Practice

This document has 13 questions.

Question-1

Statement

An image is a collection of pixels. A pixel is stored as a float value and typically occupies 4 bytes of memory. Consider a dataset of 1000 images, where each image has dimensions 100×100. Approximately, how much memory does the entire dataset occupy?

Options

(a)

4 KB

(b)

4 MB

(c)

40 MB

(d)

4 GB

Answer

(c)

Solution

We require $100 \times 100 = 10000$ float values to represent one image. Since each float value occupies 4 bytes of memory, a single image occupies 40000 bytes of memory. Roughly, this corresponds to 40 KB. The entire dataset would occupy 1000×40 KB or 40 MB of memory. Here, we have used the following facts:

- 1 KB ≈ 1000 bytes
- 1 MB ≈ 1000 KB

Question-2

Statement

Consider a dataset that has 100 points that belong to $\mathbb{R}^3$. All of them are found to lie on a line that passes through the origin. We use a unit vector along the line as a representative and the coefficients with respect to it to represent the individual data-points. Compute the percentage decrease in the size of the dataset if we move to this new representation. Assume that it takes one unit of space to store one feature. Enter your answer correct to two decimal places; it should be in the range $[0, 100]$.

Answer

65.66

Range: $[65, 66]$

Solution

The size of the dataset in its original form is:

$$S_1 = 100 \times 3 = 300$$

The size of the dataset after moving to the new representation:

$$S_2 = 3 + 100 = 103$$

The percentage decrease in the size of the dataset is therefore:

$$\frac{S_2 - S_1}{S_1} \times 100 = \frac{300 - 103}{300} \times 100 \approx 65.66\%$$

Common Data for questions (3) and (4)

Statement

Consider the following dataset that has four points, all of which lie on a line:

$$S = \left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 3 \\ 4 \end{bmatrix}, \quad \begin{bmatrix} 6 \\ 8 \end{bmatrix}, \quad \begin{bmatrix} 1/5 \\ 4/15 \end{bmatrix} \right\}$$

Answer the questions that follow:

Question-3

Statement

Among the vectors given below, choose a representative that has unit length.

Options

(a)

$$\begin{bmatrix} 3 \\ 4 \end{bmatrix}$$

(b)

$$\begin{bmatrix} 1/15 \\ 4/15 \end{bmatrix}$$

(c)

$$\begin{bmatrix} 3/5 \\ 4/5 \end{bmatrix}$$

(d)

$$\begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Answer

(c)

Solution

The length of a vector $\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$ is given by:

$$||\mathbf{w}|| = \sqrt{\mathbf{w}^T\mathbf{w}} = \sqrt{w_1^2 + w_2^2}$$

We need to find that vector which has $||\mathbf{w}|| = 1$. From the options, we see that the required vector is:

$$\begin{bmatrix} 3/5 \\ 4/5 \end{bmatrix}$$

Question-4

Statement

With respect to the representative in the previous question, compute the coefficients for these four points. The i^{th} element from the left in each option is the coefficient for the i^{th} element from the left in the set S.

Options

(a)

$\{0, \quad 5, \quad 10, \quad 1/3\}$

(b)

$\{0, \quad 1, \quad 2, \quad 1/3\}$

(c)

$\{0, \quad 5, \quad 10, \quad 3\}$

(d)

$\{0, \quad 1, \quad 10, \quad 1/3\}$

Answer

(a)

Solution

The representative and the dataset are given below:

$$\mathbf{w} = \begin{bmatrix} 3/5 \\ 4/5 \end{bmatrix}, \quad S = \left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 3 \\ 4 \end{bmatrix}, \quad \begin{bmatrix} 6 \\ 8 \end{bmatrix}, \quad \begin{bmatrix} 1/5 \\ 4/15 \end{bmatrix} \right\}$$

The coefficient of a point $\mathbf{x}$ with respect to $\mathbf{w}$ is:

