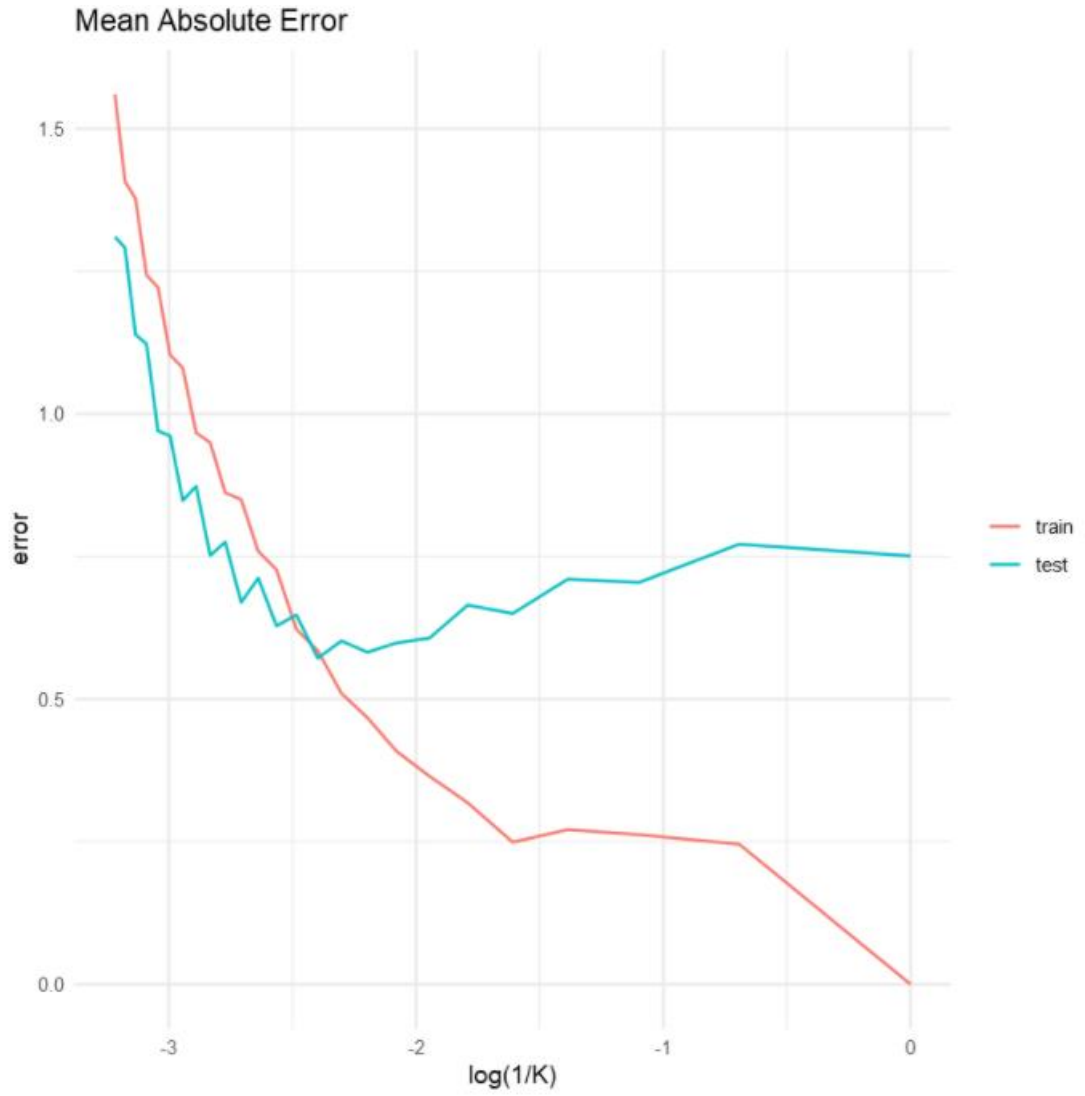
Question 1



Here, I've chosen the Mean Absolute Error(MAE) as the error function. MAE measures the average magnitude of errors in prediction irrespective of their direction.

