1.

$$f_{initial}(x_0) = x \cdot W$$

$$(0.5, -0.2, 0.1) \cdot \begin{pmatrix} 0.1 & -0.2 & 0.3 & -0.4 \\ 0.4 & -0.3 & 0.2 & -0.1 \\ 0.3 & 0.2 & -0.1 & -0.4 \end{pmatrix}$$

$$= \begin{pmatrix} 0.1 \cdot 0.5 + 0.4 \cdot -0.2 + 0.3 \cdot 0.1 \\ -02 \cdot 0.5 + -0.3 \cdot -0.2 + 0.2 \cdot 0.1 \\ 0.3 \cdot 0.5 + 0.2 \cdot -0.2 + -0.1 \cdot 0.1 \\ -0.4 \cdot 0.5 + -0.1 \cdot -0.2 + -0.4 \cdot 0.1 \end{pmatrix} = \begin{pmatrix} 0 \\ -0.02 \\ 0.1 \\ -0.22 \end{pmatrix} = h_1$$

$$f(h, x_1) = x_1 \cdot W + h \cdot W_{hidden}$$

$$(-0.1, 0.3, 0.4) \cdot \begin{pmatrix} 0.1 & -0.2 & 0.3 & -0.4 \\ 0.4 & -0.3 & 0.2 & -0.1 \\ 0.3 & 0.2 & -0.1 & -0.4 \end{pmatrix} + (0, -0.02, 0.1, -0.22) \cdot \begin{pmatrix} -0.1 & 0.2 & -0.2 & 0.1 \\ 0.3 & -0.4 & 0.1 & -0.2 \\ -0.3 & 0.1 & 0.2 & -0.1 \\ 0.2 & 0.3 & -0.4 & 0.3 \end{pmatrix}$$

$$= \begin{pmatrix} 0.23 \\ 0.01 \\ 0.05 \\ -0.15 \end{pmatrix} + \begin{pmatrix} -0.08 \\ -0.048 \\ 0.106 \\ -0.072 \end{pmatrix} = \begin{pmatrix} 0.15 \\ -0.038 \\ 0.156 \\ -0.222 \end{pmatrix} = h_2$$

$$f(h, x_2) = x_2 \cdot W + h \cdot W_{hidden}$$

$$(0.2, -0.5, -0.3) \cdot \begin{pmatrix} 0.1 & -0.2 & 0.3 & -0.4 \\ 0.4 & -0.3 & 0.2 & -0.1 \\ 0.3 & 0.2 & -0.1 & -0.4 \end{pmatrix} + (0.15 \cdot -0.038 \cdot 0.156 \cdot -0.222) \cdot \begin{pmatrix} -0.1 & 0.2 & -0.2 & 0.1 \\ 0.3 & -0.4 & 0.1 & -0.2 \\ -0.3 & 0.1 & 0.2 & -0.1 \\ 0.2 & 0.3 & -0.4 & 0.3 \end{pmatrix}$$

$$= \begin{pmatrix} -0.27 \\ 0.05 \\ -0.01 \\ 0.09 \end{pmatrix} + \begin{pmatrix} -0.1176 \\ -0.0058 \\ 0.0862 \\ -0.0594 \end{pmatrix} = \begin{pmatrix} -0.3876 \\ 0.0442 \\ 0.0762 \\ 0.0304 \end{pmatrix} = h_3$$

$$h_3 \cdot W_{output}$$

$$(-0.3876, 0.0442, 0.0762, 0.0304) \cdot \begin{pmatrix} -0.3 & 0.3 \\ 0.2 & -0.2 \\ -0.1 & 0.1 \\ 0.4 & -0.4 \end{pmatrix} = \begin{pmatrix} 0.1311 \\ -0.1311 \end{pmatrix}$$

$$\text{softmax}(Z)_{ij} = \frac{e^{z_{ij}}}{\sum_{k=1}^{n} e^{z_{ik}}}$$

$$\begin{pmatrix} 0.1311 \\ -0.1311 \end{pmatrix} \rightarrow \begin{pmatrix} \dfrac{e^{0.1311}}{0.1311} \\ \dfrac{e^{-0.1311}}{-0.1311} \end{pmatrix} = \boxed{\begin{pmatrix} 8.69627 \\ -6.69054 \end{pmatrix} = y}$$

