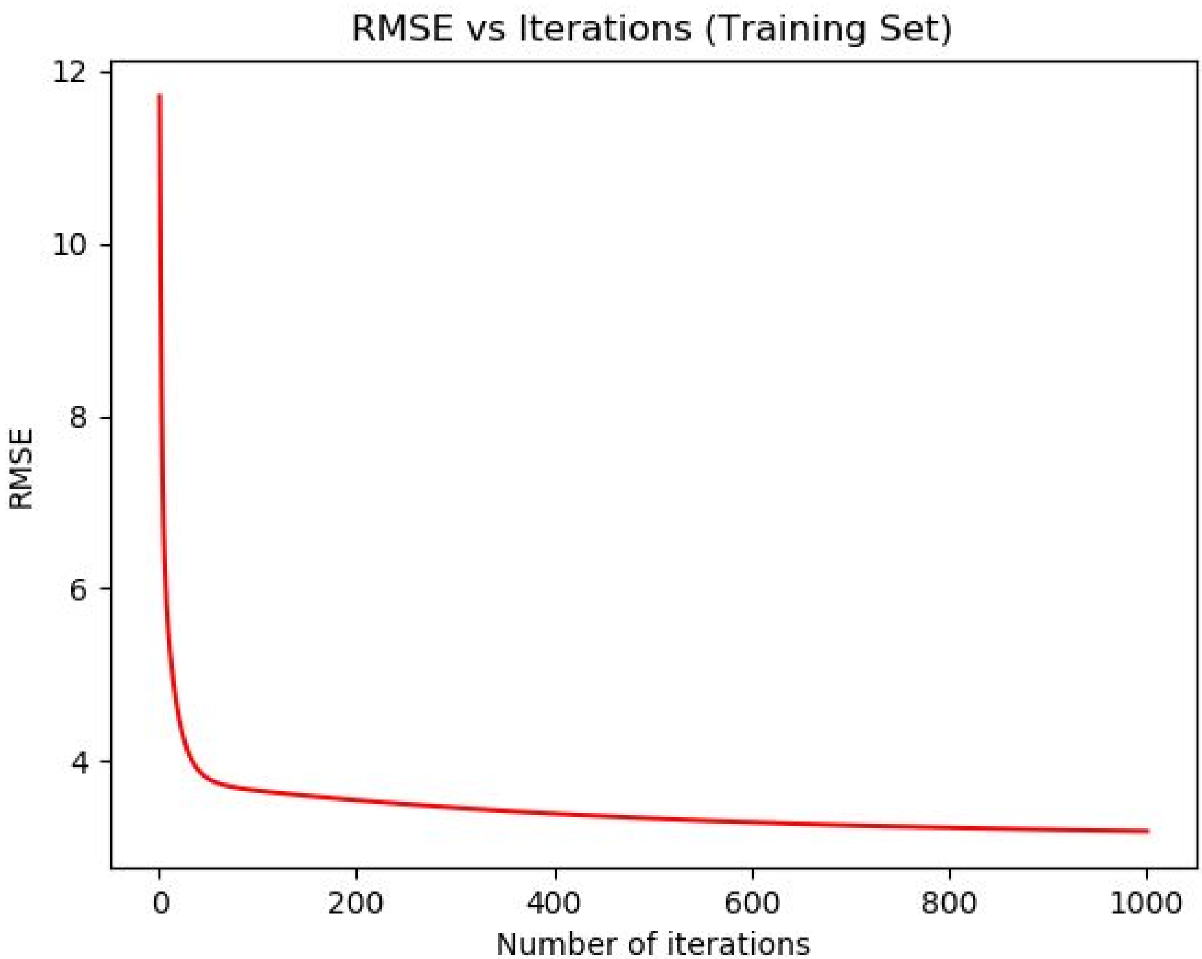
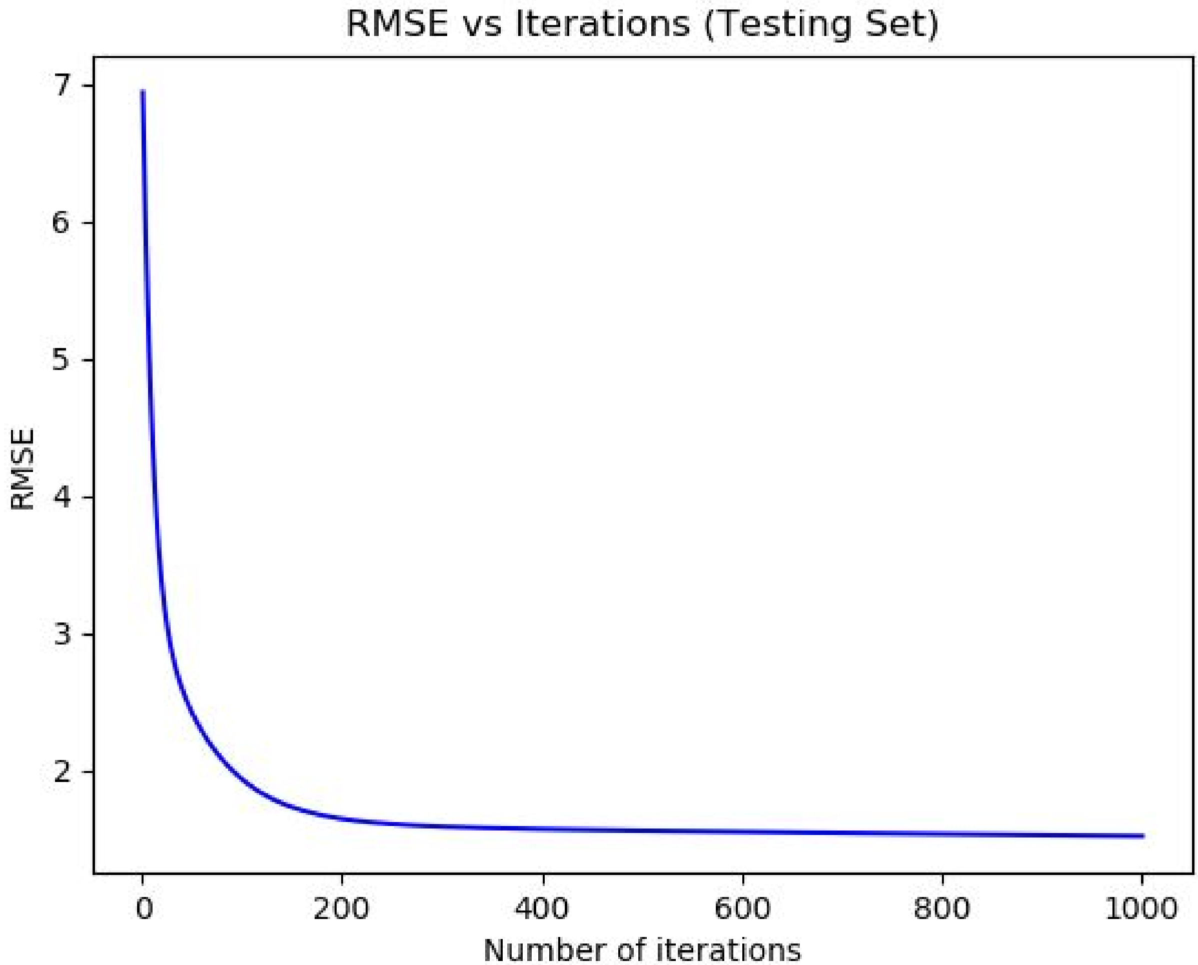
ML Assignment 1