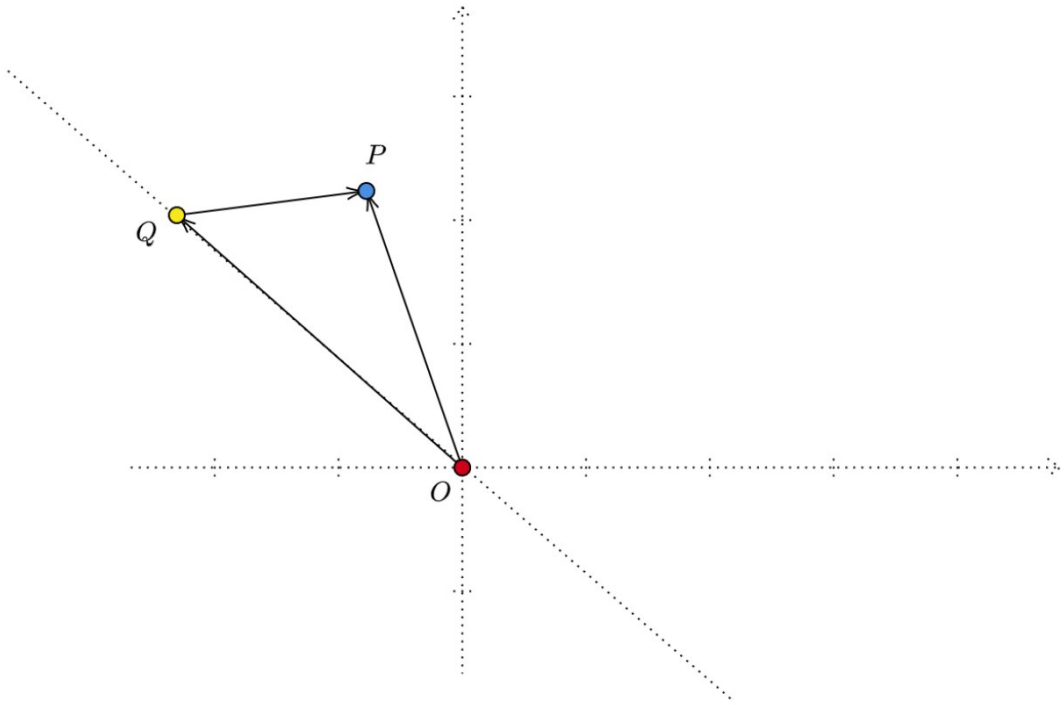
$$\mathbf{x}^T \mathbf{w}$$

Note that this equation holds only if $||\mathbf{w}|| = 1$. What will change if $||\mathbf{w}|| \neq 1$? Think about this.
The coefficients are therefore:

$$\{0, \quad 5, \quad 10, \quad 1/3\}$$

Common Data for questions (5) to (7)

Consider the following image. P is a point in 2D space. Q is a proxy for this point on a line passing through the origin. The image is drawn to scale.



Answer the questions that follow:

Question-5

Statement

Which of the following is the error vector?

Options

(a)

$\overrightarrow{OP}$

(b)

$\overrightarrow{OQ}$

(c)

$\overrightarrow{QP}$

Answer

(c)

Solution

Given these three vectors:

- a point
- its proxy
- the error vector

The following relationship holds:

$$\text{error-vector} = \text{point} - \text{proxy}$$

The error-vector is the difference between the original point and its proxy. Replacing the words with vector notation, we have:

$$\overrightarrow{QP} = \overrightarrow{OP} - \overrightarrow{OQ}$$

We have used the concept of vector addition which was covered in maths-2.

Question-6

Statement

Is Q the "best" representation of P on the line?

(a)

Yes

(b)

No

Answer

(b)

Solution

No, Q is not the best representation of P on the line. Geometrically, the best representation would be the one for which the error vector is perpendicular to the line. From the figure, we see that the line segment QP does not satisfy this property.

Question-7

Statement

If Q' is the "best" representation of P on the line, then which of the following statements are true? Notation: $|\overrightarrow{AB}|$ is the length of the vector $\overrightarrow{AB}$.

Options

(a)

$|\overrightarrow{QP}| < |\overrightarrow{Q'P}|$

(b)

$|\overrightarrow{Q'P}| < |\overrightarrow{QP}|$

(c)

$|\overrightarrow{QQ'}| = 0$

(d)

$|\overrightarrow{QQ'}| > 0$

Answer

(b), (d)

Solution

Computationally, the best representation would have the lowest reconstruction error. The smallest reconstruction error is achieved by the point Q', the tip of the projection of P onto the line. The reconstruction error is the square of the length of the error-vector. These are the terms $|\overrightarrow{Q'P}|^2$ and $|\overrightarrow{QP}|^2$. But we directly compare the lengths of the two error vectors. Think about why this is true.

