CSL407 Machine Learning

HW1

Jaskaran Singh (2011cs1012)

Ans1

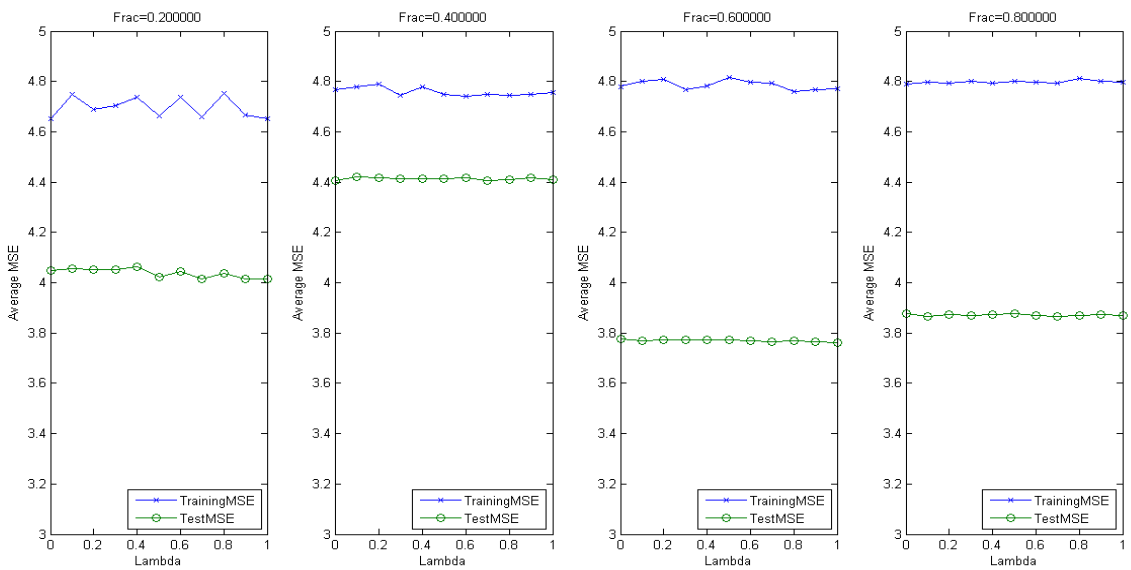
Fractions for partition into training/test set 0.25, 0.50, and 0.75

Lambda values taken 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1.

6) Total of 11 parameters (including the intercept term W_0)

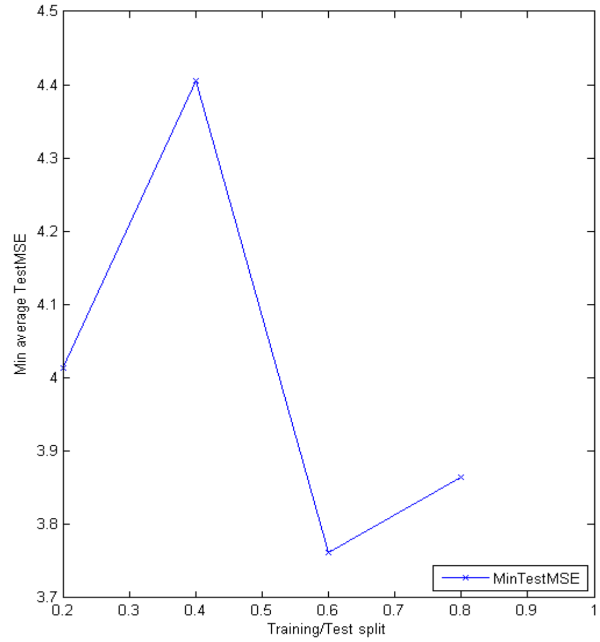
For a fixed value of lambda, the maximum weight of the regression model is the one corresponding to the intercept (W_0).The 2 least significant attributes are W_4(Length) and W_5(Diameter).Removing them doesn't cause much loss in information as the graphs formed are very similar. Thus, removing these extra parameters decreases the training time without much change to the results.

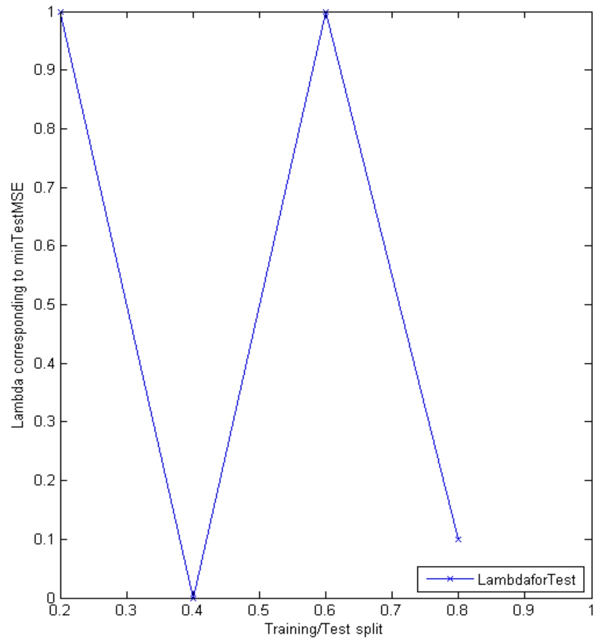
7) The following figure shows the effect of Training Fraction and Lambda on training MSE and test MSE.