Question.1

1.1.A





1.1.B

RMSE after getting the optimal parameters for training set:

Fold 1: 2.8001345177966335
Fold 2: 3.2810170944896013
Fold 3: 3.055790521800698
Fold 4: 3.2036032361116593
Fold 5: 3.17923904917155

RMSE after getting the optimal parameters for validation set:

Fold 1: 2.0585627949208027
Fold 2: 1.1460196002055758
Fold 3: 1.6554692341633384
Fold 4: 1.3473531103475564
Fold 5: 1.403877903028291

1.1.C

RMSE of the training set from Part A: 3.439031199776404
RMSE of the training set from Part B: 3.1039568838740283

RMSE of the validation set from Part A: 1.708249505765918
RMSE of the validation set from Part B: 1.5222565285331127

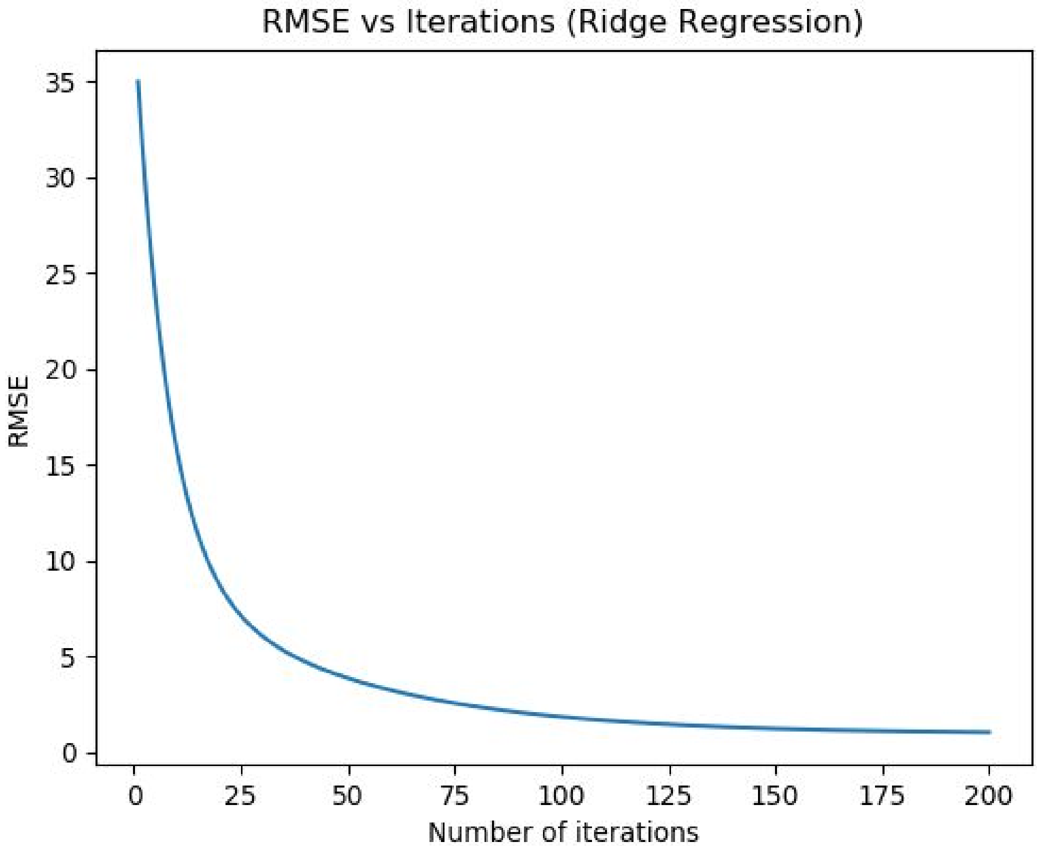
After getting the optimal parameters from normal equation we can see that the RMSE value for both validation and test set decreases as compared to linear regression without optimal parameters. Therefore using the normal equation for finding theta is a good method in this case.

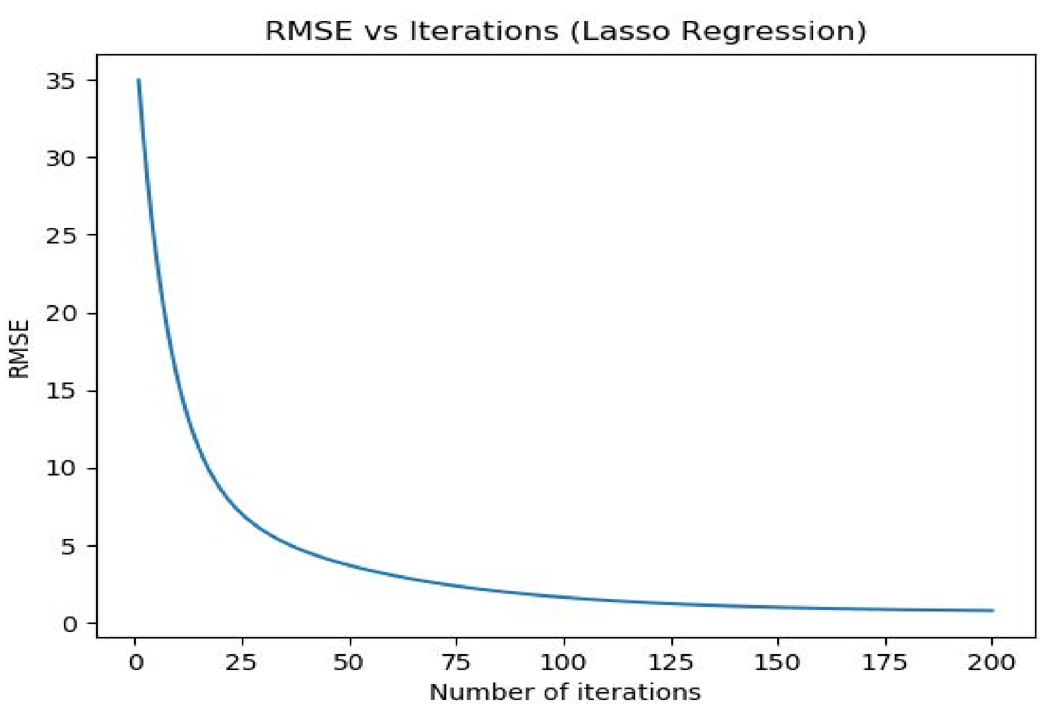
1.2.A

Regularisation Parameter for Ridge Regression: 4
RMSE on Test set for Ridge Regression= 1.0572313559730855

