IITM-CS5691 : Pattern Recognition and Machine Learning
Release Date: October 9, 2023
Assignment II
Due Date : October 23, 2023, 23:59

Roll No: CS22Z121
Name: Sandeep Kumar Suresh

Collaborators (if any):

References/sources:https://medium.com/@avijit.bhattacharjee1996/implementing-k-fold-cross-validation-from-scratch-in-python-ae413b41c80d

Solution: 1.
a.
For the given dataset given below

x	y
-1	-1
1	$+1$
20	$+1$

$z_i := z(x_i) := wx_i$
0-1 loss function error is given by $\sum_{i=1}^{n}(1 - \text{sign}(y_i z_i))/2$
$L(w) = \sum_{i=1}^{n}(1 - \text{sign}(y_i w x_i))/2$

For w = 0
$z_i = 0$
Therefore for datapoints x = -1 , x = 1 , x = 20
z_i = 0 $\text{sign}(y_i z_i) = 0$
$L(w) = \sum_{i=1}^{n}(0)/2$
$L(w) = 0$

For w = 1
$z_i = x_i$
Therefore for datapoints x = -1 z_i = -1, x = 1 z_i = 1 , x = 20 z_i = 20
$L(w) = \sum_{i=1}^{n}(1 - \text{sign}(y_i x_i))/2$
$L(w) = 0$

Since Loss for both the value is zero , we are not able to determine the actual classifier value of w that makes it a good classifier.

b.

The squared loss function is given by $\sum_{i=1}^{n}(y_i - z_i)^2$ and the logistic loss function is given by $\sum_{i=1}^{n} \log(1 + \exp(-y_i z_i))$

Let L2 be the Squared loss function and L3 be the Logistic Loss Function .

Let us take the case of L2

$\sum_{i=1}^{n}(y_i - z_i)^2$
For w = 0 , $z_i = wx_i \rightarrow z_i = 0$

L(2) $= \sum_{i=1}^{n}(y_i - z_i)^2$ = $\sum_{i=1}^{n}(y_i)^2$ = 3

For w = 1 , $z_i = x_i$

L(2) $= \sum_{i=1}^{n}(y_i - z_i)^2$ = $\sum_{i=1}^{n}(y_i - x_i)^2$ = 361

Similarly

L(3) = $\sum_{i=1}^{n} \log(1 + \exp(-y_i z_i))$

For w = 0 $\Rightarrow z_i = 0$

Therefore L(3) = $\sum_{i=1}^{n} \log(1 + \exp(0)) \Rightarrow$ L(3) =$\sum_{i=1}^{n} \log(2)$

L(3) = 3 $* log(2) = 0.903$

For w = 1 $\Rightarrow z_i = x_i$

L(3) = $\sum_{i=1}^{n} \log(1 + \exp(-y_i x_i))$

$\Rightarrow \log(1 + \exp(-1 * -1 * -1) + \log(1 + \exp(-1 * 1 * 1)) + \log(1 + \exp(-1 * 1 * 20))$

$\Rightarrow \log(1 + \exp(-1) + \log(1 + \exp(-1)) + \log(1 + \exp(-20))$

$\Rightarrow \log(1 + 0.367) + \log(1 + 0.367) + \log(1 + 2.06 * 10^{-9})$

$\Rightarrow$ 0.271

For the case of Logistic Loss function Error for w = 0 is greater than Error for w =1 . Therefore w = 1 will be value for Logistic Loss Function.

In the case of Outliers like x = 20 , Squared Error Loss will have higher value and is not preffered since it will predict wrong value.

The 0-1 Loss cannot be used to make any distinction between the datapoint , whereass Logistic Regression can make a good distinction .

Therefore Logistic Regression is preffered .

Solution: 2.a

The Gradient Descent Formula for the logistic regression model is given by the equation.