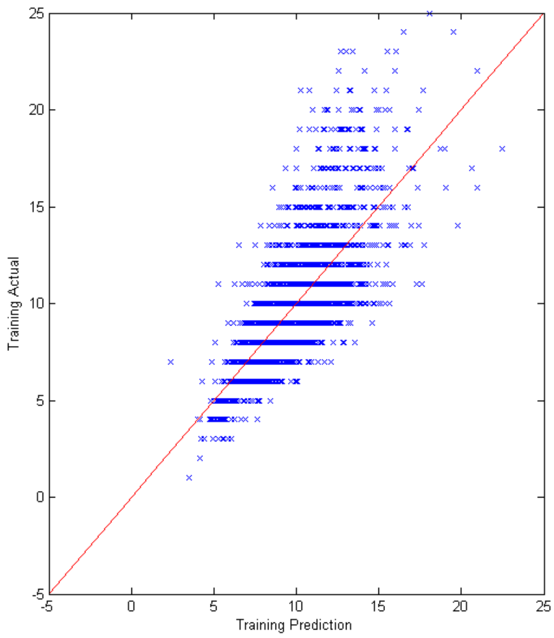
From the figure we see that the minimum Test MSE occurs in the case of training fraction=0.6 and lambda=1.0

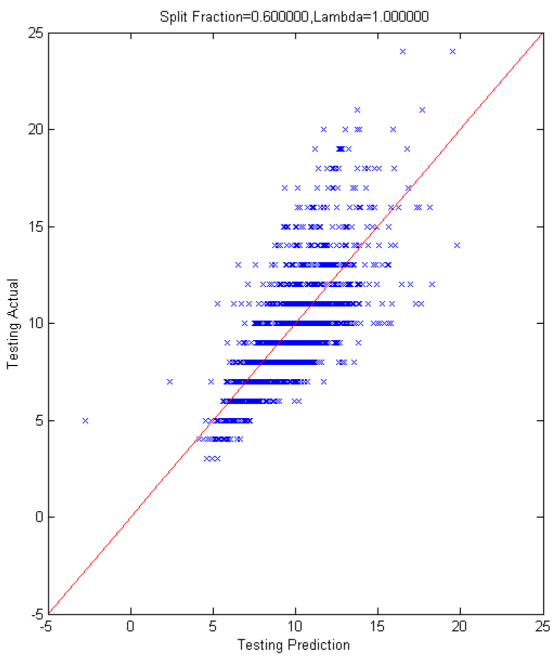
A good model is one which has the least value of Test MSE.





The above is a graph of Minimum average Test MSE for different Training fractions and the corresponding lambda values at which the minimum is achieved. Here, we can see clearly that at 0.6 fraction and lambda=1 we get minimum average Test MSE out of all cases.





The above figure shows the data points with their actual vs predicted labels. The red line has a slope of 45 degrees.

Ans2

a) The quartic regression RSS for training set would be lower than that of linear regression RSS for training set. It is because the quartic curve, being of a higher degree, would overfit the training set resulting in a lower RSS.

b) For the test set, since we know that the relationship between X and T is linear, the linear regression RSS would be lower than that of quartic regression RSS since the linear regression model generalises the problem well whereas the quartic regression model overfits the training set.

c) The quartic regression RSS for the training set would be lower than that of the linear regression RSS since the quartic curve has higher powers which make it flexible in fitting the training set.

d) It depends on how far the actual relationship is from the linear regression model and the quartic regression model. If the actual relationship is very close to linear, then the linear regression model would give a lower RSS as compared to a quartic model which would not generalize well. If the actual relationship is far from linear and is closer to quartic, then the quartic model would fit well resulting in a lower RSS as compared to the linear regression model.

Ans3

Consider the weight matrix(R) as a diagonal matrix where the R_{nn} entry is the weight of the nth observation t_n .The weighted sum of squares equation is:

$$1) E_D(W) = \frac{1}{2}\sum_{n=1}^{N} r_n\{t_n - W^T x_n\}^2$$

So, the derivative of the weighted sum of squares equation wrt $W_j$is

$$2) \frac{\partial E_D(W)}{\partial W_j} = \sum_{n=1}^{N} r_n\{t_n - W^T x_n\}x_n^{(j)}$$

Where W_j is the jth parameter to be learnt from training.

Writing the gradient in matrix form and equating it to zero for minimising $E_D(W)$,

$$3) \frac{\partial E_D(W)}{\partial W} = X^T R(XW - T) = 0$$

Solving it for W,

$$W = (X^T RX)^{-1} X^T RT$$