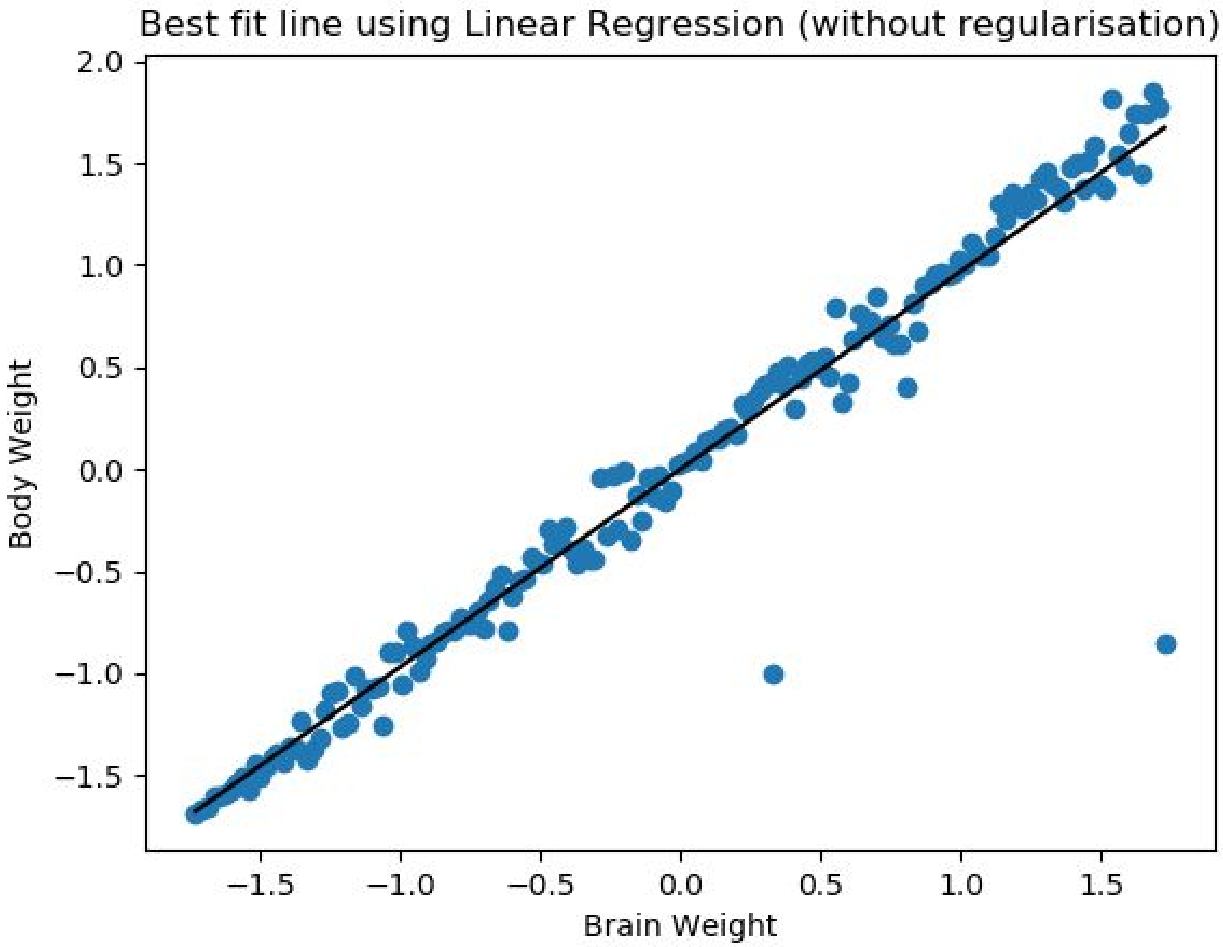
1.2.B

Regularisation Parameter for Lasso Regression: 0.001
RMSE on Test set for Lasso Regression= 0.8091992466374807


RMSE vs Iterations (Ridge Regression)
RMSE
Number of iterations
0
5
10
15
20
25
30
35
0
25
50
75
100
125
150
175
200


RMSE vs Iterations (Lasso Regression)
RMSE
Number of iterations
0
5
10
15
20
25
30
35
0
25
50
75
100
125
150
175
200

1.3.A



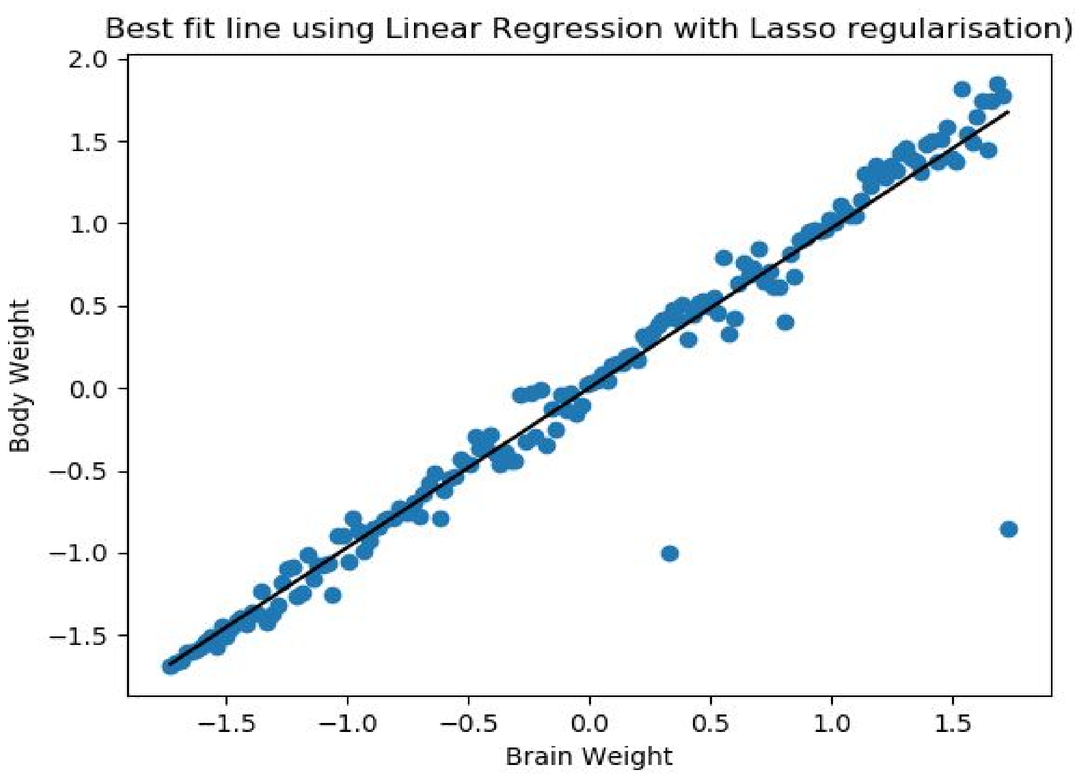
1.3.B

Regularisation Parameter for Ridge Regression: 4

1.3.C

Regularisation Parameter for Lasso Regression: 0.001


Best fit line using Linear Regression with Ridge regularisation)
Body Weight
Brain Weight
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5


Best fit line using Linear Regression with Lasso regularisation)
Body Weight
Brain Weight
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

On using regression without regularisation, minimum RMSE= 0.02955844
On using regression Ridge regularisation, minimum RMSE= 0.04765235
On using regression Lasso regularisation, minimum RMSE= 0.05349572

