Umair Afzal
HW4
CS589

Collaborators:
Debankita,
Bill

1. Linear Regression and Beyond [40 points + 10 Extra Credit]

1.1.

1.1 Finding the w that minimizes $\|y - Xw\|_2^2 + \lambda \|w\|_2^2 \Leftarrow$ Ridge Reg

$\frac{\partial}{\partial w}$ (loss function) $= 0$

$f(w,\lambda) = (y - Xw)^T (y - Xw) + \lambda w^T w$

$\frac{\partial}{\partial w}(f(w,\lambda)) = 2(X^T X) w - 2X^T y + 2\lambda w$

As we know the derivative of loss function is equal to '0'.

So, $2(X^T X) w - 2X^T y + 2\lambda w = 0$

$2X^T y = 2(X^T X) w + 2\lambda w$

$\cancel{2} X^T y = \cancel{2}\,(X^T X + \lambda I) w$

$w^* = (X^T X + \lambda I)^{-1} X^T y$

1.2.

$(PB^T)(BPB^T + R)^{-1}$

ndx

1.2 $x \to \phi(x)$ Replace x with $\phi(x)$

$w = (\phi^T \phi + \lambda I)^{-1} \phi^T y$

Expression of w^* change for kernel Ridge Regression in a way that w is in the span of data.

1.3.

1.3 Matrix Inversion Lemma $(X^T X + \lambda I)^{-1} X^T y = X^T (X X^T + \lambda I)^{-1} y$

$\hat{y}$ can be written entirely in term of $\phi(x_i)^T \cdot \phi(x_{new})$

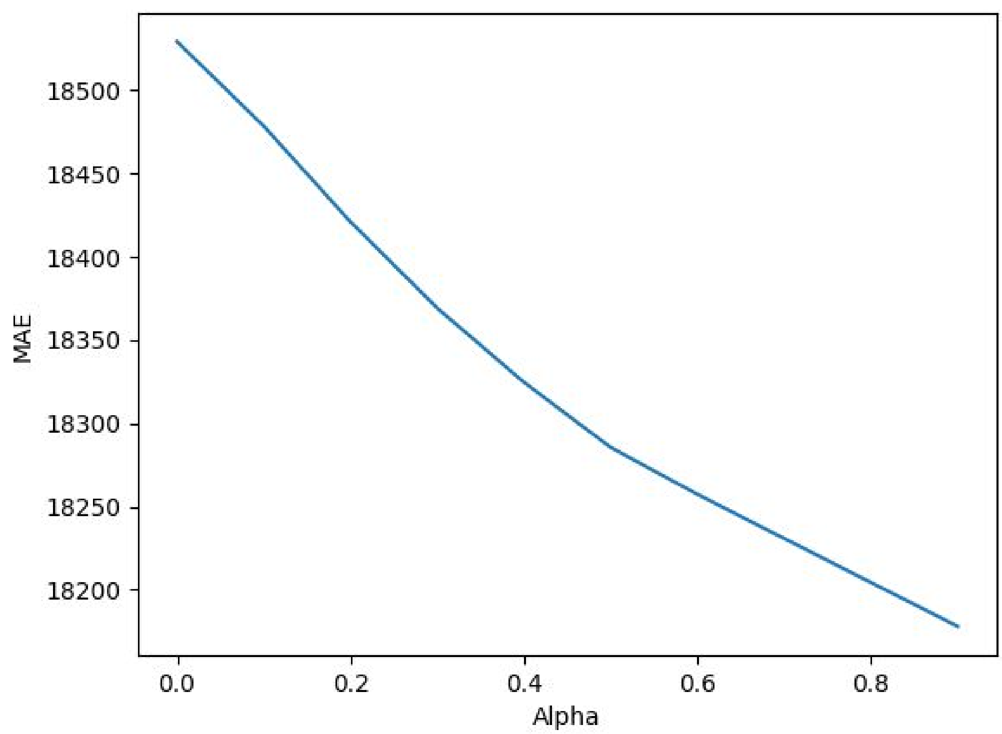
If we project $x_{new} \to w$

1.4.

