Practice Assignment

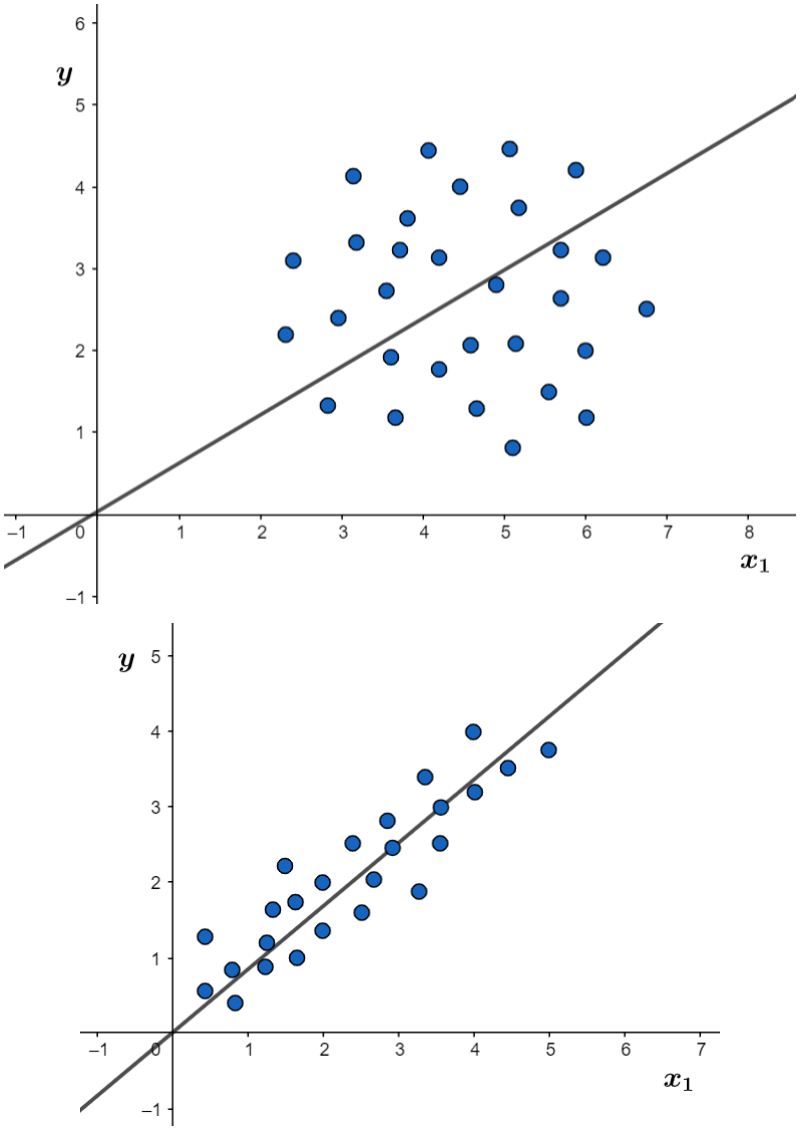
Note:

1. In the following assignment, X denotes the data matrix of shape (d, n) where d and n are the number of features and samples, respectively.
2. x_i denotes the i^{th} sample and y_i denotes the corresponding label.
3. w denotes the weights (parameter) in the linear regression model.

Question 1

Statement

Consider the following two models for two different datasets:



Models are represented by the line and both the graphs are on the same scale. Which model will give the more training error?

Options

(a)

Model 1

(b)

Model 2

Answer

(a)

Solution

Since the error for a point is the perpendicular distance of that point from the model (line in this case), model 1 has a larger perpendicular distance than that model 2. That is why model 1 will give more training error.

Common data for Questions 2 and 3

Statement

Consider the following linear regression model:

$$y_i|x_i = w^T x_i + \epsilon$$

where the noise ϵ follows the following distribution:

$$f_{\mathrm{E}}(\epsilon; \mu, b) \propto \exp(\frac{-|\epsilon - \mu|}{b})$$

with $\mu = 0$. b is a parameter.

Question 2

Statement

Find the log-likelihood function for the parameters w if the samples are taken from the above model.