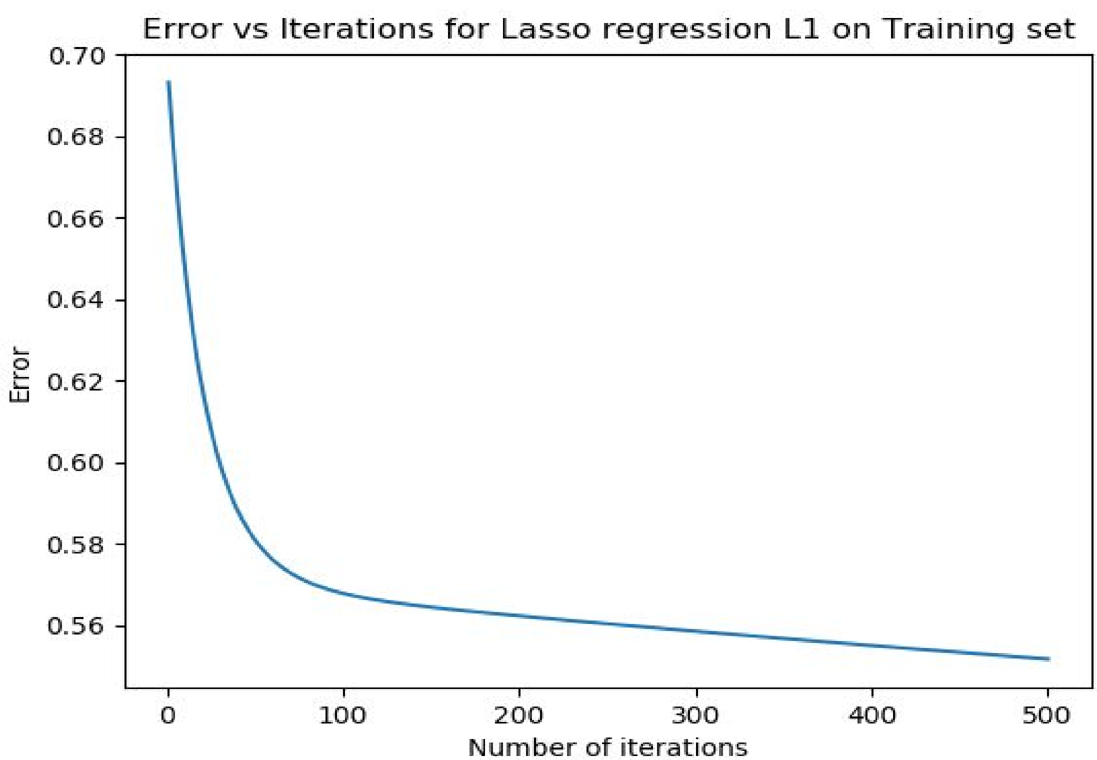
Ridge regularisation performed better than Lasso and overall without using any regularisation performed the best. Although, the visually three of them are almost alike.

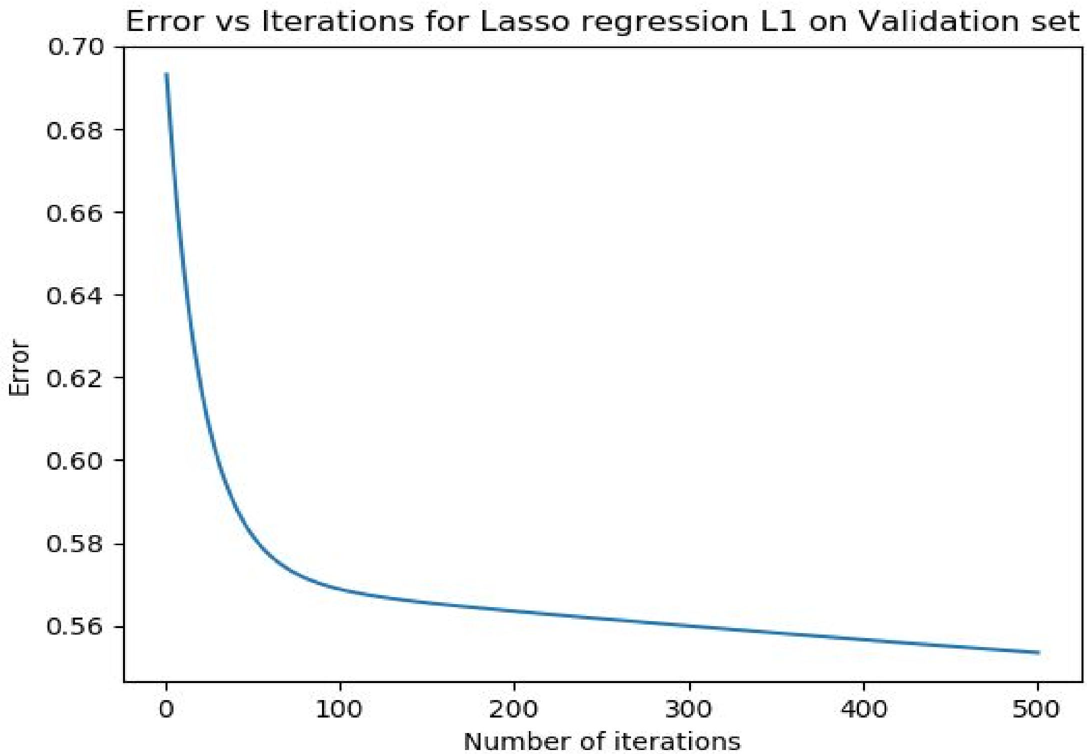
Question. 2

2.1.1

Hyperparameter for Lasso Regularisation: 0.0004887332907777671

Accuracy with Lasso Regularisation on Tain set: 75.9749678809731
Accuracy with Lasso Regularisation on Validation set: 75.7791777188329
Accuracy with Lasso Regularisation on Test set: 76.39285477123315