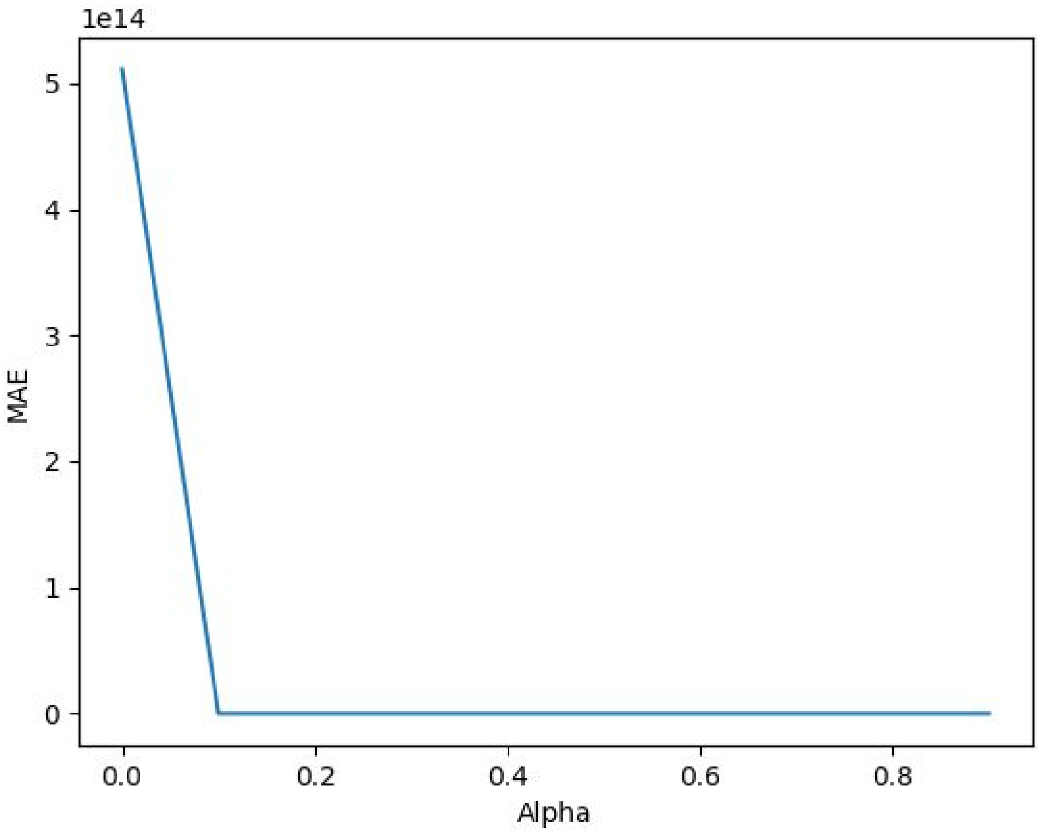
- Mean Absolute Error for OLS LinearRegression() with {'fit_intercept': True, 'normalize': False} = 18619.22403859135
- Mean Absolute Error for Lasso(alpha=0.9, normalize=True) with {'alpha': 0.9, 'max_iter': 1000, 'normalize': True} = 18177.971845335334
- Mean Absolute Error for Ridge(alpha=0.1, max_iter=1000, normalize=True) with {'alpha': 0.1, 'max_iter': 1000, 'normalize': True} = 17825.506218126695

	OLS	Lasso	Ridge
MAE	18619.22403859135	18177.971845335334	17825.506218126695

Lasso Regression Model Graph



Ridge Regression Model Graph



2. Fully Connected Neural Network [40 points]

2.1.

2.1

$$z^{(1)} = f_1(W_1 x_1 + b_1) = \sigma(W_1 x_1 + b_1)$$

$$z^{(2)} = f_2(W_2 z^{(1)} + b_2) = \sigma(W_2 \sigma(W_1 x_1 + b_1) + b_2)$$

$$z^{(3)} = f_3(W_3 z^{(2)} + b_3)$$

$$= f_3(W_3 \sigma(W_2 \sigma(W_1 x_1 + b_1) + b_2) + b_3)$$

$$= \sigma(W_3 \sigma(W_2 \sigma(W_1 x_1 + b_1) + b_2) + b_3)$$

2.2.