Note: If $X \sim \mathrm{Laplace}(\mu, b)$, then $aX + c \sim \mathrm{Laplace}(a\mu + c, |a|b)$

Options

(a)

$$\log(L(w;\ x_1, x_2, \ldots x_n, y_1, y_2, \ldots, y_n)) = \sum_{i=1}^{n} \left(\frac{-w^T x_i}{b} \right)$$

(b)

$$\log(L(w;\ x_1, x_2, \dots x_n, y_1, y_2, \dots, y_n)) = \sum_{i=1}^{n} \left(\frac{-|w^T x_i|}{b} \right)$$

(c)

$$\log(L(w;\ x_1, x_2, \dots x_n, y_1, y_2, \dots, y_n)) = \sum_{i=1}^{n} \left(\frac{w^T x_i - y_i}{b} \right)$$

(d)

$$\log(L(w;\ x_1, x_2, \dots x_n, y_1, y_2, \dots, y_n)) = \sum_{i=1}^{n} \left(\frac{-|w^T x_i - y_i|}{b} \right)$$

Answer

(d)

Solution

Given that

$$y_i|x_i = w^T x_i + \epsilon$$

Where the error ϵ follows the below distribution

$$f_{\mathrm{E}}(\epsilon; \mu, b) \propto \exp(\frac{-|\epsilon - \mu|}{b})$$

Therefore,

$$f_{y_i|x_i}(y_i) \propto \exp\left(\frac{-|y_i - w^T x_i|}{b} \right)$$

For the sample $y_1, y_2, \dots, y_n$, the likelihood function is defined as (Constant terms are avoided)

$$\begin{aligned} L(w;\ x_1, x_2, \dots x_n, y_1, y_2, \dots, y_n) &= \prod_{i=1}^{n} f_{y_i|x_i}(y_i) \\ &= \prod_{i=1}^{n} \exp\left(\frac{-|y_i - w^T x_i|}{b} \right) \end{aligned}$$

Taking $\log_e$ we got

$$\begin{aligned} \log(L(w;\ x_1, x_2, \dots x_n, y_1, y_2, \dots, y_n)) &= \log\left(\prod_{i=1}^{n} \exp\left(\frac{-|y_i - w^T x_i|}{b} \right) \right) \\ &= \sum_{i=1}^{n} \left(\frac{-|w^T x_i - y_i|}{b} \right) \end{aligned}$$

Question 3

Statement

Choose the correct statement.

Options

(a)

ML estimator assuming noise following the above distribution is the same as linear regression with squared error.

(b)

ML estimator assuming noise following the above distribution is the same as linear regression with absolute error.

Answer

(b)

Solution

Let $\hat{w}$ be the ML estimate for w, then

$$\begin{aligned}\hat{w} &= \underset{w}{\operatorname{argmax}} \log(L(w;\ x_1, x_2, \ldots x_n, y_1, y_2, \ldots, y_n)) \\ &= \underset{w}{\operatorname{argmax}} \sum_{i=1}^{n} \left(\frac{-|w^T x_i - y_i|}{b} \right) \\ &= \underset{w}{\operatorname{argmin}} \sum_{i=1}^{n} \left(\frac{w^T x_i - y_i|}{b} \right)\end{aligned}$$

It implies that ML estimator assuming noise following the above distribution is the same as linear regression with absolute error.

Common data for Questions 4, 5, and 6

Consider the following dataset with one feature and corresponding label:

x_1	**Label (y)**
2	2.2
0	-0.1
-3	-2.5
1	1

Question 4

Statement

Fit the linear regression model $y = wx_1$ using squared error.

Options

(a)

$y = 1.5x_1$

(b)

$y = 2.9x_1$

(c)

$y = 0.9x_1$

(d)

$y = 1.6x_1$

Answer

(c)

Solution

The weight vector w is given as

$$w = (XX^T)^{-1}Xy$$

Here, $X = [2, 0, -3, 1]$ and $y = [2.2, -0.1, -2.5, 1]$

Doing the matrix multiplication, we get

$$w = [0.9]$$

Therefore, the fit model is given as

$$y = 0.9x_1$$

Question 5

Statement

What will be the prediction for the point $x_1 = 4$? Write your answer correct to two decimal places.

Answer

3.6 Range = [3.5, 3.8]

Solution

The model is given by

$$y = 0.9x_1$$

at $x = 4$, we have

$$y = 3.6$$

Question 6

Statement

Find the root mean squared error (RMSE) for the training dataset. Write your answer correct to two decimal places.

$$\text{RMSE} = \left(\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \right)^{1/2}$$

Answer

0.23 Range = [0.1, 0.2]

Solution

x_1	Label (y)	$\hat{y} = 0.9x_1$
2	2.2	1.8
0	-0.1	0
-3	-2.5	-2.7
1	1	0.9

Therefore, the RMSE is given by

$$\begin{aligned} \text{RMSE} &= \left(\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \right)^{1/2} \\ &= \frac{1}{4} \left((2 - 1.8)^2 + (0)^2 + (-3 + 2.7)^2 + (1 - 0.9)^2 \right)^{1/2} \\ &= 0.23 \end{aligned}$$

Question 7

Statement

What are the possible issues with the gradient descent?

Options

(a)

Gradient descent can never converge to the global minima.