$W_{t+1} = W_t + \eta \sum_{i=1}^{n} \sigma(-y_i w^\mathsf{T} x_i)(y_i x_i)$

For $\eta = 1$

$W_{t+1} = W_t + \sum_{i=1}^{n} \sigma(-y_i w^\mathsf{T} x_i)(y_i x_i)$

$$w_1 = \tfrac{1}{2}\left[\begin{bmatrix} -1 \\ -5 \end{bmatrix} + \begin{bmatrix} 0 \\ -10 \end{bmatrix} + \begin{bmatrix} -1 \\ -13 \end{bmatrix} + \begin{bmatrix} 0 \\ -2 \end{bmatrix} + \begin{bmatrix} -1 \\ -3 \end{bmatrix} + \begin{bmatrix} 0 \\ 5 \end{bmatrix} + \begin{bmatrix} 0 \\ 2 \end{bmatrix} + \begin{bmatrix} 1 \\ 10 \end{bmatrix} + \begin{bmatrix} 0 \\ 10 \end{bmatrix} + \begin{bmatrix} 0 \\ 3 \end{bmatrix}\right]$$

$$w_1 = \tfrac{1}{2}\begin{bmatrix} -2 \\ -3 \end{bmatrix}$$

Similarly we need to calculation of w2

$$\sigma\left(-y_i w_1^\mathsf{T} x_i\right)\left(y_i x_i\right)$$

$$= \tfrac{1}{1+\exp(8\cdot 5)}\begin{bmatrix} 1 \\ 5 \end{bmatrix} + \tfrac{1}{1+\exp(15)}\begin{bmatrix} 0 \\ 10 \end{bmatrix} + \tfrac{1}{1+\exp(205)}\begin{bmatrix} 1 \\ 13 \end{bmatrix}$$

$$+ \tfrac{1}{1+\exp(5\cdot 5)}\begin{bmatrix} 1 \\ 3 \end{bmatrix} + \tfrac{1}{1+\exp(-7.5)}\begin{bmatrix} 0 \\ 5 \end{bmatrix} + \tfrac{1}{1+\exp(-3)}\begin{bmatrix} 0 \\ 2 \end{bmatrix}$$

$$+ \tfrac{1}{1+\exp(-16)}\begin{bmatrix} 1 \\ 10 \end{bmatrix} + \tfrac{1}{1+\exp(-15)}\begin{bmatrix} 0 \\ 10 \end{bmatrix} + \tfrac{1}{1+\exp(-4\cdot 5)}\begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} -1 \\ -29.76 \end{bmatrix}$$

$$w_2 = w_1 + \eta \sum_{i=1}^{n} \sigma\left(-y_i w^\top xi\right)\left(y_i \times x_i\right)$$

$$w_2 = \begin{bmatrix} -1 \\ -1.5 \end{bmatrix} - \begin{bmatrix} -1 \\ -29.76 \end{bmatrix} = \begin{bmatrix} 0 \\ 28.26 \end{bmatrix}$$

Solution: 2.b

Prediction of a new datapoint is given by $(w_{t+1}^\mathsf{T})X$

$$\begin{bmatrix} 0 & 28 & 26 \end{bmatrix}\begin{bmatrix} 0 \\ 20 \end{bmatrix} = 585.2$$

$$P(y = 1 \mid X) = \sigma(w^\mathsf{T} X)$$

$$P\left(y = 1 \mid x = \begin{bmatrix} 0 & 20 \end{bmatrix}^\top\right) = \tfrac{1}{1+\exp(5852)} \approx 0$$

$$\therefore P(y = -1 \mid x) = 1 - P(y = 1 \mid x) \approx 1$$

The point belongs to class Y = -1

Solution: 2.c
Logistic Regression gives a probability estimation to the Classification problem .Therefore it is a good choice to evaluate using Logistic Regression.

Solution: 3.
For problem i. , ii. , iii. we need to understand the soft margin parameter C. For high value of C , the given problem becomes hard-margin problem , for lower value of C , there it becomes a soft-margin problem where there can be points for which $\zeta > 0$.

i. Linear Kernel with C = 1 ⇒ b
ii. Linear Kernel with C = 10 ⇒ f
iii. Linear Kernel with C = 0.1 ⇒ c

For problem iv. , v. , vi. , the equation of the Kernel is given by

$$K(u, v) = e^{-k||u-v||^2}. \tag{1}$$

where $k = 1/\sigma^2$
From the Kernel equation , we can say that if the points are similar then ,

$$||u - v|| \tag{2}$$

is small . If $k > 0$, it follows that

$$-k||u - v||^2 \tag{3}$$

will be larger.