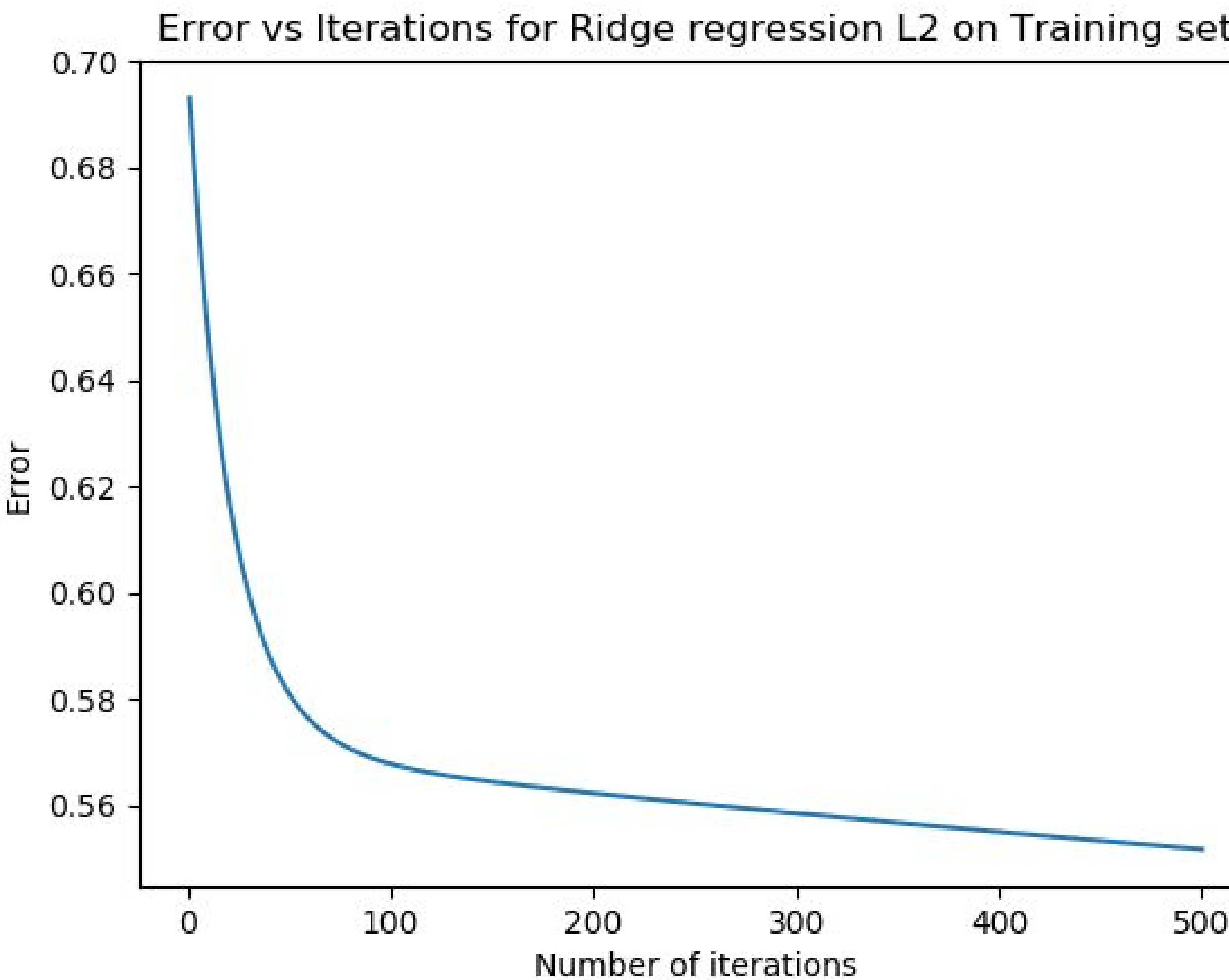

Error vs Iterations for Lasso regression L1 on Validation set
0.70
0.68
0.66
0.64
0.62
0.60
0.58
0.56
Error
0
100
200
300
400
500
Number of iterations

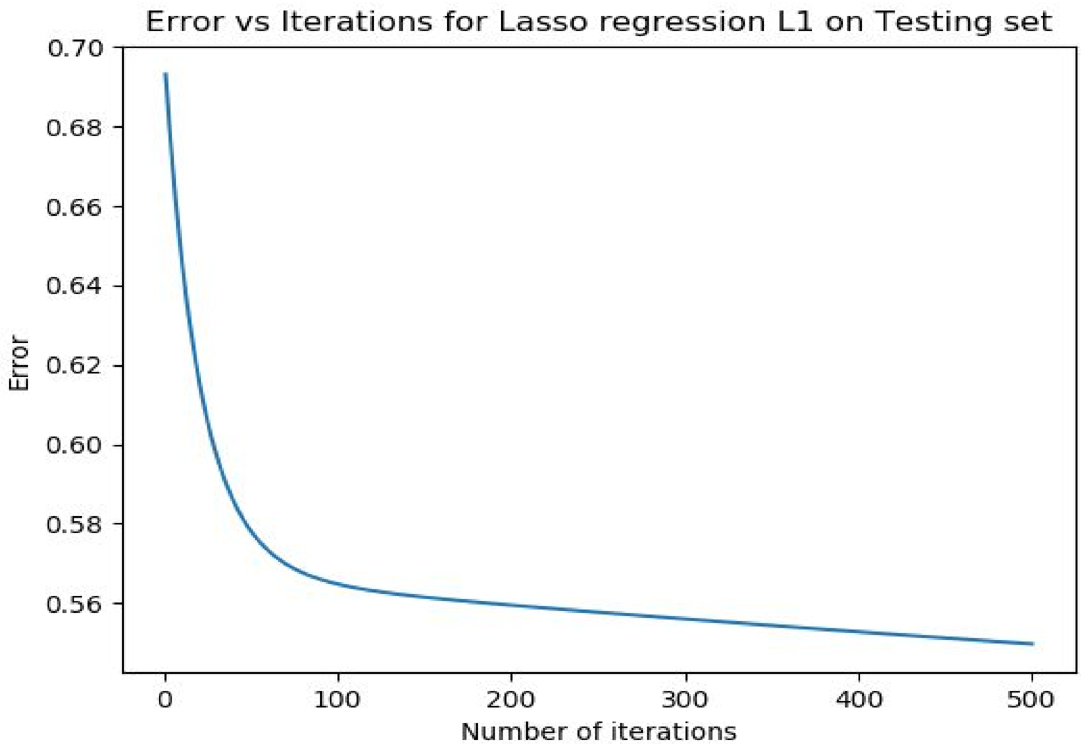

Error vs Iterations for Lasso regression L1 on Testing set
0.70
0.68
0.66
0.64
0.62
0.60
0.58
0.56
Error
0
100
200
300
400
500
Number of iterations

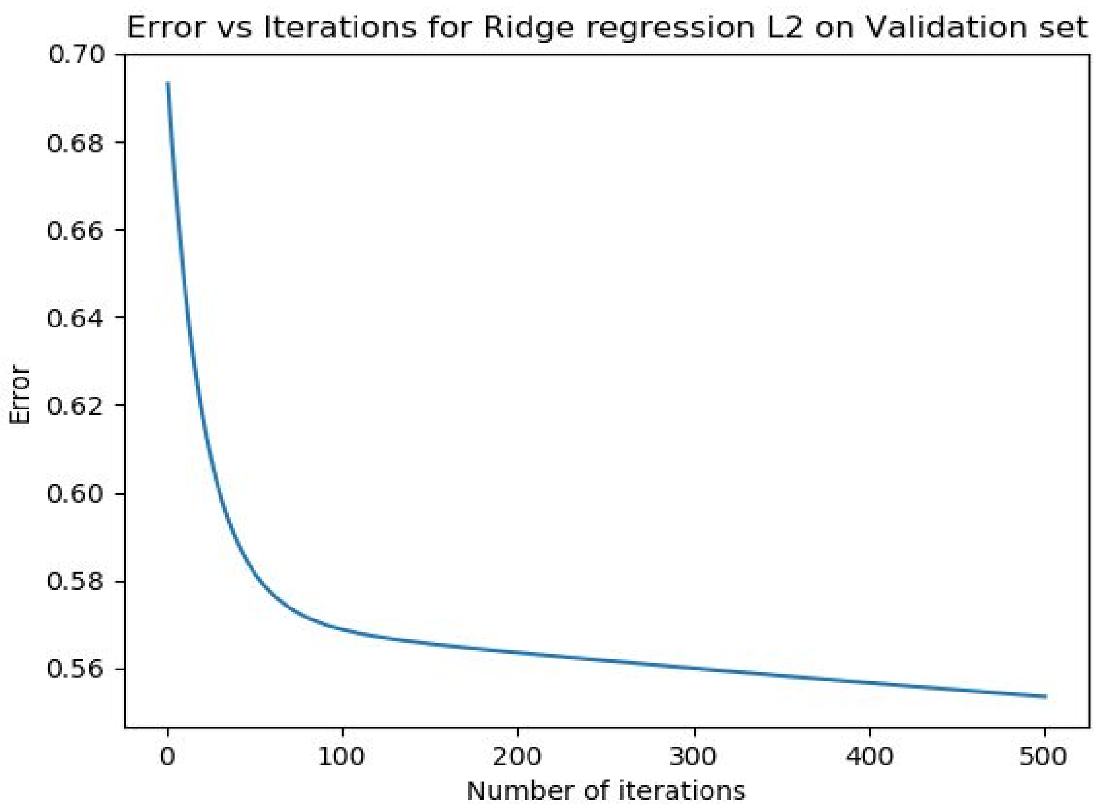
2.1.2

Hyperparameter for Ridge Regularisation: 0.1

Accuracy with Ridge Regularisation on Tain set: 75.9749678809731
Accuracy with Ridge Regularisation on Validation set: 75.7791777188329
Accuracy with Lasso Regularisation on Test set: 76.39285477123315