(b)

If the number of training samples is large, then the gradient descent assuming constant learning will take a long time to converge because a weight update is only happening once per data cycle.

(c)

The larger your dataset, the more nuanced the gradients become, and the more time is used, and eventually, there will not be much learning.

Answer

(b), (c)

Solution

Statement (a) is false as if we initialize the weight vector such that the loss function value corresponding to the same weight is near the global minima, the gradient descent will converge to the global minima.

Statement (b) is true since it will take more time to update the weights in each iteration when the number of samples becomes large. Even the matrix multiplications such as XX^Tw become very computationally large.

Statement (c) is true since the larger the dataset, the more time is used to update the weights and the gradient descent becomes nuanced.

Question 8

Statement

Gaussian kernel regression with parameter $\sigma^2 = 1/2$ was applied to the following dataset with two features:

$$X = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix} \qquad y = [2.1, 1, 2, 1.2]^T$$

The weight vector can be written as $w = X\alpha$. The vector α is given by $[1.4, -1.4, 2, 0]^T$ which is obtained as $(K)^{-1}y$, where K is the kernel matrix. What will be the prediction for point $[1, 1]^T$?

Answer

2

Solution

The prediction is given by

$$\sum_{i=1}^{n} k(x_i, x_{\text{test}})\alpha_i \tag{1}$$

The kernel function is given by

$$\begin{aligned} k(x_i, x_j) &= \exp\left(\frac{-||x_i - x_j||^2}{2(\sigma^2)}\right) \\ &= \exp\left(-||x_i - x_j||^2\right) (\because \sigma^2 = 1/2) \end{aligned}$$

Now,

$$\begin{aligned} k(x_1, x_{\text{test}}) &= k([1,0],[1,1]) \\ &= \exp(-(0+1)) &= e^{-1} \\ k(x_2, x_{\text{test}}) &= k([0,1],[1,1]) \\ &= \exp(-(1+0)) &= e^{-1} \\ k(x_3, x_{\text{test}}) &= k([1,1],[1,1]) \\ &= \exp(-(0+0)) &= 1 \\ k(x_4, x_{\text{test}}) &= k([0,0],[1,1]) \\ &= \exp(-(1+1)) &= e^{-2} \end{aligned}$$

Putting the values in eq (1), we get

$1.4e^{-1} - 1.4e^{-1} + 1(2) + 0(e^{-2}) = 2$

Question 9

Statement

Is the following statement true or false?

The line (or hyperplane in higher dimension) that passes through the origin will incur the minimum error out of all linear functions.

Options

(a)

True

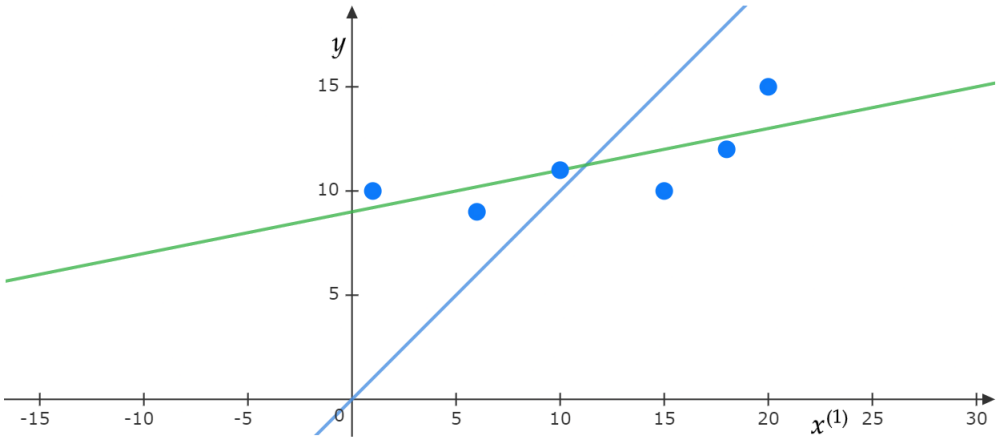
(b)

False

Answer

(b)

Consider the dataset as given in the following image:



We can see that the model that passes through the origin (blue line) incurs more loss than the model that doesn't pass through the origin (green line). Therefore, the given statement is false.

Question 10

Statement

Since the best fit line need not pass through the origin for some datasets, the model $y_i = wx_i$ may not give the best fit solution. What should be the better way to tackle this problem?

Options

(a)

mean-center the dataset.

(b)

Add a dummy feature $x_0 = 1$ in the dataset and learn the model $y_i = w^T x_i + w_0$.

Answer

(b)

Solution

If we allow our model to have an intercept on the y-axis (on the label axis), our model need not always pass through origin, we can tackle the above problem.