Question 1.3

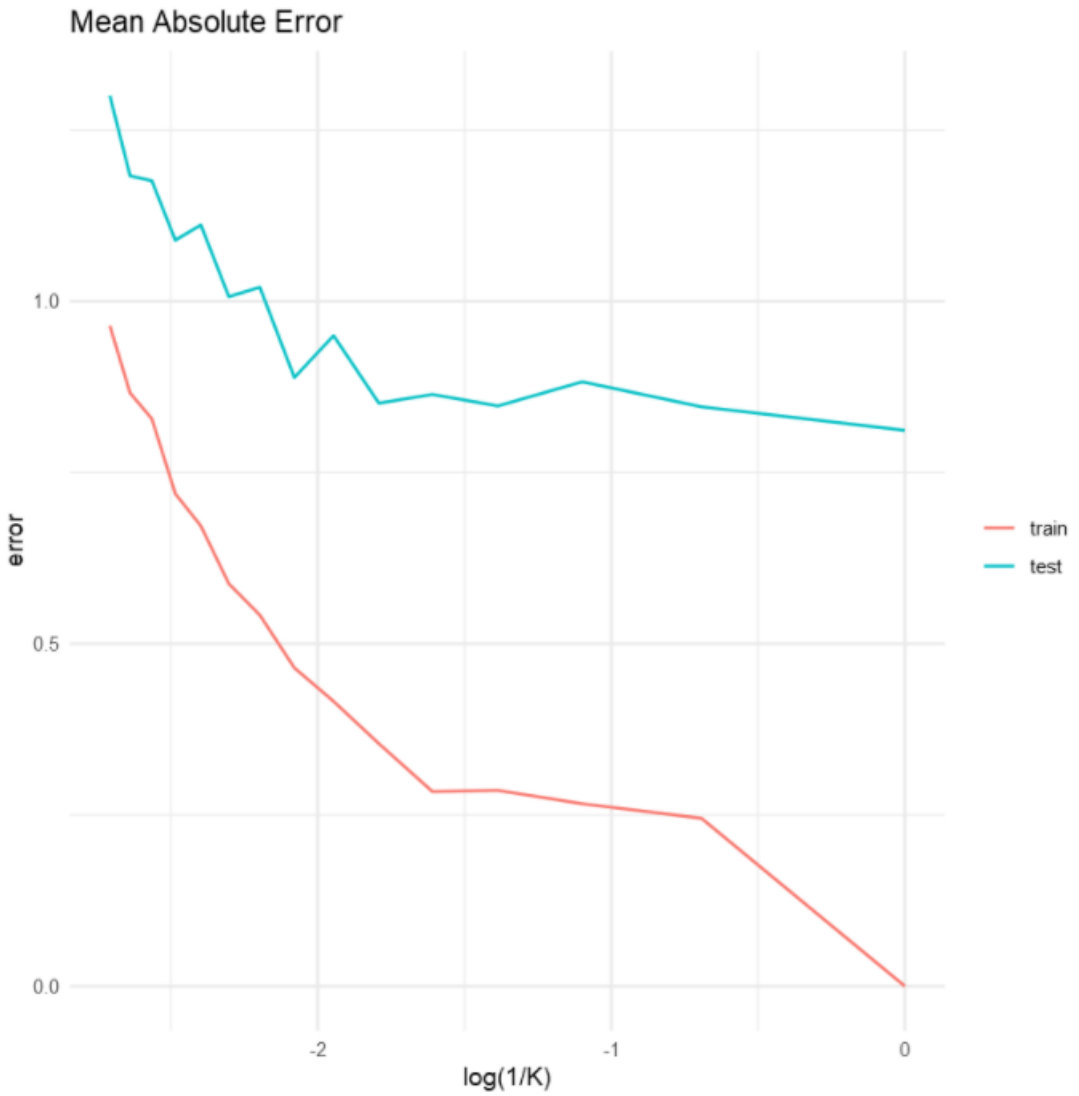
The optimum value of K in terms of testing error is 11 (using K=11 gives the lowest magnitude of testing error).

Increasing the model complexity i.e. decreasing K leads to overfitting as model memorizes the training data and the training error reduces to 0 whereas the testing error increases.