Accuracy without Regularisation: 76.2402088772846

2.2.1

Accuracy of Train Set in case of L1 Regularisation: 92.56833333333333
Accuracy of Test Set in case of L1 Regularisation: 92.30000000000001

Accuracy of 0 in Training set for L1: 97.06229951038326
Accuracy of 0 in Testing set for L1: 98.26530612244898

Accuracy of 1 in Training set for L1: 97.30050430139424
Accuracy of 1 in Testing set for L1: 97.97356828193833

Accuracy of 2 in Training set for L1: 90.1477005706613
Accuracy of 2 in Testing set for L1: 88.95348837209302

Accuracy of 3 in Training set for L1: 90.57250040776383
Accuracy of 3 in Testing set for L1: 90.99009900990099

Accuracy of 4 in Training set for L1: 93.7521396781924
Accuracy of 4 in Testing set for L1: 93.48268839103869

Accuracy of 5 in Training set for L1: 86.69987087253274
Accuracy of 5 in Testing set for L1: 85.08968609865471

Accuracy of 6 in Training set for L1: 96.04596147347077
Accuracy of 6 in Testing set for L1: 95.09394572025052

Accuracy of 7 in Training set for L1: 93.3439744612929
Accuracy of 7 in Testing set for L1: 92.41245136186771

Accuracy of 8 in Training set for L1: 89.14715433259272
Accuracy of 8 in Testing set for L1: 89.42505133470226

Accuracy of 9 in Training set for L1: 90.48579593208943
Accuracy of 9 in Testing set for L1: 90.08919722497522

2.2.2

Accuracy of Train Set in case of L2 Regularisation: 92.89833333333334
Accuracy of Test Set in case of L2 Regularisation: 92.30000000000001

Accuracy of 0 in Training set for L2: 96.85969947661658
Accuracy of 0 in Testing set for L2: 97.55102040816327

Accuracy of 1 in Training set for L2: 97.44882824087807
Accuracy of 1 in Testing set for L2: 98.06167400881057

Accuracy of 2 in Training set for L2: 91.18831822759316
Accuracy of 2 in Testing set for L2: 89.92248062015504

Accuracy of 3 in Training set for L2: 92.02413961833307
Accuracy of 3 in Testing set for L2: 91.88118811881188

Accuracy of 4 in Training set for L2: 93.08456008216365
Accuracy of 4 in Testing set for L2: 92.4643584521385

Accuracy of 5 in Training set for L2: 88.37852794687328
Accuracy of 5 in Testing set for L2: 87.66816143497758

Accuracy of 6 in Training set for L2: 96.28252788104089
Accuracy of 6 in Testing set for L2: 94.98956158663883

Accuracy of 7 in Training set for L2: 92.00319233838786
Accuracy of 7 in Testing set for L2: 90.46692607003891

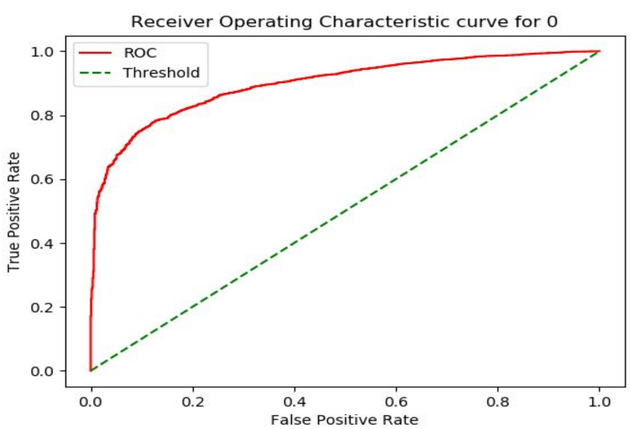
Accuracy of 8 in Training set for L2: 88.85660570842592
Accuracy of 8 in Testing set for L2: 87.37166324435319

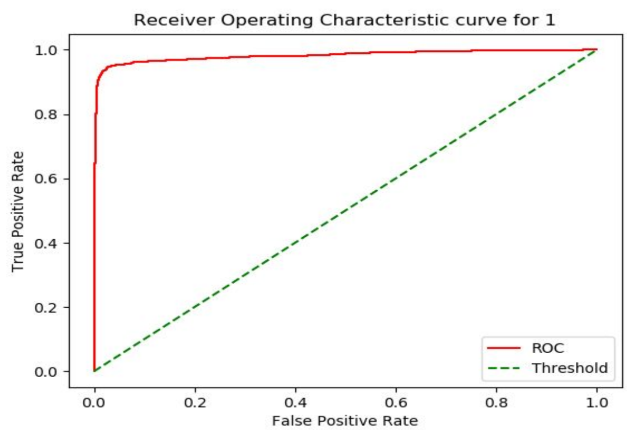
Accuracy of 9 in Training set for L2: 91.89779794923517
Accuracy of 9 in Testing set for L2: 91.57581764122894

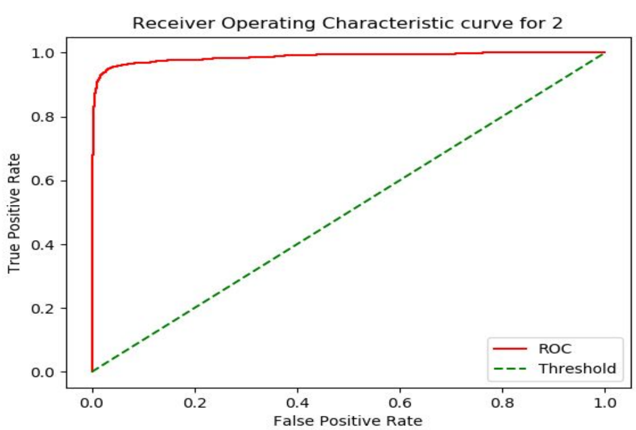
2.2.3

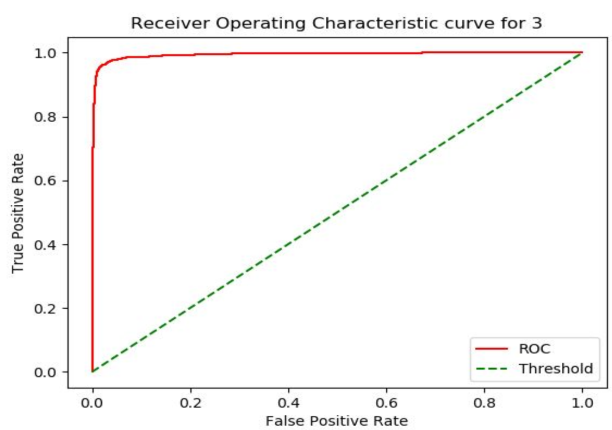
After introducing L1, L2 regularisation parameters we do not see much change in the accuracy of each digit as it was in case of without regularisation. Therefore it is a good fit for the model. Neither underfit nor overfit.