<table>
<tr>
<td>**2.2c**</td>
<td>
2.2c)
$\frac{\partial L}{\partial b_3} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial b_3}$
$= \left(\frac{-y}{\hat{y}} + \frac{(1-y)}{1-\hat{y}}\right)\left(\hat{y}(1-\hat{y})\right)$
$= \frac{-y(1-\hat{y}) + \hat{y}(1-y)}{\cancel{(\hat{y})}\cancel{(1-\hat{y})}} \cancel{(\hat{y})}\cancel{(1-\hat{y})}$
$= -y + \cancel{y\hat{y}} + \hat{y} - \cancel{y\hat{y}}$
$= \hat{y} - y$

$\frac{\partial L}{\partial \hat{y}} = \frac{-y}{\hat{y}} + \frac{(1-y)}{1-\hat{y}}$
$\frac{\partial \hat{y}}{\partial b_3} = \frac{\partial \sigma(W_3 z^{(2)} + b_3)}{\partial W_3 z^{(2)} + b_3} \cdot \frac{\partial (W_3 z^{(2)} + b_3)}{\partial b_3}$
$= \hat{y}(1-\hat{y})(1)$
</td>
</tr>
<tr>
<td>**2.2d**</td>
<td>
2.2 d) $\frac{\partial L}{\partial W_2} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z^{(2)}} \frac{\partial z^{(2)}}{\partial W_2}$
★ • $\frac{\partial L}{\partial \hat{y}}$ From part a solution (given)
$\frac{-y(1-\hat{y}) + \hat{y}(1-y)}{(\hat{y})(1-\hat{y})} = \frac{-y + \cancel{y\hat{y}} + \hat{y} - \cancel{y\hat{y}}}{(\hat{y})(1-\hat{y})}$
★ $= \frac{\hat{y} - y}{\hat{y}(1-\hat{y})}$

★
• $\frac{\partial \hat{y}}{\partial z^{(2)}} = \frac{\partial \sigma(W_3 z^{(2)} + b_3)}{W_3 z^{(2)} + b_3} \frac{\partial (W_3 z^{(2)} + b_3)}{\partial z^{(2)}}$
$= (1-\hat{y})(\hat{y}) \left(\frac{\partial W_3 z^{(2)}}{\partial z^{(2)}} + \cancel{\frac{\partial b_3}{\partial z^{(2)}}}\right)$
$= (1-\hat{y})(\hat{y}) W_3^T$ ★
• $\frac{\partial z^{(2)}}{\partial W_2} = \frac{\partial (\sigma(W_2 z^{(1)} + b_2))}{\partial (W_2 z^{(1)} + b_2)} \frac{\partial (W_2 z^{(1)} + b_2)}{\partial W_2}$
$= (1-z^{(2)})(z^{(2)})(z^{(1)})^T$

...
$\frac{\partial L}{\partial W_2} = \frac{\hat{y} - y}{\cancel{\hat{y}(1-\hat{y})}} \left(\cancel{(1-\hat{y})}\cancel{(\hat{y})} W_3^T\right)\left((1-z^{(2)})(z^{(2)})(z^{(1)})^T\right)$
$= (\hat{y} - y)(W_3)^T (1-z^{(2)})(z^{(2)})(z^{(1)})^T$
</td>
</tr>
</table>

2.2e

2.2 e) As calculated already in 2.2 part d. I am going to use those results marked ★ from previous part.

- $\frac{\partial L}{\partial \hat{y}} = \frac{\hat{y} - y}{\hat{y}(1-\hat{y})}$
- $\frac{\partial \hat{y}}{\partial z^{(2)}} = (1-\hat{y})(\hat{y}) W_3^T$
- $\frac{\partial z^{(2)}}{\partial b_2} = \frac{\partial \sigma(W_2 z^{(1)} + b_2)}{\partial (W_2 z^{(1)} + b_2)} \frac{\partial (W_2 z^{(1)} + b_2)}{\partial b_2}$

$$= z^{(2)}(1 - z^{(2)})$$

2.2e) continued

$$\frac{\partial L}{\partial b_2} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z^{(2)}} \frac{\partial z^{(2)}}{\partial b_2}$$

$$= \frac{\hat{y} - y}{\cancel{\hat{y}(1-\hat{y})}} \cancel{(1-\hat{y})}\cancel{(\hat{y})}(W_3)^T \left(z^{(2)}(1 - z^{(2)})\right)$$

$$= (\hat{y} - y)(W_3)^T (z^{(2)})(1 - z^{(2)})$$

2.3.