Therfore we can that closer vectors will have larger RBF Kernel value than the points that are farthest away.

From equation 1 , σ^2 is the variance of the Gaussian Distribution.

Hence For Larger value of σ^2 , the distribution will be wider , and hence the k will be smaller. Smaller value of σ^2 , the distribution will be narrower and hence the value of k will be larger.

iv . RBF Kernel with k =1 , C =3 ⇒ d
$v. RBF Kernel with k = 0.1, C = 15 \Rightarrow a$
$vi. RBF Kernel with k = 10, C = 1 \Rightarrow e$

Solution: 4.a

Based on the train and validation test errors we can find if the model suffers from the high variance or high bias.The model suffers from high variance by determining if the training error is close to zero and the test/validation error is high . Or in other words we can say during training the model has zero errors and when testing the dataset , the error is high .

The model suffers from a high bias if the if there are high training and test errors. We can say that the model is unable to learn from the features of the data.

Solution: 4.b
Code of the above question have been attached.

Solution: 4.c.i

Uploaded the Code

4.c.ii
The below is the table for Average Training and Validation Error

Lambda	Average Training Error	Average Test Error
1×10^{-15}	0.89	2.10
1×10^{-9}	0.90	1.78
1×10^{-6}	0.91	1.58
0.001	0.94	1.49
0.01	0.97	1.45
0.1	1.03	1.46
1	1.13	1.52
10	1.39	1.74
100	1.99	2.30
1000	2.87	3.16
10^6	3.71	3.97
10^9	4.40	4.64
10^{15}	4.99	5.21

Table 1: Average Training and Testing Error for Different Regularisation Values

Solution: 4.d

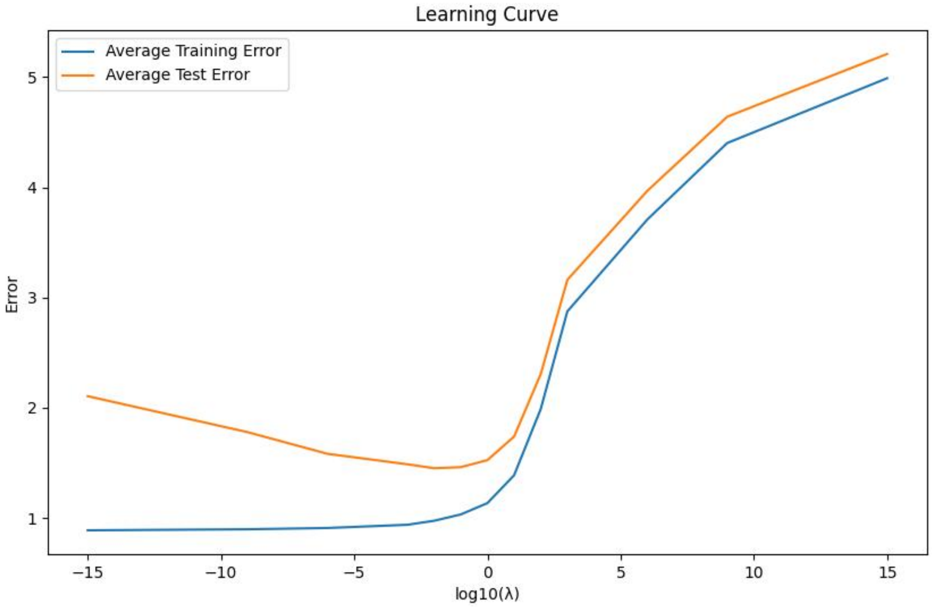

Figure 1: Learning Curve

Based on the Learning Curve, the model can be said to have
i) high bias : For $\lambda = 10^{15}$, the training error is high . The model is unable to fit the training data itself .

ii) highest variance : For $\lambda = 10^{-15}$, there is a gap between cross-validation error and training error which can be seen as the variance of the model. Since the training error is low and test error is high , the model is said to have high variance.

iii) Best Model : The model that works best on unseen data will be for $\lambda = 0.01$. Here both the training and the validation error is small.