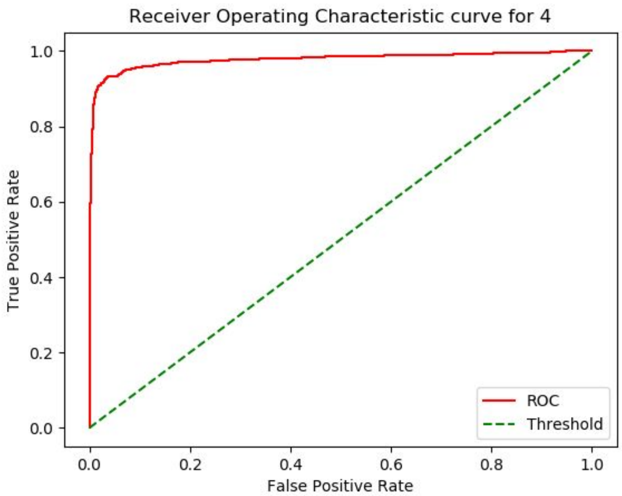
2.3

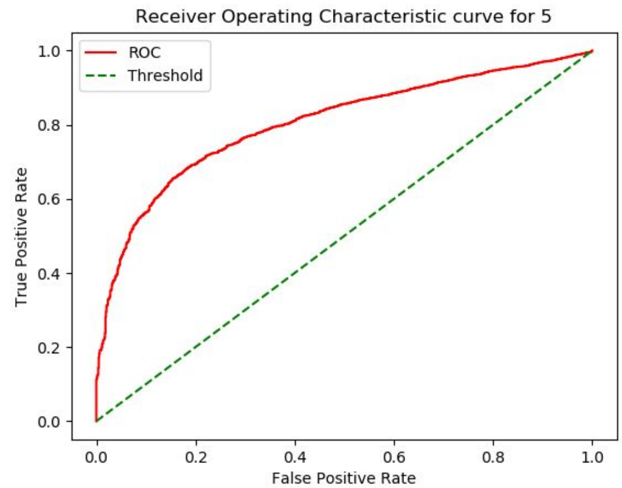


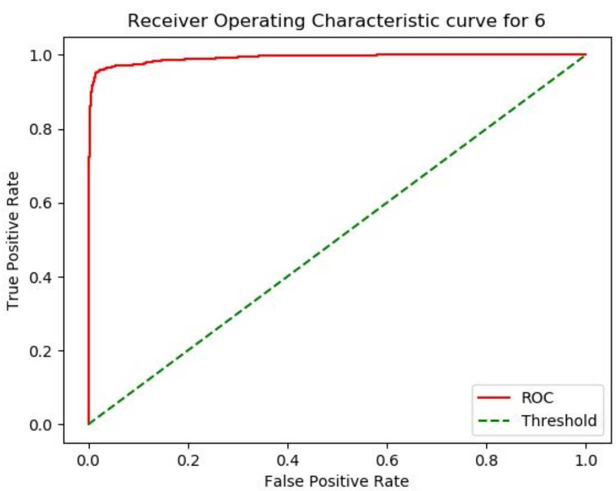


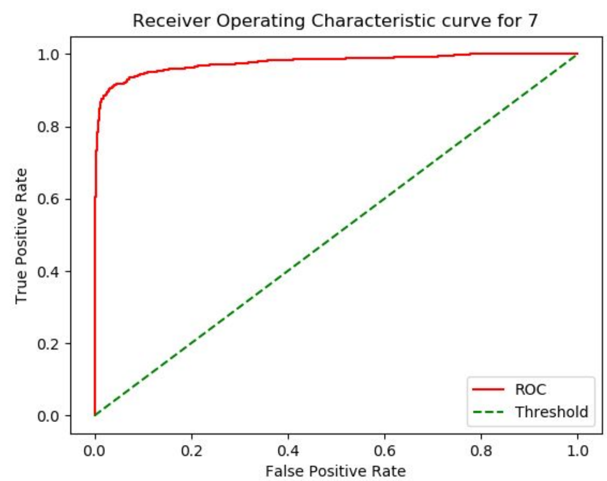


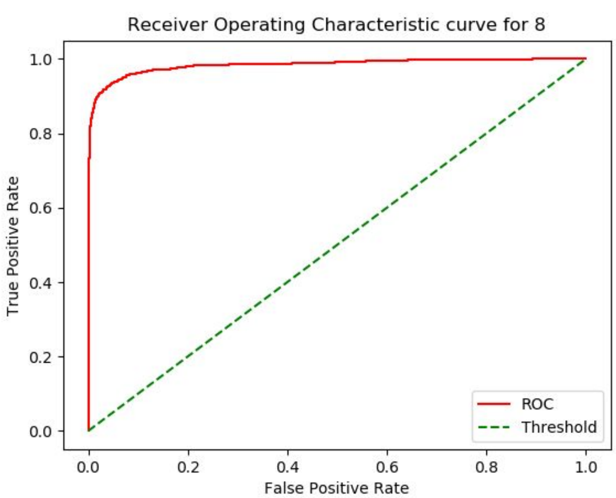


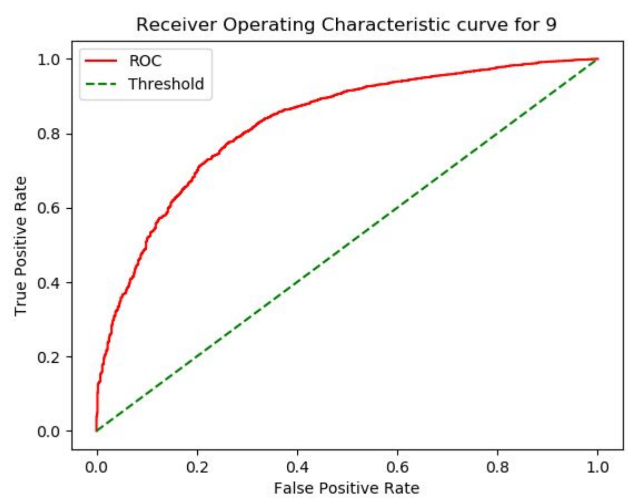

Receiver Operating Characteristic curve for 4
True Positive Rate
False Positive Rate
ROC
Threshold
0.0
0.2
0.4
0.6
0.8
1.0


Receiver Operating Characteristic curve for 5
True Positive Rate
False Positive Rate
ROC
Threshold
0.0
0.2
0.4
0.6
0.8
1.0


Receiver Operating Characteristic curve for 6
True Positive Rate
False Positive Rate
ROC
Threshold
0.0
0.2
0.4
0.6
0.8
1.0