Whereas, decreasing the model complexity i.e increasing K leads to underfitting as the model neither fits well to training data and nor to the test data (model has high test and training error).

The right side of the graph log(1/K) = (-0.5, 0) shows overfitted models while the left side of the graph log(1/K) = (-3.5, -3) shows underfitted models.

Question 2



Here, I've chosen the Mean Absolute Error(MAE) as the error function. MAE measures the average magnitude of errors in prediction irrespective of their direction.

Question 2.3

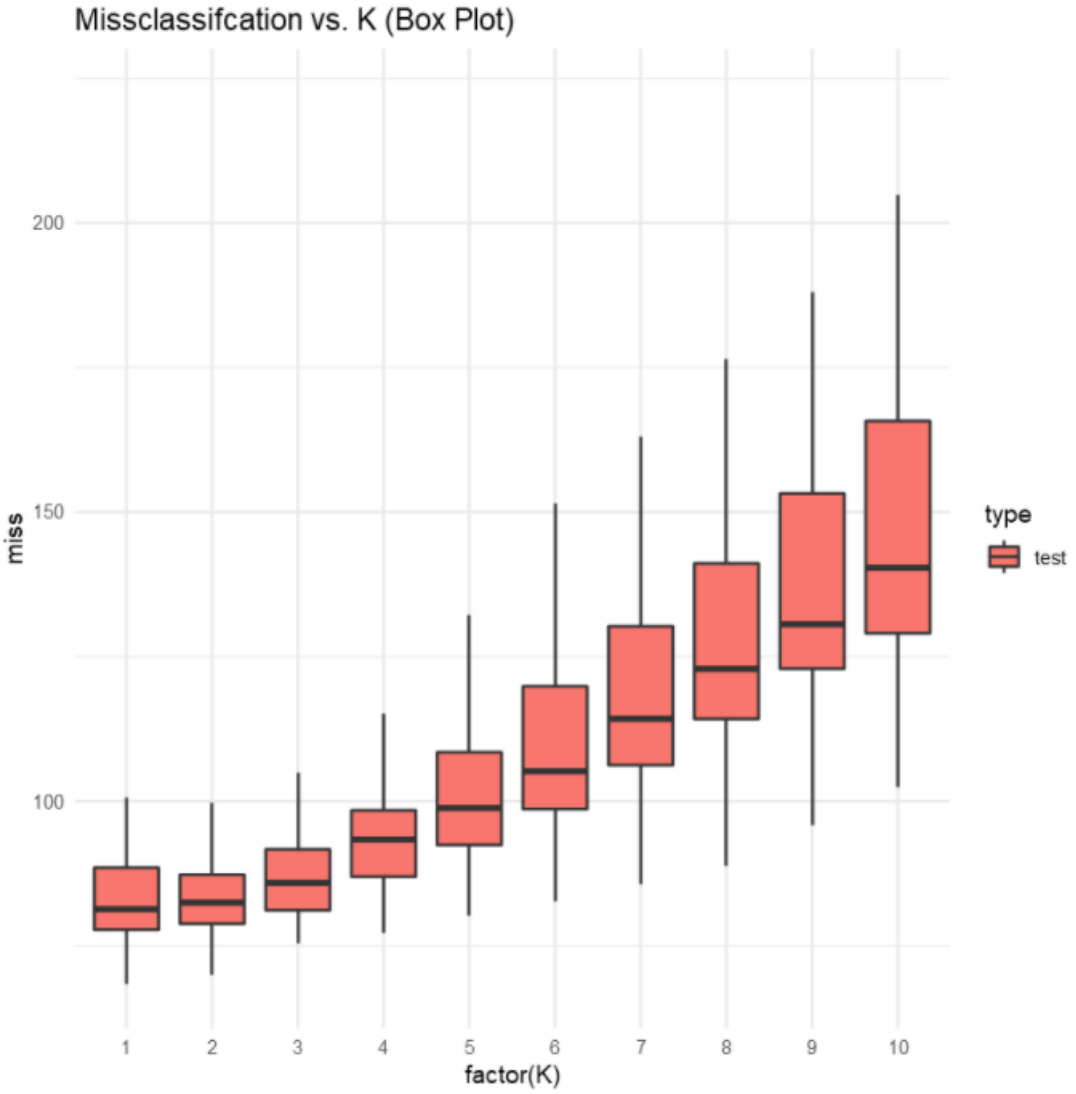
The optimum value of K in terms of testing error is 4. Here, I observe that the the testing error further falls on decreasing the value of k which is quite interesting

Increasing the model complexity i.e. decreasing K leads to overfitting as model memorizes the training data and the training error reduces to 0.