Solution:

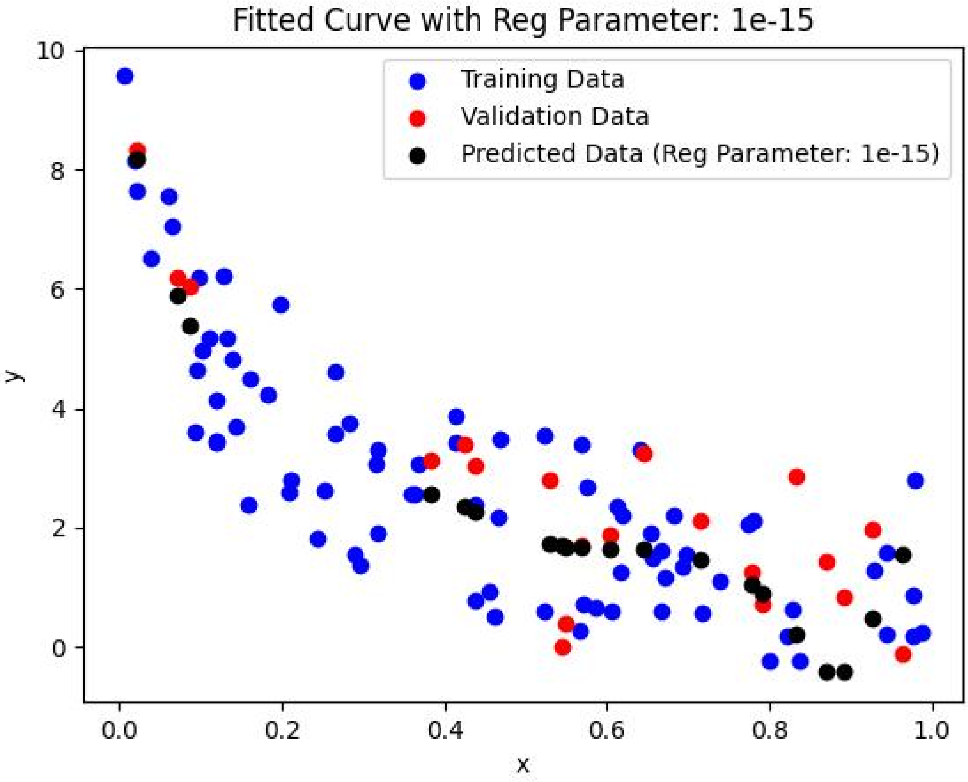

Figure 2: Fitted Curve with Reg Parameter:1e-15

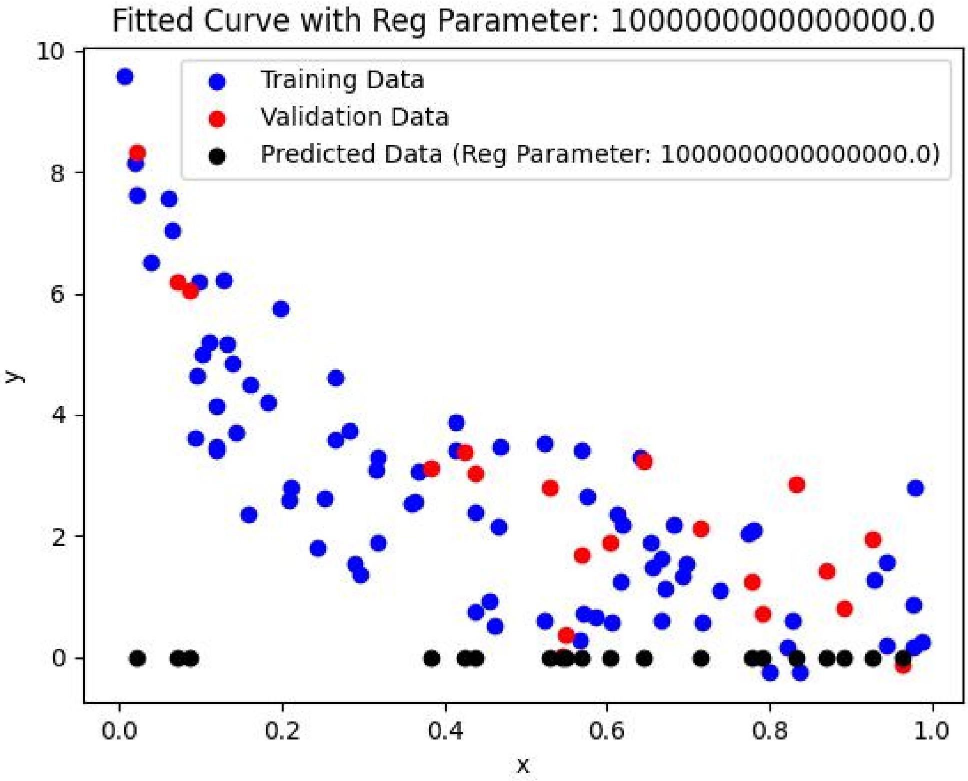