Receiver Operating Characteristic curve for 7
True Positive Rate
False Positive Rate
ROC
Threshold
0.0
0.2
0.4
0.6
0.8
1.0


Receiver Operating Characteristic curve for 8
True Positive Rate
False Positive Rate
ROC
Threshold
0.0
0.2
0.4
0.6
0.8
1.0


Receiver Operating Characteristic curve for 9
True Positive Rate
False Positive Rate
ROC
Threshold
0.0
0.2
0.4
0.6
0.8
1.0

Question 3

Harshit Rai
2017152

M.L. Assignment-1

classmate
Date
Page

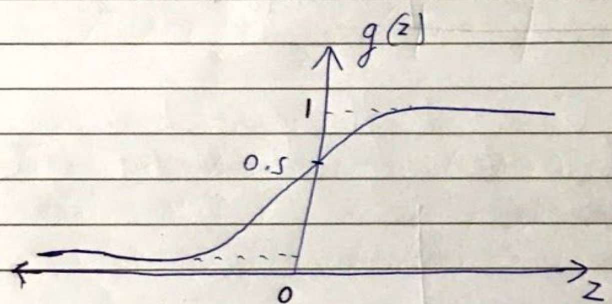
Ans.3) ① $P(y=1 / x, w) = g(w_0 + w_1 x)$.

and, $g(z) = \dfrac{1}{1+e^{-z}}$ (Sigmoid Function).

$\therefore g(w_0 + w_1 x) = \dfrac{1}{1+(e)^{-(w_0 + w_1 x)}}$

We know that the range of Sigmoid Function, $g(z)$ is between 0 to 1

=)



Range of $g(z) = (0, 1)$

when, $z \to \infty$, $g(z) \to 1$ $\left[\lim_{z\to\infty} g(z) = 1\right]$

and, $z \to -\infty$, $g(z) \to 0$ $\left[\lim_{z\to-\infty} g(z) = 0\right]$

When $z = w_0 + w_1 x$, the range of $g(w_0 + w_1 x)$ will not change.

P.T.O.

Date
Page

(2)

=> Range of $g(w_0 + w_1 x)$ will not change because as z and x have a linear relationship and that doesn't affect the range of sigmoid function.

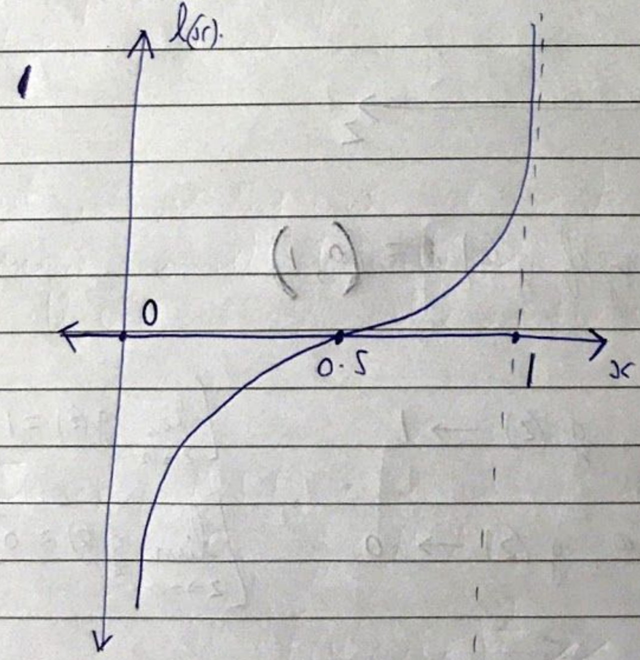
Therefore => when, $w_0 + w_1 . x \rightarrow -\infty \Rightarrow g(w_0 + w_1 x) \rightarrow 0$

and, when, $w_0 + w_1 x \rightarrow \infty \Rightarrow g(w_0 + w_1 x) \rightarrow 1$

$\boxed{\text{Range of } P = (0, 1)}$ where, $P(y=1 | x, w) = g(w_0 + w_1 x)$

Ans

(ii)



=> Logit Function,

$$l(x) = \ln\left(\frac{x}{1-x}\right)$$

Question 4

(3)

$\boxed{\text{Range of logit function in } [0,1] = (-\infty, \infty)}$ ← Ans

because from the graph we can see that the logit function spans all real numbers from $(-\infty, \infty)$ in the range from $x=0$ to $x=1$

Ans.4) (A) $$M.S.E. = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{x}_i)^2$$

$$M.A.E. = \frac{1}{N} \sum_{i=1}^{N} |x_i - \hat{x}_i|$$

The mean Absolute Error (M.A.E.) is harder to optimize than M.S.E. as the first order differentiation of M.A.E. takes 2 different values => -1, 1

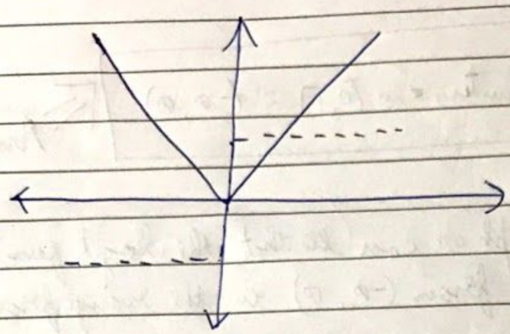
$$\Rightarrow \frac{d}{d\hat{x}_i}\left[\frac{1}{N}\sum_{i=1}^{N}|x_i - \hat{x}_i|\right] = -1, \text{ when } x_i - \hat{x}_i < 0$$

$$\text{and, } \frac{d}{d\hat{x}_i}\left[\frac{1}{N}\sum_{i=1}^{N}|x_i - \hat{x}_i|\right] = 1, \text{ when } x_i - \hat{x}_i > 0$$

~~In other words, M.A.E is not differentiable at y~~

P.T.O.

(4)



From the graph we can see that the M.A.E. is not differentiable at $y_i = \hat{y}_i$

Moreover, the gradient is same throughout which means that the gradient will be large even for small loss values.

Therefore, Mean Absolute error to optimize as it converge to a point with fixed learning rate. M.S.E. is also unique everywhere.

(B) We will use this when we want the loss function more robust to outliners.
Outliners are the extreme +ve or -ve values in training set.

In case of M.S.E, error will be much higher than in case of M.A.E,

$$\left[(M.S.E)e\right]^2 >> \left[(M.A.E)e\right]$$

$$\text{or, } e^2 >> e$$

Therefore we use M.A.E. rather than M.S.E. in such cases.

(c) Quantile regression is used for estimating a conditional "quantile" of a response variable.

It's an extension to M.A.E,

$$M.A.E = \frac{1}{N}\sum_{i=1}^{N} |x_i - \hat{x}_i| .$$

$$L_\gamma(x_i, \hat{x}_i) = \sum_{i: x_i < \hat{x}_i} (\gamma - 1)\cdot|x_i - \hat{x}_i| + \sum_{i: x_i \geq \hat{x}_i} (\gamma)\cdot|x_i - \hat{x}_i|$$

By using quantile loss, we can give different penalties based on a chosen (γ) quantile.

If $\gamma = 0.25$, more penalty will be given to negative errors.

Quantile regression comes in handy when we have to estimate an interval instead of a point prediction.

This method is useful when there is non-constant variance among data points.