We now have:

$$|\overrightarrow{Q'P}| < |\overrightarrow{QP}|$$

Since Q and Q' are two different points on the line, we have $\overrightarrow{QQ'} > 0$.

Question-8

Statement

Is the following statement true or false?

The projection of $\mathbf{x}$ onto $\mathbf{w}$ was derived to be $(\mathbf{x}^T\mathbf{w})\mathbf{w}$, where $\mathbf{w}$ is a unit vector. Since this derivation was done for the special case of 2D vectors, this formula is not applicable in the general case of d-dimensional vectors.

Options

(a)

True

(b)

False

Answer

(b)

Solution

This formula still holds for any two d-dimensional vectors. The geometry of 2D space is generalized to d-dimensional space. In 2D space, we can visually see what it means for the error-vector/residue to be perpendicular to the line. Though this visualization is not possible for higher dimensional spaces, the basic ideas still stand. For instance, the dot-product between two vectors in $\mathbb{R}^d$ is:

$$\mathbf{x}^T\mathbf{y} = \sum_{i=1}^{d} x_i y_i$$

Likewise, the length of a vector is given by:

$$||\mathbf{x}|| = \sqrt{\mathbf{x}^T\mathbf{x}}$$

Question-9

Statement

Consider a mean-centered dataset of n points where each point belongs to $\mathbb{R}^d$. $\mathbf{w_1}, \cdots, \mathbf{w_k}$ are the first k principal components obtained by running PCA on the dataset, where $k < d$. The following relationship is observed:

$$\mathbf{x}_i - \left[\sum_{j=1}^{k} (\mathbf{x}_i^T \mathbf{w}_j)\mathbf{w}_j\right] = 0, \quad 1 \leq i \leq n$$

Which of the following statement about the dataset is true?

Options

(a)

The dataset lies in a d-dimensional subsapce of $\mathbb{R}^n$

(b)

The dataset lies in a k-dimensional subspace of $\mathbb{R}^d$

(c)

The dataset lies in a d-dimensional subspace of $\mathbb{R}^k$

(d)

The dataset lies in a k-dimensional subspace of $\mathbb{R}^n$

Answer

(b)

Solution

First, the dataset has d features and n examples. So, it doesn't make sense to talk about $\mathbb{R}^n$ as n is the number of examples. Secondly, we note that each principal component, $\mathbf{w}_i$, is a vector in $\mathbb{R}^d$. Thirdly, we know that the k principal components are orthogonal, and hence linearly independent. It follows that $S = \text{span}\left(\{\mathbf{w}_1, \cdots, \mathbf{w}_k\}\right)$ is a k-dimensional subspace of $\mathbb{R}^d$. Finally, since each data-point in the dataset is a linear combination of these k principal components, we see that all of them should lie in S.

Question-10

Statement

In the context of PCA, given n data-points in $\mathbb{R}^d$ that are mean-centered, after estimating $\mathbf{w_1}$ in the first round, what is the mean of the residues?

Options

(a)

$\mathbf{w}_1$

(b)

$\mathbf{0}$

Answer

(b)

Solution

The mean of the residuals is the zero vector in $\mathbb{R}^d$:

$$\sum_{i=1}^{n} \mathbf{x}_i' = \sum_{i=1}^{n} \mathbf{x}_i - (\mathbf{x}_i^T \mathbf{w_1}) \mathbf{w_1}$$

$$= \mathbf{0} - \left[\left(\sum_{i=1}^{n} \mathbf{x}_i \right)^T \mathbf{w_1} \right] \mathbf{w_1}$$

$$= \mathbf{0}$$

Here, we have used the fact that $\frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i = \mathbf{0}$ as the data is mean-centered.

Question-11

Statement

Consider two ways of representing n datapoints that belong to $\mathbb{R}^d$ in the form of a matrix:

Approach-1: A matrix $\mathbf{X_1}$ of dimension $n \times d$

Approach-2: A matrix $\mathbf{X_2}$ of dimension $d \times n$

Assume that the dataset is mean-centered. Select all correct expressions for the covariance matrix.