Whereas, decreasing the model complexity i.e increasing K leads to underfitting as the model neither fits well to training data and nor to the test data (model has high test and training error).

The right side of the graph log(1/K) = (-0.5, 0) shows overfitted models while the left side of the graph log(1/K) = (-3.0, -2.5) shows underfitted models.

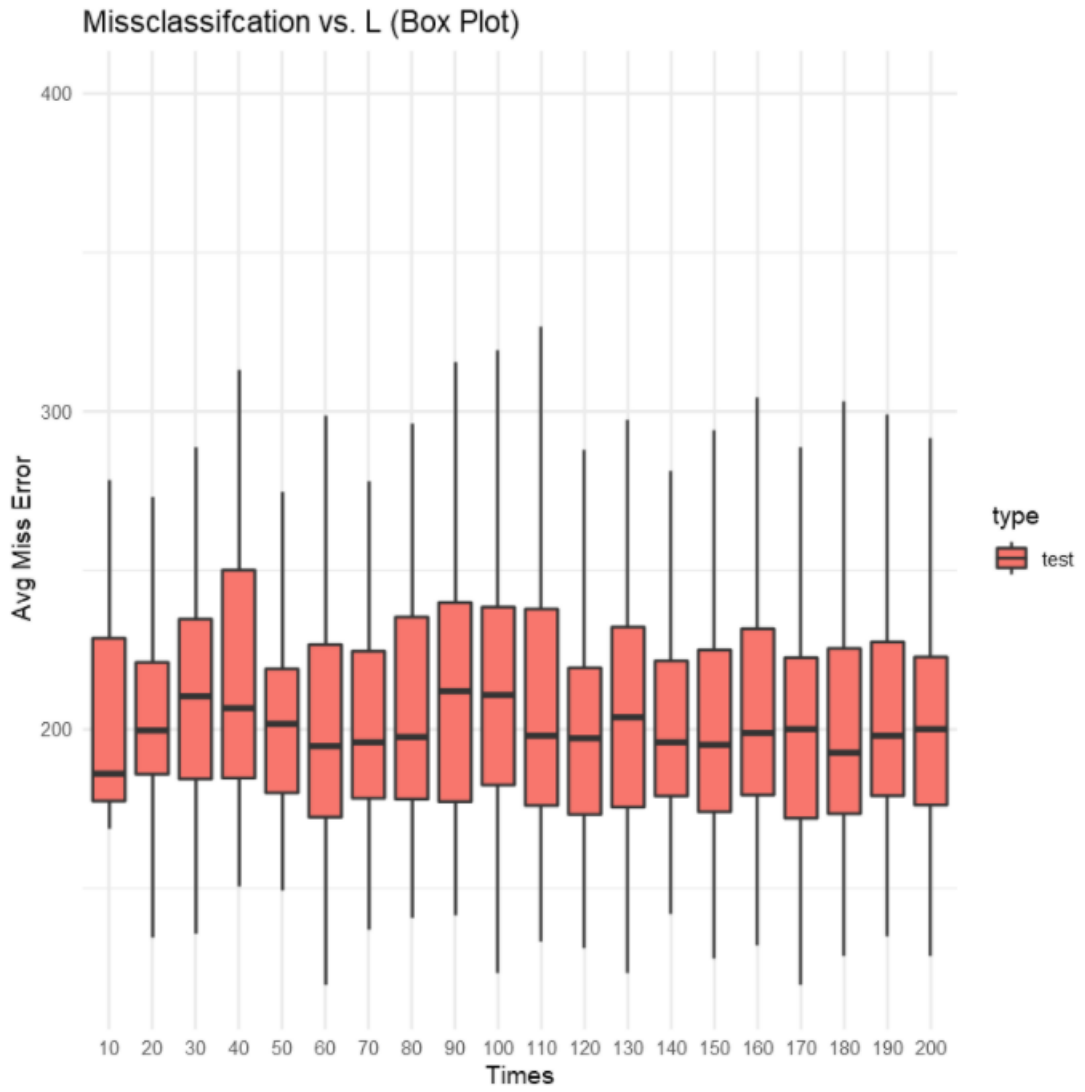
Question 3



Question 3.3

Here, we observe that the mean absolute error increases as K increases.