2.3 b	2.3 b) $\frac{\partial L}{\partial w_3} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial w_3}$ $= \frac{\hat{y} - y}{\cancel{\hat{y}(1-\hat{y})}} \cancel{(1-\hat{y})(\hat{y})} (z^{(2)})^T$ $= (\hat{y} - y)(z^{(2)})^T$ • As calculated in previous parts and from given hints $\frac{\partial L}{\partial \hat{y}} = \frac{\hat{y} - y}{(\hat{y})(1-\hat{y})}$ • $\frac{\partial \hat{y}}{\partial w_3} = \frac{\partial(\sigma(w_3 z^{(2)} + b_3))}{\partial(w_3 z^{(2)} + b_3)} \frac{\partial(w_3 z^{(2)} + b_3)}{\partial(w_3}$ $= (1-\hat{y})(\hat{y})(z^{(2)})^T$
2.3c	2.3c) Same result as I calculated in 2.2 c).. Because the deravatie of ReLU doesn't get calulated. $\frac{\partial L}{\partial b_3} = \hat{y} - y$ I am not doing the calculation again because I already did it in 2.2 c)
2.3 d	2.3 d) Following from 2.2 part d, I am using my calculation marked ★ from that part. $\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z^{(2)}} \frac{\partial z^{(2)}}{\partial w_2}$ $= \begin{cases} (\hat{y} - y)(w_3)^T (z^{(1)})^T & z^{(2)} > 0 \\ 0 & z^{(2)} \leq 0 \end{cases}$ ★ $\frac{\partial L}{\partial \hat{y}} = \frac{\hat{y} - y}{\hat{y}(1-\hat{y})}$ ★ $\frac{\partial \hat{y}}{\partial z^{(2)}} = (1-\hat{y})(\hat{y})(w_3)^T$ • $\frac{\partial z^{(2)}}{\partial w_2} = \frac{\partial(ReLU(w_2 z^{(1)} + b_3))}{\partial(w_2 z^{(1)} + b_2)} \frac{\partial(w_2 z^{(1)} + b_2)}{\partial w_2}$ $= \begin{cases} (z^{(1)})^T & z^{(2)} > 0 \\ 0 & z^{(2)} \leq 0 \end{cases}$ $1 \times \frac{\partial(w_2 z^{(1)} + b_2)}{\partial w_2} = (z^{(1)})^T$ $0 \times \frac{\partial(w_2 z^{(1)} + b_2)}{\partial w_2} = 0$

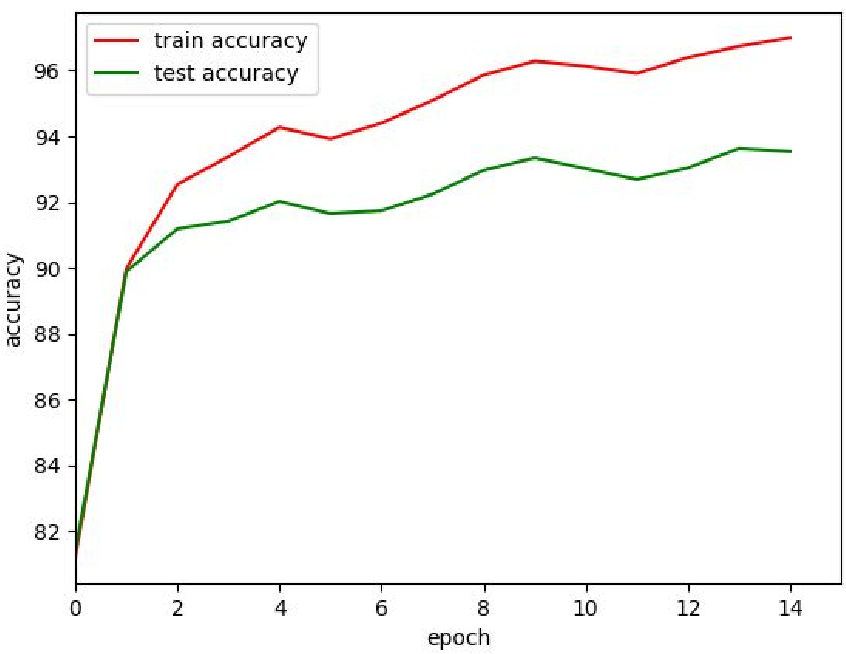
2.4.

2.4 We can use the partial derivatives to optimize the neural network model. When we utilize the weights that are linked with each neuron. Considering the fact that we can get the values of $z^{(2)}$ from the layer before it and W_3 and b_3 can change the result of $z^{(3)}$. Also, we know that the loss function is dependant on $z^{(3)}$ and b_3. We can calculate the partial derivative of the expression and use the gradients in line with W_3 and b_3 to find the weight and biases direction.

