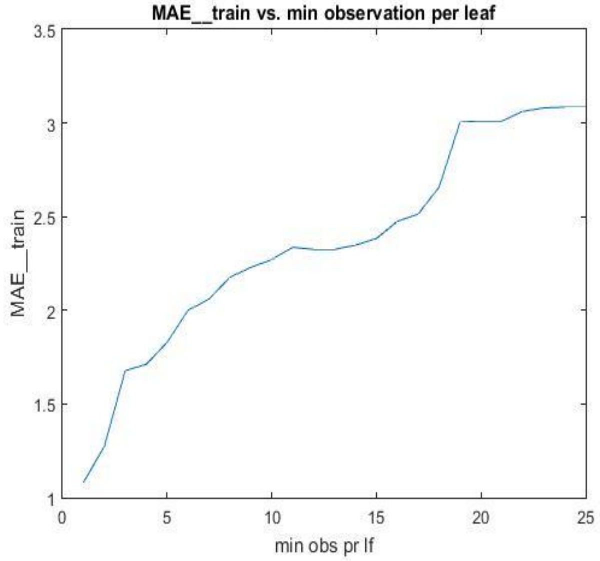
a)



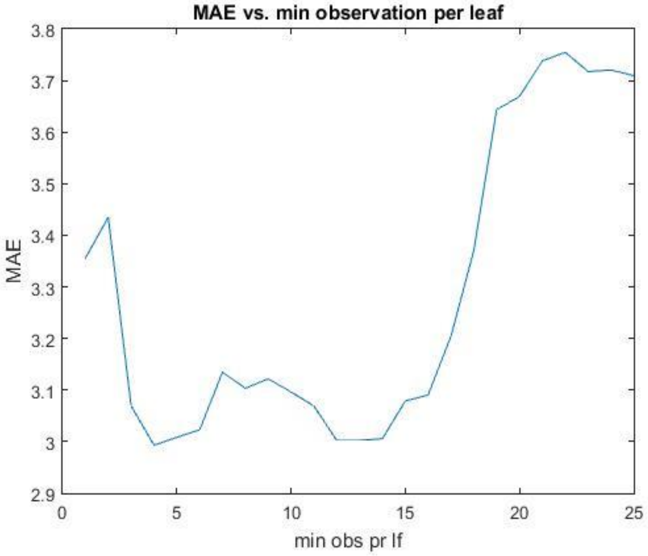
b) The estimated MEDV value for the test feature vector: 22.0476

c)

For train data set



For test data set



For training data: as we can see, by decreasing the number of min observation per leaf we get less value in mean absolute error. This makes sense because when we make leaf smaller and smaller, for training data we get the exact prediction. By increasing the min observation in each leaf we will have more points in each sub region so diversity increases, as a result, error increases.

For test data: as it is shown in the picture, when the number of min observation per leaf is very large, MAE is also large, which is logical. Because, it means that we did not partition or data enough to get the accurate answer for prediction.

However, when the number of min observation per leaf is very small, again we get a large error. The reason is that, when we divide our training data to very small sub region, actually, we don't learn anything we just memorizing data (overfitting problem). That reason make the error large when we have a small number for min observation per leaf.

EC503: Learning From Data
Matlab Assignment 3
April 14, 2016

Athar Roshandelpoor
U46981822

Problem 2

a) i) Yes it is unique. Because it has only one variant (x) and matrix sigma(x) is invertible. So the optimization problem has unique solution.

ii) w_{OLS}= 0.0221, b_{OLS} = 2.9743 ,MSE= 0.2588 ,MAE= 0.3174

b) i)