Moreover, for large K we can see that the size of the boxplot increases as K increases. This means that the uncertainty around the error increases as K increases i.e a higher K causes an increased variation in test errors.



Question 3.5

When plotting the average misclassification against the number of subsets in bootstrapping we observe that there isn't much variation in errors across the number of datasets drawn.

The range of avg. errors for the boxplots seem to be approximately equal while for L=20 and 40, the range of errors seem to be a bit low.

Question 4

Suppose we have one red, one blue, and one yellow box. In the red box we have 3 apples and 5 oranges, in the blue box we have 4 apples and 4 orange, and in the yellow box we have 1 apples and 1 orange. Now suppose we randomly selected one of the boxes and picked a fruit. If the picked fruit is an apple, what is the probability that it was picked from the yellow box? Note that the chances of picking the red, blue, and yellow boxes are 50%, 30%, and 20% respectively and the selection chance for any of the pieces from a box is equal for all the pieces in that box. Please show your work in your PDF report.

$$P(Red - box) = 0.5$$

$$P(Blue - box) = 0.3$$

$$P(Yellow - box) = 0.2$$

$$P(Apple|Red - box) = 3/8$$

$$P(Apple|Blue - box) = 4/8$$

$$P(Apple|Yellow - box) = 1/2$$

$$P(Yellow - box|Apple) = \frac{P(Apple|Yellow-box)P(Yellow-box)}{P(Apple)}$$

$$P(Yellow - box|Apple)$$

$$= \frac{P(Apple|Yellow-box)P(Yellow-box)}{P(Apple|Yellow-box)P(Yellow-box)+P(Apple|Blue-box)P(Blue-box)+P(Apple|Red-box)P(Red-box)}$$

$$P(Yellow - box|Apple) = \frac{0.5\times0.2}{0.5\times0.2+0.375\times0.5+0.5\times0.3}$$

$$P(Yellow - box|Apple) = \frac{8}{35}$$

$$P(Yellow - box|Apple) = 0.2285$$

Question 5

Question 5.1

Error function for regression:

$E_{reg} = \sum_{i=1}^{N} (t_i - \boldsymbol{w}.\phi(x_i))^2$

For L_2 regularization, penalty term:

$\lambda \sum_{m=0}^{M-1} \boldsymbol{w}_m^2$, where λ is the regularization parameter and M is the complexity of the linear model.

So now, the error function with L_2 regularization is:

$Error = \sum_{i=1}^{N} (t_n - \boldsymbol{w}.\phi(x_i))^2 + \lambda \sum_{m=0}^{M-1} \boldsymbol{w}_m^2$