Options

(a)

$\frac{1}{n}\mathbf{X}_1^T\mathbf{X}_1$

(b)

$\frac{1}{n}\mathbf{X}_2^T\mathbf{X}_2$

(c)

$\frac{1}{n}\mathbf{X}_1\mathbf{X}_1^T$

(d)

$\frac{1}{n}\mathbf{X}_2\mathbf{X}_2^T$

Answer

(a), (d)

Solution

Let the i^{th} data-point be $\mathbf{x}_i$. The expression for the covariance matrix is:

$$\mathbf{C} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T$$

There are two ways to arrange the n data-points. We have a $d \times n$ matrix, where each column corresponds to one data-point. This form is particularly important as we will be using this extensively in the second week of the course:

$$\mathbf{X}_2 = \begin{bmatrix} \vdots & & \vdots \\ \mathbf{x}_1 & \cdots & \mathbf{x}_n \\ \vdots & & \vdots \end{bmatrix}$$

Then, we have:

$$\mathbf{X}_2\mathbf{X}_2^T = n \cdot \mathbf{C}$$

On the other hand, we have a $n \times d$ matrix, where each row corresponds to one data-point:

$$\mathbf{X}_1 = \begin{bmatrix} \cdots & \mathbf{x}_1 & \cdots \\ & \vdots & \\ \cdots & \mathbf{x}_n & \cdots \end{bmatrix}$$

Since $\mathbf{X}_1^T = \mathbf{X}_2$, we have:

$$\mathbf{X}_1^T \mathbf{X}_1 = n \cdot \mathbf{C}$$

Question-12

Statement

Consider a mean-centered dataset obtained from the banking domain that has 100 data-points, each of which is described by 7 features. The dataset is represented as a 100×7 matrix, $\mathbf{X}$. You run PCA on this dataset and observe that the residues vanish completely after k iterations.

A little later, a domain expert makes the following observations. If $\mathbf{c}_i$ represents the i^{th}column of $\mathbf{X}$, then:

- The set of vectors $\{\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3, \mathbf{c}_4\}$ are linearly independent.
- The following relations are satisfied:
  - $\mathbf{c}_5 = \mathbf{c}_1 + \mathbf{c}_3$
  - $\mathbf{c}_6 = 2\mathbf{c}_3 - 3\mathbf{c}_4$
  - $\mathbf{c}_7 = \mathbf{c}_2 + 3\mathbf{c}_4$

What is the value of k? Assume that the dataset is already mean-centered.