$\left(z^{(3)} = W_3 z^{(2)} + b_3 \right)$

$b_3 \Rightarrow b_{3,t+1} = b_{3,t} - \eta \frac{\partial L}{\partial b_3}$ | $W_3 \Rightarrow W_{3,t+1} = W_{3,t} - \eta \frac{\partial L}{\partial W_3}$

2.5.



OUTPUT FROM Q2_TEMPLATE

Train Epoch: 0 Loss: 2.301180

Train Epoch: 0 Loss: 1.706761

Train Epoch: 0 Loss: 1.822461

Train Epoch: 0 Loss: 1.665592

Train set Accuracy: 81%
Test set Accuracy: 81%

Train Epoch: 1 Loss: 1.640361
Train Epoch: 1 Loss: 1.600710
Train Epoch: 1 Loss: 1.692290
Train Epoch: 1 Loss: 1.567268
Train set Accuracy: 90%
Test set Accuracy: 90%

Train Epoch: 2 Loss: 1.596981
Train Epoch: 2 Loss: 1.555976
Train Epoch: 2 Loss: 1.577573
Train Epoch: 2 Loss: 1.548556
Train set Accuracy: 93%
Test set Accuracy: 91%

Train Epoch: 3 Loss: 1.581413
Train Epoch: 3 Loss: 1.577446
Train Epoch: 3 Loss: 1.560676
Train Epoch: 3 Loss: 1.548975
Train set Accuracy: 93%
Test set Accuracy: 91%

Train Epoch: 4 Loss: 1.529510
Train Epoch: 4 Loss: 1.553287
Train Epoch: 4 Loss: 1.538871
Train Epoch: 4 Loss: 1.552762
Train set Accuracy: 94%
Test set Accuracy: 92%

Train Epoch: 5 Loss: 1.517335
Train Epoch: 5 Loss: 1.530920
Train Epoch: 5 Loss: 1.543072
Train Epoch: 5 Loss: 1.526103
Train set Accuracy: 94%
Test set Accuracy: 92%

Train Epoch: 6 Loss: 1.495716
Train Epoch: 6 Loss: 1.522538
Train Epoch: 6 Loss: 1.495137
Train Epoch: 6 Loss: 1.508056
Train set Accuracy: 94%

Test set Accuracy: 92%

Train Epoch: 7 Loss: 1.484285
Train Epoch: 7 Loss: 1.544632
Train Epoch: 7 Loss: 1.527293
Train Epoch: 7 Loss: 1.501924
Train set Accuracy: 95%
Test set Accuracy: 92%

Train Epoch: 8 Loss: 1.472607
Train Epoch: 8 Loss: 1.557030
Train Epoch: 8 Loss: 1.497054
Train Epoch: 8 Loss: 1.476057
Train set Accuracy: 96%
Test set Accuracy: 93%

Train Epoch: 9 Loss: 1.464883
Train Epoch: 9 Loss: 1.502371
Train Epoch: 9 Loss: 1.494863
Train Epoch: 9 Loss: 1.471656
Train set Accuracy: 96%
Test set Accuracy: 93%

Train Epoch: 10 Loss: 1.464039
Train Epoch: 10 Loss: 1.489569
Train Epoch: 10 Loss: 1.466958
Train Epoch: 10 Loss: 1.464883
Train set Accuracy: 96%
Test set Accuracy: 93%

Train Epoch: 11 Loss: 1.462607
Train Epoch: 11 Loss: 1.524543
Train Epoch: 11 Loss: 1.489454
Train Epoch: 11 Loss: 1.466048
Train set Accuracy: 96%
Test set Accuracy: 93%

Train Epoch: 12 Loss: 1.463362
Train Epoch: 12 Loss: 1.483198
Train Epoch: 12 Loss: 1.461668
Train Epoch: 12 Loss: 1.464724
Train set Accuracy: 96%
Test set Accuracy: 93%

Train Epoch: 13 Loss: 1.461373
Train Epoch: 13 Loss: 1.501119
Train Epoch: 13 Loss: 1.481827
Train Epoch: 13 Loss: 1.463881
Train set Accuracy: 97%
Test set Accuracy: 94%

Train Epoch: 14 Loss: 1.462344
Train Epoch: 14 Loss: 1.465648
Train Epoch: 14 Loss: 1.477745
Train Epoch: 14 Loss: 1.484976
Train set Accuracy: 97%
Test set Accuracy: 94