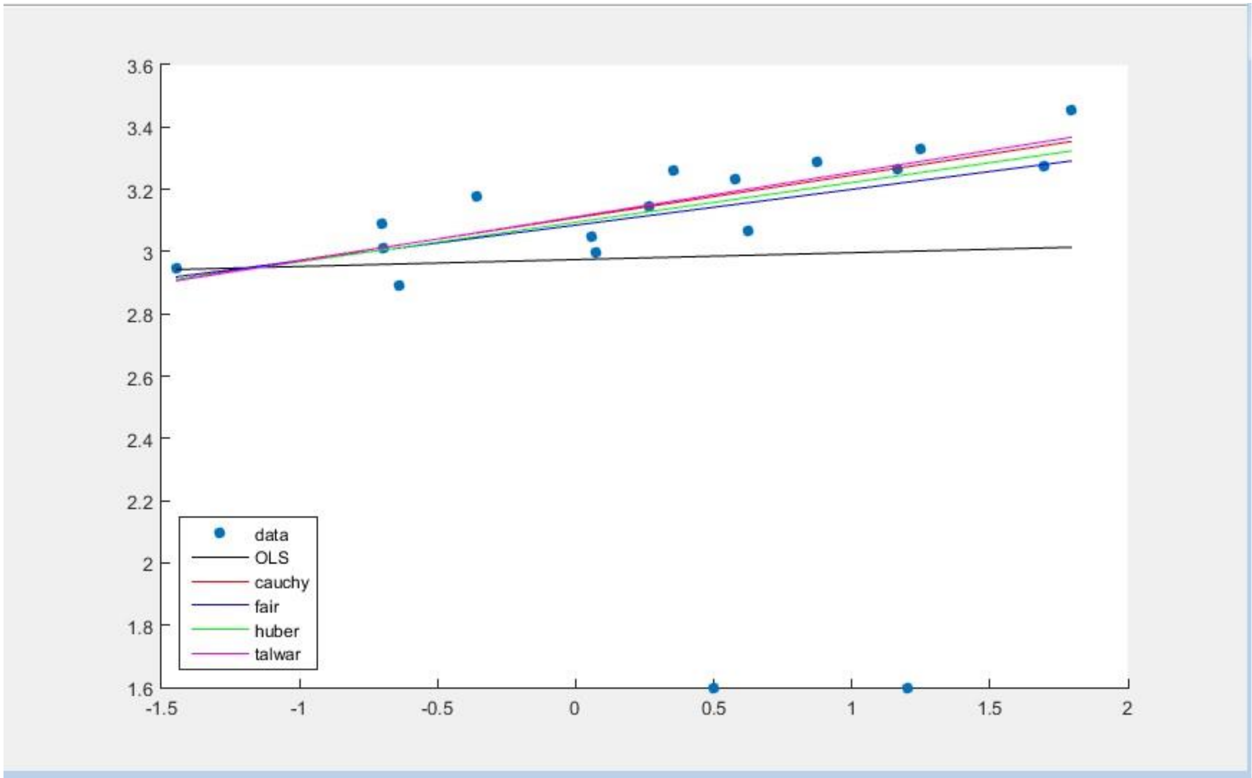
Method	Mean Absolute Value	Mean Squared Error
OLS	0.2588	0.3174
Cauchy	0.2427	0.2995
fair	0.2468	0.2861
Huber	0.2451	0.2922
Talwar	0.2435	0.3024

ii) w_{huber}= 0.1275 ,b_{huber}=3.0944

EC503: Learning From Data
Matlab Assignment 3
April 14, 2016

Athar Roshandelpoor
U46981822

iii)



As it is shown in the picture, we have two outliers ([0.5, 1.6] ,[1.2, 1.6]). These outliers have some effects on OLS and pull the OLS line down. Because OLS is not robust enough compare to robust linear regression line. On the other hand, robust fit is more robust and as a result, it follows the trend of the majority of the data points. In the sense of robustness is in the order: talwar> Cauchy> Huber> fair

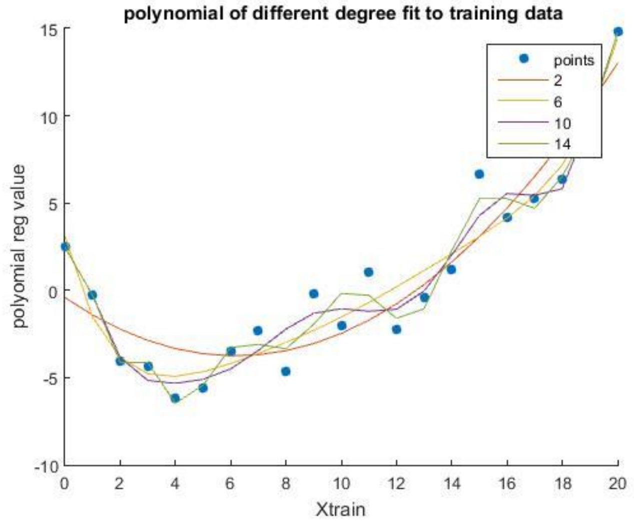
EC503: Learning From Data
Matlab Assignment 3
April 14, 2016