Figure 3: Fitted Curve with Reg Parameter:1000000000000000.0.

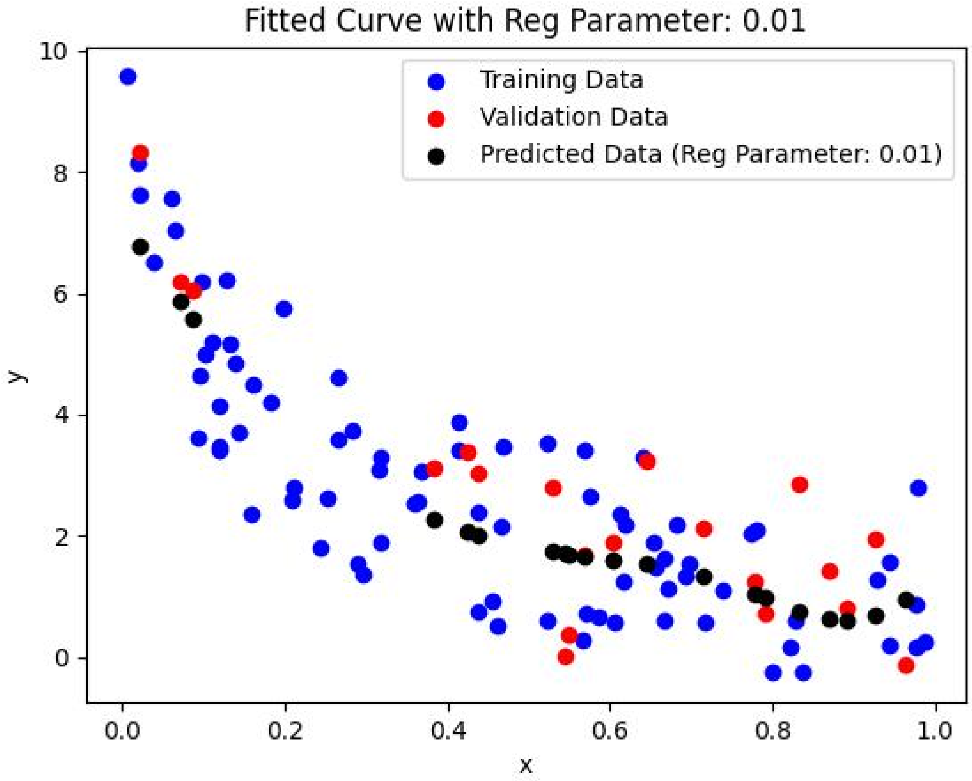

Figure 4: Fitted Curve with Reg Parameter:0.01