Answer

4

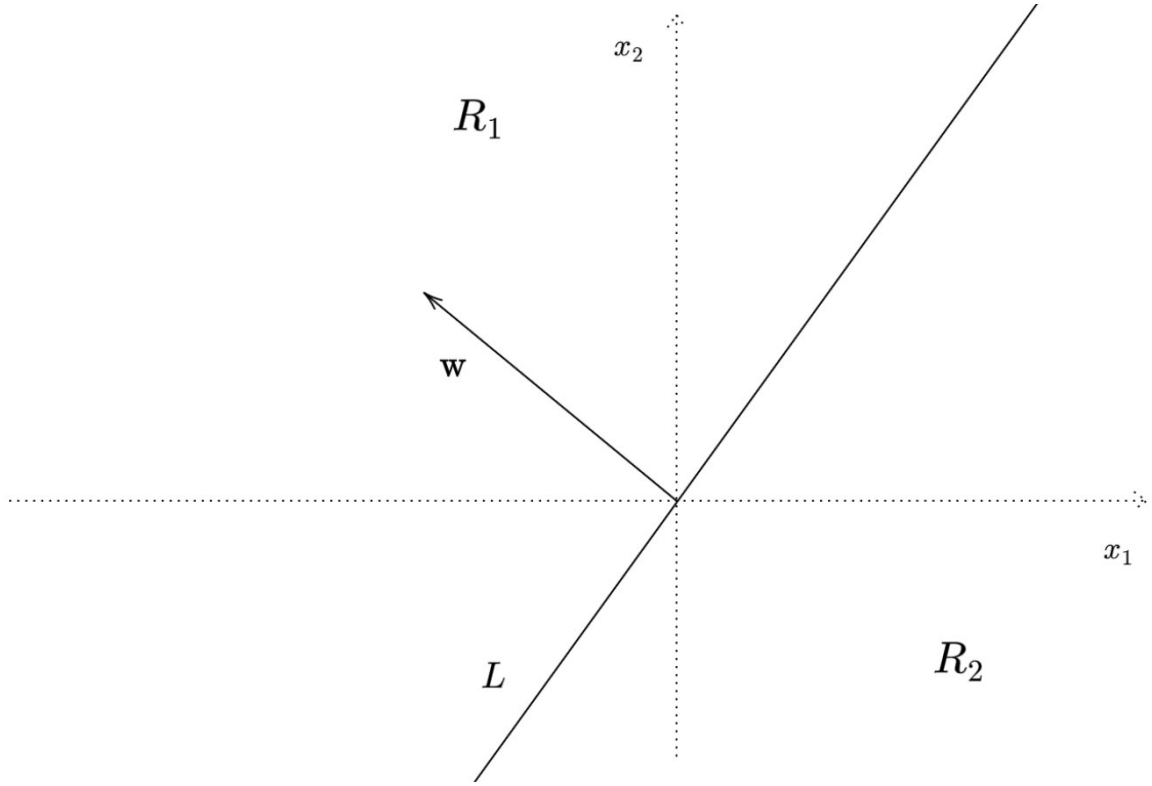
Solution

Since the last three columns are linear combinations of the first four, and since the first four columns are linearly independent, the rank of the matrix $\mathbf{X}$ is 4. This means that the rows of the matrix (the data-points) belong to a four dimensional subspace of $\mathbb{R}^7$. Intuitively, we see that PCA should terminate after four iterations and the principal components will form a basis of this subspace. For now, we shall skip the proof of this statement.

Question-13

Statement

Consider the following image:



Here, $\mathbf{w}$ is a vector and L is a line perpendicular to $\mathbf{w}$ that passes through the origin. R_1 and R_2 are two regions on either side of the line L. If $\mathbf{x}$ is an arbitrary vector in the plane, select all correct statements.

Options

(a)

$$R_1 : \mathbf{w}^T\mathbf{x} > 0$$

(b)

$$R_2 : \mathbf{w}^T\mathbf{x} > 0$$

(c)

$$R_1 : \mathbf{w}^T\mathbf{x} < 0$$

(d)

$$R_2 : \mathbf{w}^T\mathbf{x} < 0$$

(e)

$$L : \mathbf{w}^T\mathbf{x} = 0$$

Answer

(a), (d), (e)

Solution

All points (vectors) in the region R_1 make an acute angle with $\mathbf{w}$. Hence, $\mathbf{w}^T\mathbf{x} > 0$ for these points. All points (vectors) in the region R_2 make an obtuse angle with $\mathbf{w}$. Hence, $\mathbf{w}^T\mathbf{x} < 0$ for these points. All points on the line L make a right angle with $\mathbf{w}$. Hence, $\mathbf{w}^T\mathbf{x} = 0$ for these points.

Practice assignment

Question: 1

Statement

Assume that $w_k;\ k = 1, 2, \ldots d$ are d principal components corresponding to nonzero eigenvalues of the D-dimensional centered data points $x_i;\ i = 1, 2, \ldots, n$.

Statement 1: each x_i can be written as a linear combination of $w_k s$.

Statement 2: each w_k can be written as a linear combination of $x_i s$.

Options

(a)

Statement 1 is correct but statement 2 is incorrect.

(b)

Statement 1 is incorrect but statement 2 is correct.

(c)

Both statements are correct.

(d)

Both statements are incorrect.

Answer:

(c)

Solution

In the first week, we have seen that residues after d iterations become zero that is

$$\begin{aligned} & x_i - (x_i^T w_1 + x_i^T w_2 + \cdots + x_i^T w_d) = 0 \\ \Rightarrow \quad & x_i = x_i^T w_1 + x_i^T w_2 + \cdots + x_i^T w_d \end{aligned}$$

it implies that each x_i can be written as a linear combination of $w_k s$.

We know that the eigenvectors of the covariance matrix C are the principal components of the dataset and by the definition of eigenvectors, we have

$$Cw_k = \lambda_k w_k$$
$$\Rightarrow \quad \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T\right) w_k = \lambda_k w_k$$
$$\Rightarrow \quad w_k = \sum_{i=1}^{n}\left(\frac{x_i^T w_k}{n\lambda_k}\right) x_i$$

That is ach w_k can be written as a linear combination of $x_i s$.

Question: 2

Statement

A transformation mapping ϕ is defined as

$$\phi : \mathbb{R} \to \mathbb{R}^4$$
$$\phi(x) = [x^3, \sqrt{3}x^2, \sqrt{3}x, 1]^T$$

Which of the following options are the same as $\phi(x_1)^T \phi(x_2)$ for two points $x_1, x_2 \in \mathbb{R}$?

Hint: Rather than doing the calculation, try to figure out the appropriate kernel function.