2.

a) The error on the training set

i) 入 = 0

The ridge regression line and the lasso regression line equal the least squares line, so we get the same overfitting line. With the same overfitting line, the error on the training set is decreased.

ii) 入 = optimal

Error on the training set increases. This is because regularization avoids overfitting. Regularization makes our curve simple, in doing so, it introduces some error in the training set but this error enables us to move towards a more generalized model.

iii) 入 = too large

The error on the training set is larger compared to when 入= 0 and when 入 is optimal. When 入 is too large, the model grew to be so simple that it is asymptotically parallel to the x axis and we have a very biased model.

b) The error on the testing set

i) 入 = 0

The ridge regression line and the lasso regression line equal the least squares line, so we get the same overfitting line. With the same overfitting line, the error on the testing set is increased.

ii) 入 = optimal

MSE on the testing set decreases from when 入 = 0. This is because regularization avoids overfitting. Since there is no more overfitting of the line, the MSE for the testing data decreases.

iii) 入 = too large

The MSE on the testing data is higher than when 入 = optimal but less than when 入 = 0. When 入 is too large, the model grew to be so simple that is is asymptotically parallel to the x axis and we have a very biased model.

c) The regression coefficients

i) 入 = 0

1) When 入 = 0, there is no penalty term added to the cost function. The model essentially reverts to the ordinary least squares regression. This can lead to overfitting.

ii) 入 = large but not too large

As the value of 入 increases, the penalty also increases, which in turn, causes the coefficients to shrink in value in order to minimize the cost function.

iii) 入 = too large

1) In ridge: As the value of 入 increases, the penalty also increases, which in turn, causes the coefficients to shrink in value in order to minimize the cost function. The value of the coefficients shrink until the value is asymptotically 0.
2) In lasso: As the value of 入 increases, the penalty also increases, which in turn, causes the coefficients to shrink in value in order to minimize the cost function. The value of the coefficients shrink until the value is 0.

d) The number of non-zero elements in ŵ

i) 入 = 0

When 入 = 0, there is no penalty term added to the cost function. The coefficients in ŵ remain the same; the number of non-zero elements in ŵ remains the same.

