Graded 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 weight vector (parameter) in the linear regression model.

Question 1

Statement

An ML engineer comes up with two different models for the same dataset. The performances of these two models on the training dataset and test dataset are as follows:

- Model 1: Training error = 0.9; Test error = 0.1
- Model 2: Training error = 0.1; Test error = 10

Which model you would select?

Options

(a)

Model 1

(b)

Model 2

Answer

(a)

Solution

In model 1, the test error is very low compared to model 2 even though the training error is high in model 1. We choose model 1 as it worked well on unseen data.

Question 2

Statement

Consider a model h for a given d-dimensional training data points $\{x_1, x_2, \ldots, x_n\}$ and corresponding labels $\{y_1, y_2, \ldots, y_n\}$ as follows:

$$h : \mathbb{R}^d \to \mathbb{R}$$
$$h(x_i) = \overline{y}$$

where $\overline{y}$ is the average of all the labels. Which of the following error function will always give the zero training error for the above model?

Options

(a)

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

(b)

$\sum_{i=1}^{n}|(h(x_i)-y_i)|$

(c)

$\sum_{i=1}^{n}(h(x_i)-y_i)$

(d)

$\sum_{i=1}^{n}(h(x_i)-y_i)^3$

Answer

(c)

Solution

The sum of squared error and absolute error will give zero error only if predicted values are the same as actual values for all the examples.

But for option (3), we have

$$\begin{aligned}\sum_{i=1}^{n}(h(x_i)-y_i) &= \sum_{i=1}^{n}(\overline{y}-y_i)\\ &= n\overline{y}-\sum_{i=1}^{n}y_i\\ &= n\overline{y}-n\overline{y}=0\end{aligned}$$

This error function will give zero error for the above model.

Common data for questions 3 and 4

Consider the following dataset with one feature x_1:

x_1	**label (y)**
-1	5
0	7
1	6

Question 3

Statement

We want to fit a linear regression model of the form $y_i = w^T x_i$. Assume that the initial weight vector is $w = [2]$. What will be the weight after one iteration using the gradient descent algorithm assuming the squared loss function? Assume the learning rate is $\eta = 1$.

Answer

-4 No range is required

Solution

At iteration $t = 0$, we have $w^0 = [2]$

At $t = 1$, we have

$$w^1 = w^0 - \eta[2XX^T w^0 - 2Xy]$$

Here $w^0 = [2]$

$X = [-1, 0, -1]$

$y = [5, 7, 6]^T$

Put the values and we get

$$w^1 = [-4]$$

Question 4

Statement

If we stop the algorithm at the weight calculated in question 1, what will be the prediction for the data point $x_1 = -2$?

Answer

8 No range is required

Solution

The model is given as

$$y_i = -4x_i$$

at $x = -2$,

$y = (-4)(-2) = 8$

Question 5

Statement

Assume that w^t denotes the updated weight after the t^{th} iteration in the stochastic gradient descent. At each step, a random sample of the data points is considered for weight update. What will be the final weight w after T iterations?

Options

(a)

w^T

(b)

$w^1 + w^2 + \ldots + w^T$

(c)

$\frac{1}{T}\sum_{i=1}^{T} w^i$

(d)

any of the w^t

Answer

(c)

Solution

The final weight is given by the average of all the weights updated in all the iterations. That is why option (c) is correct.

Common data for questions 6 and 7

Kernel regression with a polynomial kernel of degree two is applied on a data set $\{X, y\}$. Let the weight vector be

$$w = X[0.3, 1.6, 4.2, -0.5, 0.9]$$

Question 6

Statement

Which data point plays the most important role in predicting the outcome for an unseen data point? Write the data point index as per matrix X assuming indices start from 1.

Answer

3, No range is required

Solution

Since w is written as $X[0.3, 1.6, 4.2, -0.5, 0.9]$, the data point which is associated with the highest weight (coefficient) will have the most importance. The third data point is associated with the highest coefficient (4.2) therefore, the third data point has the highest importance.

Question 7

What will be the prediction for the data point $[0, 0, 0, 0, 0]^T$?

Answer

6.5 No range is required

Solution

The polynomial kernel of degree 2 is given by

$$k(x_i, x_j) = (x_i^T x_j + 1)^2$$

The coefficient α vector is given as $[0.3, 1.6, 4.2, -0.5, 0.9]$.

The prediction for a point x_{test} is given by

$$\alpha_1 k(x_{\text{test}}, x_1) + \alpha_2 k(x_{\text{test}}, x_2) + \ldots + \alpha_n k(x_{\text{test}}, x_n)$$

Since, $x_{\text{test}} = [0, 0, 0, 0, 0]^T$

$$x_{\text{test}}^T x_j = [0, 0, 0, 0, 0]^T x_j = 0 \quad \forall j$$

That is

$$k(x_{\text{test}}, x_j) = 1 \quad \forall j$$

Therefore the prediction will be

$$\alpha_1 + \alpha_2 + \ldots + \alpha_n = 0.3 + 1.6 + 4.2 - 0.5 + 0.9 = 6.5$$

Question 8

Statement

If w^* be the solution to the optimization problem of the linear regression model, which of the following expression is always correct?

Options

(a)

$y^T X^T w^* = 0$

(b)

$(y - X^T w^*)^T (X^T w^*) = 0$

(c)

$(X^T w^*)(X^T w^*) = 0$

(d)

$y - X^T w^* = 0$

Answer

(b)

Solution

We know that $X^T w^*$ is the projection of labels y on the subspace spanned by the features that is $(y - X^T w^*)$will be orthogonal to $X^T w^*$. For datails, check the lecture 5.4.

Question 9

Statement

The gradient descent with a constant learning rate of $\eta = 1$ for a convex function starts oscillating around the local minima. What should be the ideal response in this case?

Options

(a)

Increase the value of η

(b)

Decrease the value of η

Answer

(b)

Solution

One possible reason of oscillation is that the weight vector jumps the local minima due to greater step size (η). That is if we decrease the value of η, the weight vector may not jump the local minima and the GD will converge to that local minima.

Question 10

Statement

Is the following statement true or false?

Error in the linear regression model is assumed to have constant variance.

Options

(a)

True

(b)

False

Answer

(a)

Solution

We make the assumption in the regression model that the error follows gaussian distribution with zero mean and a constant variance.