(a)

$(x_1 x_2 + 1)^3$

(b)

$(x_2 x_1 + 1)^3$

(c)

$(x_1 x_2 + 1)^4$

(d)

$\phi(x_2)^T \phi(x_1)$

Answer:

(a), (b), (d)

Solution

It is easy to verify that

$$\phi(x_1)^T \phi(x_2) = x_1^3 x_2^3 + 3x_1^2 x_2^2 + 3x_1 x_2 + 1 = (x_1 x_2 + 1)^3$$

It shows that the polynomial kernel of degree three refers to the given transformation ϕ. And since the dot product is commutative, we can check that options (a), (b), and (d) refer to the same expression.

Therefore the correct answers are options (a), (b), and (d).

Question: 3

Statement

Let C be the covariance matrix of n data points in d-dimensional space. Assume that the data points are mean-centered. If $2, 5,$ and 7 are the only non-zero eigenvalues of C, what will be the non-zero eigenvalues of $X^T X$, where X is the matrix of shape (d, n) containing the data points?

Options

(a)

$2, 5, 7$

(b)

$2d, 5d, 7d$

(c)

$2n, 5n, 7n$

(d)

Can not be determined

Answer: C

Solution

The covariance matrix is defined as $\frac{1}{n} XX^T$ and the nonzero eigenvalues of $\frac{1}{n} XX^T$ are given to be $2, 5,$ and 7.

$\Rightarrow$ nonzero eigenvalues of XX^T will be $2n, 5n$ and $7n$.

Since all the nonzero eigenvalues of XX^T and $X^T X$ are the same, the nonzero eigenvalues of $X^T X$ are $2n, 5n,$ and $7n$.

Common data for Questions 4 and 5

Statement

Consider an image dataset matrix X of shape (d, n) with $d > n$. The kth principal component of the dataset can be written as $w_k = \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_n x_n$, where, the vector x_i is the ith data point. The k^{th} largest eigenvalue and the corresponding eigenvector of $X^T X$ are 4 and $[\frac{1}{\sqrt{51}}, \frac{3}{\sqrt{51}}, \frac{4}{\sqrt{51}}, \frac{5}{\sqrt{51}}]^T$, respectively.

Question 4

Statement

What will be the value of α_1?

Options

(a)

4

(b)

$\frac{1}{\sqrt{51}}$

(c)

$\frac{1}{2\sqrt{51}}$

(d)

$[\frac{1}{\sqrt{51}}, \frac{3}{\sqrt{51}}, \frac{4}{\sqrt{51}}, \frac{5}{\sqrt{51}}]^T$

(e)

$[\frac{1}{2\sqrt{51}}, \frac{3}{2\sqrt{51}}, \frac{4}{2\sqrt{51}}, \frac{5}{2\sqrt{51}}]^T$

Answer:

C

Solution

We know that the k^{th} principal component can be written as a linear combination of the data points that is

$$w_k = \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_n x_n$$

And the vector $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_n]$ can be obtained by eigen decomposition of $X^T X$ as follows:

If the k^{th} largest eigenvalue and the corresponding unit eigenvector of $X^T X$ are λ_k and β_k, respectively then

$$\alpha = \frac{\beta_k}{\sqrt{\lambda_k}}$$

Therefore, by the given information, we can say that

$$[\alpha_1, \alpha_2, \ldots, \alpha_n]^T = \frac{1}{\sqrt{4}}[\frac{1}{\sqrt{51}}, \frac{3}{\sqrt{51}}, \frac{4}{\sqrt{51}}, \frac{5}{\sqrt{51}}]^T$$
$$\Rightarrow \alpha_1 = \frac{1}{2\sqrt{51}}$$

Question 5

Statement

What will be the k^{th} largest eigenvalue of the covariance matrix $\frac{1}{4}XX^T$?

Note that $n = 4$ as the length of the eigenvector of X^TX is 4.

Answer:

1

Solution

The k^{th} largest eigenvalue of $X^TX = 4$

The nonzero eigenvalues of X^TX and XX^T are the same.

$\Rightarrow$ The k^{th} largest eigenvalue of $XX^T = 4$

$\Rightarrow$ The k^{th} largest eigenvalue of $\frac{1}{4}XX^T = 1$

Question: 6

Statement

A function k is defined as

$$k : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$$
$$k([x_1, x_2]^T, [y_1, y_2]^T) = x_1^2y_1^2 + x_2^2y_2^2$$

Is k a valid kernel?