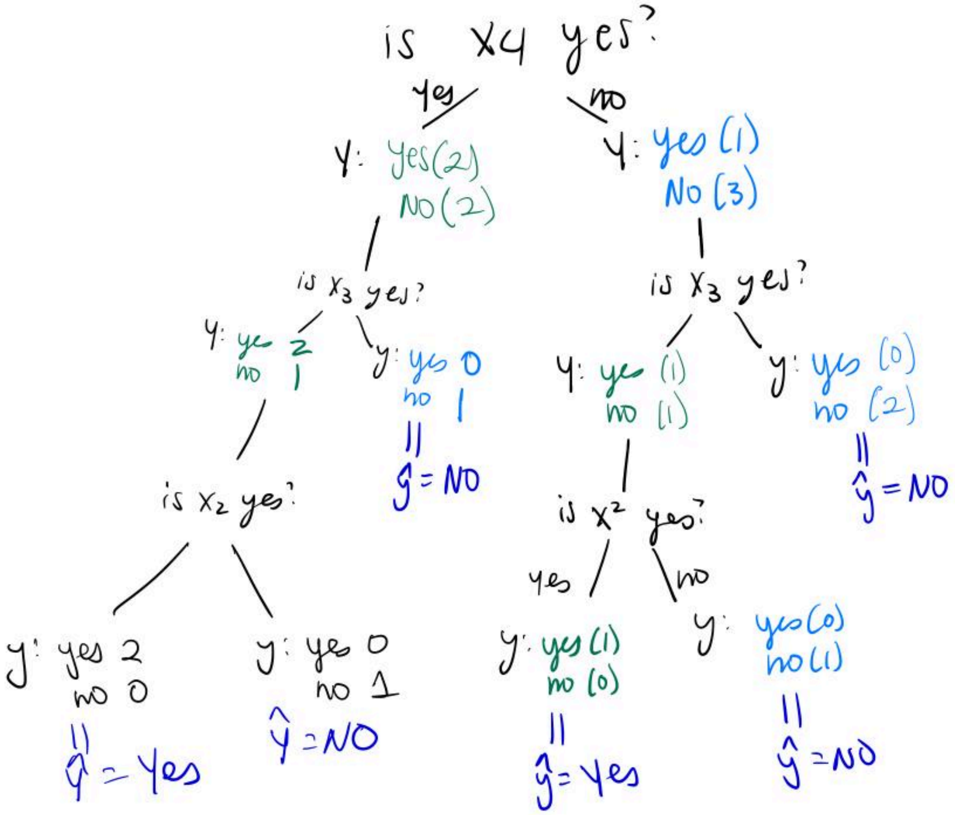
ii) 入 = too large

As the value of 入 increases, the penalty also increases, which in turn, causes the coefficients to shrink in value in order to minimize the cost function. The number of non-zero elements in ŵ increases.

iii) 入 = too large

1) In ridge: As the value of 入 increases, the penalty also increases, which in turn, causes the coefficients to shrink in value in order to minimize the cost function. The value of the coefficients shrink until the value is asymptotically 0; the number of non-zero elements decreases. Most of the values in ŵ may be 0 or asymptotically close to 0.
2) In lasso: As the value of 入 increases, the penalty also increases, which in turn, causes the coefficients to shrink in value in order to minimize the cost function. The value of the coefficients shrink until the value is 0. Most of the coefficients in ŵ will be 0 or equal to 0. Equally, the number of non-zero elements decreases the most when 入 is too large.

3


is X4 yes?
yes
no
Y: Yes(2) No(2)
Y: yes (1) No (3)
is X3 yes?
is X3 yes?
Y: yes 2 no 1
y: yes 0 no 1
|| ŷ = NO
Y: yes (1) no (1)
y: yes (0) no (2)
|| ŷ = NO
is X2 yes?
is X² yes?
yes
no
y: yes 2 no 0
|| ŷ = Yes
y: yes 0 no 1
ŷ = NO
y: yes (1) no (0)
|| ŷ = Yes
y: yes (0) no (1)
|| ŷ = NO

4.

TRee 1

- RSS

$$\sum_{n=1}^{4} (Y_i - \hat{Y})^2 = (50-50)^2 + (60-60)^2 + (70-70)^2 + (80-80)^2 = 0 \text{ RSS}$$

★ $\boxed{\alpha = 1 \rightarrow 0 + 1|4| = 4}$

- $\alpha = 10 \rightarrow 0 + 10|4| = 40$
- $\alpha = 100 \rightarrow 0 + 100|4| = 400$

TREE 2

- RSS

$$\sum_{n=1}^{4} (y_i - \hat{y})^2 = (50-55)^2 + (60-55)^2 + (70-70)^2 + (80-80)^2 = 25 + 25 = 50$$

- $\underline{\alpha = 1 \rightarrow 50 + 1|3| = 53}$
- $\alpha = 10 \rightarrow 50 + 10|3| = 80$
- $\alpha = 100 \rightarrow 50 + 100|3| = 350$

TREE 3

- RSS

$$\sum_{i=1}^{4} (y_i - \hat{y})^2 = (50-55)^2 + (60-55)^2 + (70-75)^2 + (80-75)^2 = 25 + 25 + 25 + 25 = 100$$

- $\underline{\alpha = 1 \rightarrow 100 + 1|2| = 102}$
- $\alpha = 10 \rightarrow 100 + 10|2| = 120$
- $\alpha = 100 \rightarrow 100 + 100|2| = 300$

best tree is TREE 1 with $\alpha = 1$