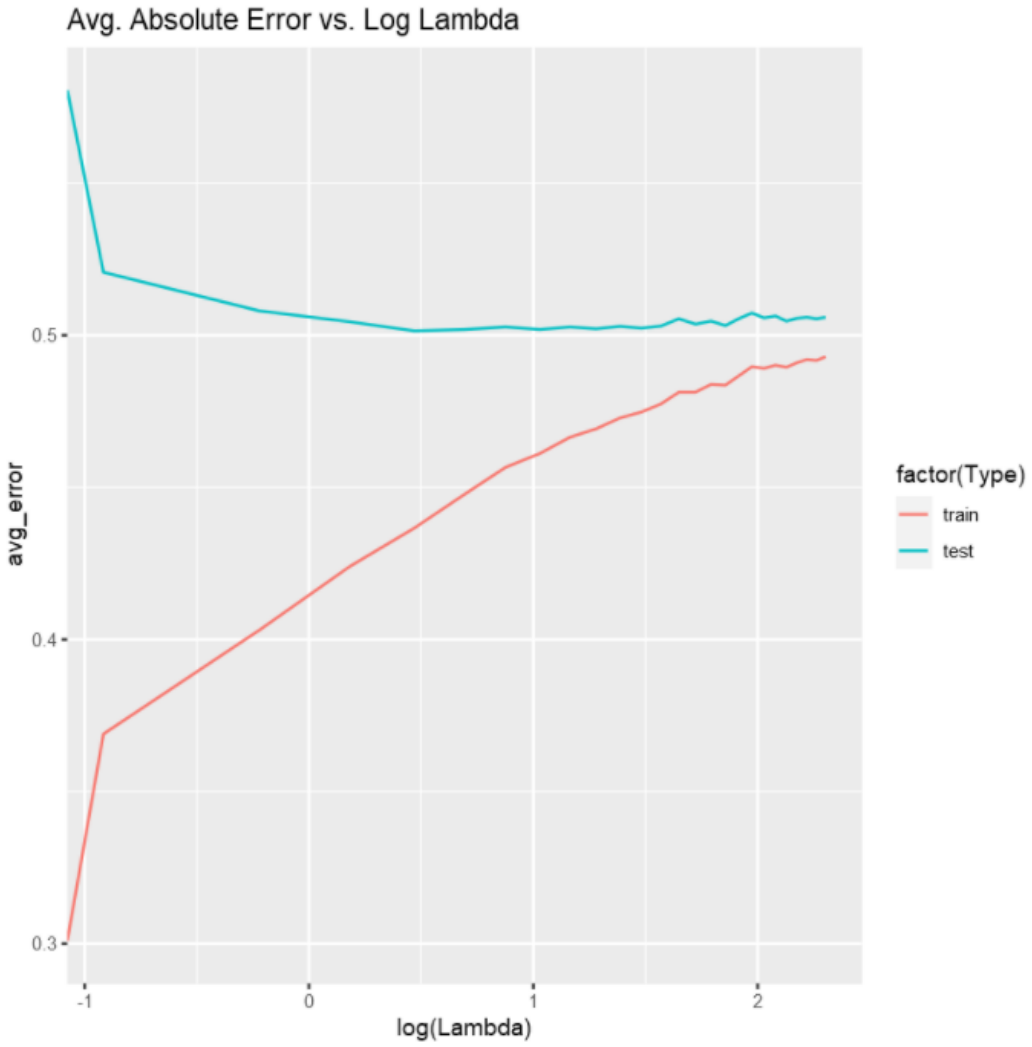
Differentiating $Error$ w.r.t. $\boldsymbol{w}$:

$\frac{\partial Error}{\partial \boldsymbol{w}} = -\sum_{i=1}^{N} \phi(x_i)(t_i - \boldsymbol{w}.\phi(x_i)) + 2\lambda \sum_{m=0}^{M-1} \boldsymbol{w}_m$

For each step we subtract the partially differentiated error with the slope value of last step. The current step is considered the solution if the step size is smaller than the threshold value. Updating weights:

$\boldsymbol{w}^{(\tau+1)} := \boldsymbol{w}^{(\tau)} - \eta \frac{\partial E(\boldsymbol{w}^{(\tau)})}{\partial \boldsymbol{w}^{(\tau)}}$

$\boldsymbol{w}^{(\tau+1)} := \boldsymbol{w}^{(\tau)} + \eta \sum_{i=1}^{N} (t_i - \boldsymbol{w}^{(\tau)}.\phi(x_i))\phi(x_i) - 2\eta\lambda \sum_{m=0}^{M-1} \boldsymbol{w}_m^{(\tau)}$



Question 5.3

We can see that the average test error first decreases and then gradually increases as we increase the value of lambda. While the average train error goes on increasing as we increase the value of lambda. We can say that the optimum value of lambda is 1.6 since it gives the lowest possible test error.

As we increase lambda, we put a heavy penalty on the error function and minimize the weights assigned to the features and thereby, oversimplify the model which leads to model underfitting. While when lambda is low, we decrease the penalty which leads to heavy weights assigned to the features thereby making the model too complex i.e. overfitting.