Athar Roshandelpoor
U46981822

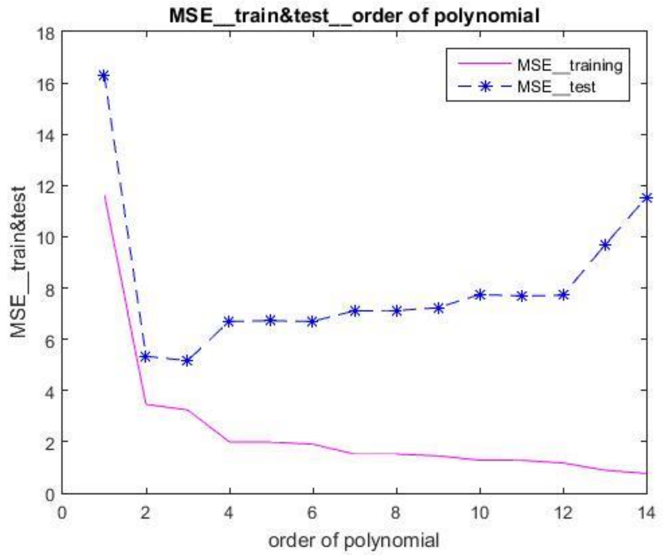
Problem3

a)

i)



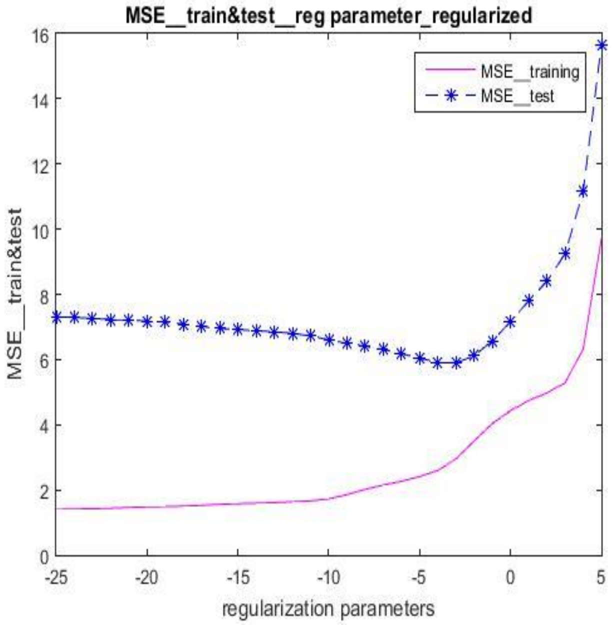
ii)



At first by increasing degree of the polynomial, MSE both for training and test data will increases. For **training data** by increasing the degree we pushes the line through the training data so always MSE decreases by increasing the degree od polynomial. However, for **test data**, at some point MSE starts increasing again, this is because of the problem of overfitting. We only have data so by increasing degree of polynomial we pushes the polynomial to the training data, in fact, we are not learning anything we just memorize the training data, so the accuracy will decrease.

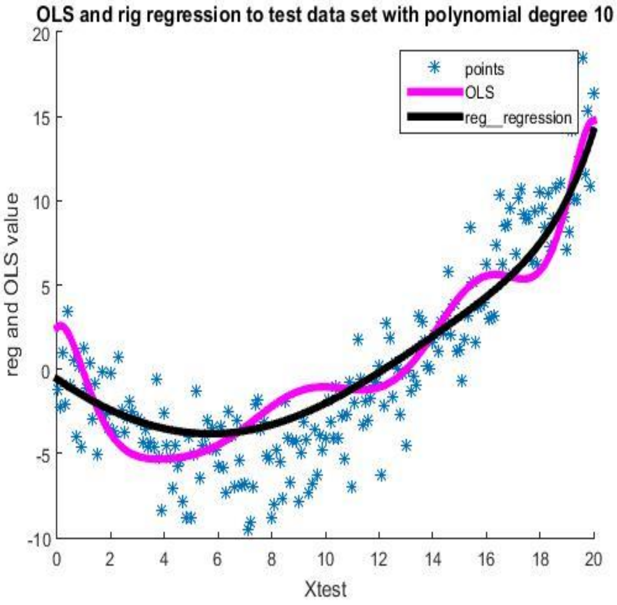
b)

i)



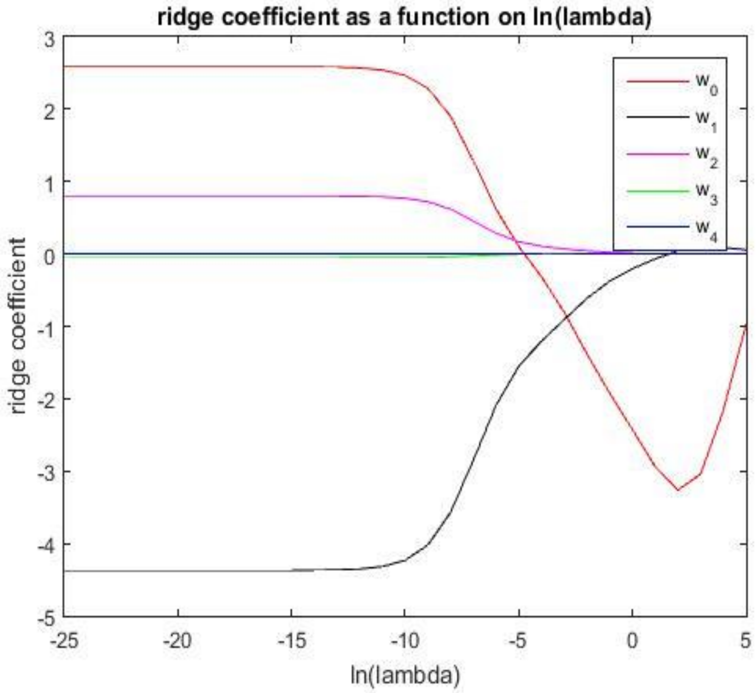
We can see that as the regularization parameter gets larger, MSE for training data increases and MSE for test data first decreases and then increases. For training data, since the regularization parameter forces to fit a lower degree polynomial to data, by increasing the parameter we do to go through all data so MSE will increase. For test data we have lower MSE when we find a good enough model for our data (at this point we are kind of avoiding the overfitting problem). However, by increasing the parameter to a very large number we miss a lot of data from training data so again MSE will increases.

ii)



As it is shown, ridge-regression gives a smoother fit compare to OLS. Although both are in degree 10 but the coefficient of rig-regression fit for high degree terms is very small and near zero. So they don’t look like in the same degree. This is reasonable since rig- regression imposes a penalty for overfitting.

c)



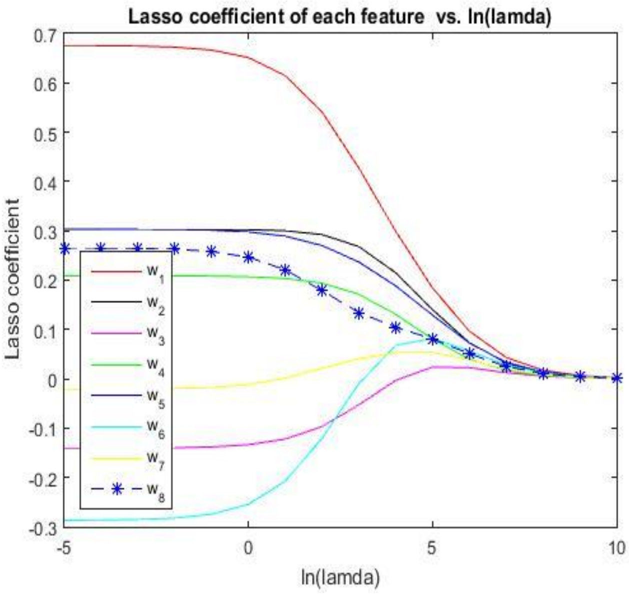
By increasing the lambda w coefficient are going to zero because of the penalty term which is become larger (l2 norm). So the minimum for our objective function achieves when W are zero. So by increasing the lambda all W go to zero. On the other hand, regularization parameter forces to fit a lower degree polynomial to data, as a result, here W_3 and W_4 are almost zero from the beginning by adding the regularization.

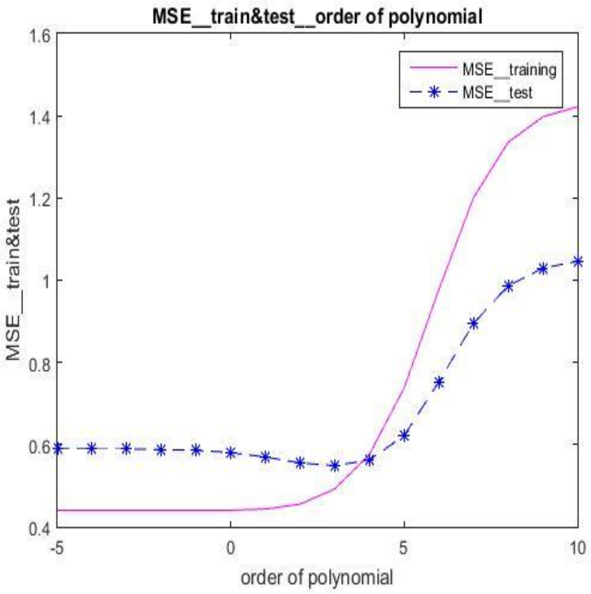
EC503: Learning From Data
Matlab Assignment 3
April 14, 2016

Athar Roshandelpoor
U46981822

Problem4

a)





b) As it is shown in the picture, Lasso coefficients converge to zero when lambda increases. This is because when we have a very large penalty term in our objective function the first term (which has w and it is influenced by data) is negligible and the penalty term is dominant. Since the penalty term in always nonnegative the minimum value for that is zero so we get all the w zero.

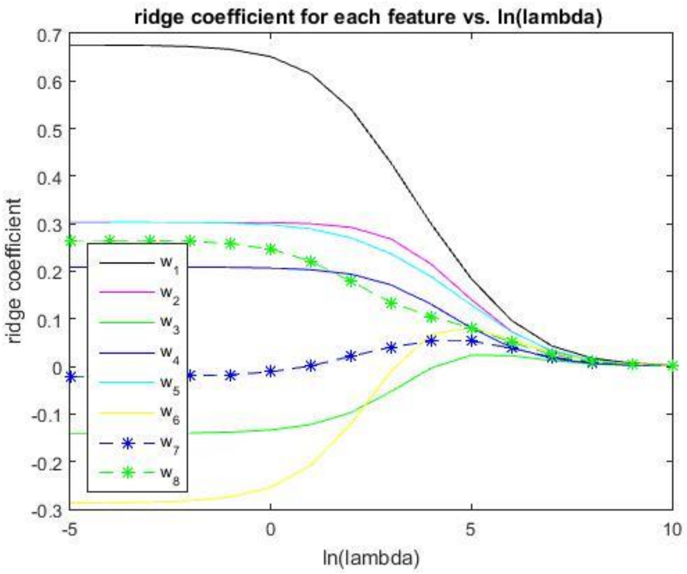
We can see that the first two W are converging last. Which means that the 2 most meaningful terms (dominant features) are lcavol - log (cancer volume), lweight -log (prostate weight). We can thus discard other predictors and only use lcavol and lweight to fit a regression model to predict lpsa (log (prostate specific anti-gen)). It means that for lpsa, among 8 features these two are the most dominant ones.

Also for Lasso coefficients, from beginning we have some almost zero coefficients and some coefficients which are not close to zero. It is a sparse matrix, which means that some of W s are zero or almost zero and coefficients do not behave in same way. But eventually, by increasing lambda, all of them go to zero.

EC503: Learning From Data
Matlab Assignment 3
April 14, 2016

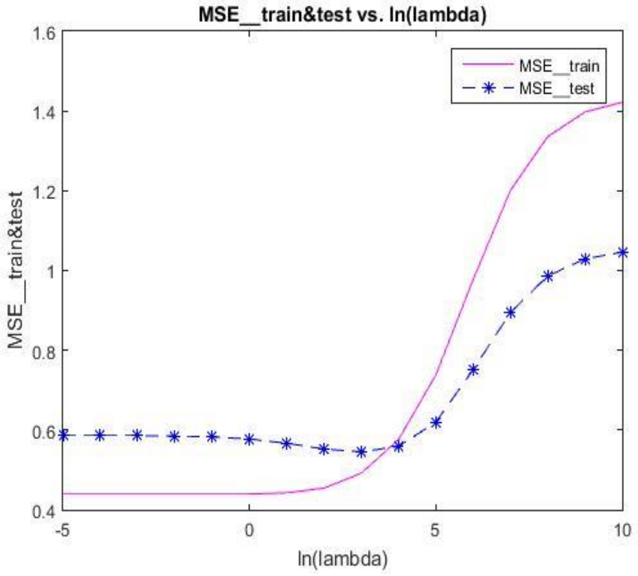
Athar Roshandelpoor
U46981822

c)



As we can see by increasing the lambda, ridge coefficient are again converges to zero. But here is much slower than Lasso coefficients. As we can see here not even exactly converging to zero (although some number really close to zero). Here all the coefficient are converging to zero approximately at same time and it is not like lasso that some w converges later than others. This is because ridge regression imposes a l2 norm penalty on objective function while lasso regression imposes a l1 norm penalty. The comparison shows that lasso regression does yield sparse estimates of the parameters.

d)



Same argument as 3-b prat i.