Hint: Try to find out the appropriate ϕ.

Options

(a)

Yes

(b)

No

Answer:

A

Solution

If we can find an appropriate transformation mapping ϕ such that

$$k(x_1, x_2) = \phi(x_1)^T\phi(x_2)$$

Then we can conclude that k is a valid kernel.

The given kernel is

$$
\begin{aligned}
k : \mathbb{R}^2 \times \mathbb{R}^2 &\to \mathbb{R} \\
k([x_1, x_2]^T, [y_1, y_2]^T) &= x_1^2 y_1^2 + x_2^2 y_2^2 \\
&= [x_1^2, x_2^2]^T [y_1^2, y_2^2]
\end{aligned}
$$

If we define a function $\phi : \mathbb{R}^2 \to \mathbb{R}^2$ such that

$$\phi([x_1, x_2]) = [x_1^2, x_2^2]$$

then we can say that

$$k(x_1, x_2) = \phi(x_1)^T \phi(x_2)$$

It implies that k is a valid kernel.

Question: 7

Statement

A dataset of 1000 second-hand cars has four features: kilometers driven (x_1), mileage (x_2), the present price of the car (x_3), and the selling price (x_4). The selling price seems to have the following relationship (approximate) with the other three features.

$$x_4 = x_1^2 x_3 + 2x_2$$

If we want to project the dataset into a lower dimensional space, which of the following task would be most appropriate?

Options

(a)

Standard PCA

(b)

Kernel PCA with a polynomial kernel of degree 2

(c)

Kernel PCA with a polynomial kernel of degree 3

(d)

Kernel PCA with a polynomial kernel of degree 4

Answer

(c)

Solution

Notice that the features are not linearly related. The feature x_4 is related to other features and the relationship is cubic in nature.

That is why if we transform the dataset into a higher dimension using the degree three polynomial, then the dataset may live in a linear subspace.

Therefore, kernel PCA with a polynomial kernel of degree 3 will be an appropriate task.

Question 8

Statement

Abhishek runs a kernel PCA on a dataset containing n examples with d features. Which of the following strategy he should follow to center the data points?

strategy 1: First center the dataset using the mean and then apply the kernel.

Strategy 2: First apply the kernel and then center the matrix.

Options

(a)

Strategy 1

(b)

Strategy 2

(c)

Both strategies are the same

Answer

(b)

Solution

Applying transformation on the centered dataset is not mandatory to give the centered transformed dataset. For example, consider a centered dataset containing four points in two dimensions.

$$X = \begin{bmatrix} 1 & -1 & 3 & -3 \\ 2 & -2 & 4 & -4 \end{bmatrix}$$

Now apply the transformation $\phi : \mathbb{R}^2 \to \mathbb{R}^2$ such that

$$\phi([x_1, x_2]) = [x_1^2, x_2^2]$$

The transformed dataset will be

$$X' = \begin{bmatrix} 1 & 1 & 9 & 9 \\ 4 & 4 & 16 & 16 \end{bmatrix}$$

Clearly, this dataset is not centered. Therefore, strategy 2 is best suited.

Question 9

Statement

A dataset containing 1000 points in 3-dimensional space is run through the kernel PCA with the polynomial kernel of degree p. If the transformed dataset lives in a ten-dimensional space, what will be the value of p?

Answer

2 (No range required)

Solution

Let x_1, x_2 and x_3 are three features. If we use $p = 2$ the features will be

$1, x_1, x_2, x_3, x_1^2, x_2^2, x_3^2, x_1x_2, x_1x_3, x_2x_3$ that is the transformed space will be ten-dimensional.

Question 10

Statement

A dataset containing 1000 examples in 10-dimensional space is projected into other dimension space using kernel PCA with the following kernel.

$$k(x_1, x_2) = \exp\left(\frac{-||x_1 - x_2||^2}{4}\right)$$

What will be the dimension of the projected dataset?

Options

(a)

10

(b)

40

(c)

∞

(d)

Can not be determined

Answer

(c)

Solution

The given kernel is the gaussian kernel, which will lead to infinite dimension.

Practice

This document has 10 questions.

Question-1

Statement

Which of the following sequences is correct for K-Means algorithm?

1. Assign each data point to the nearest cluster centres.
2. Re-assign each point to nearest cluster centres.
3. Assign cluster centres randomly.
4. Re-compute cluster centres.
5. Specify the number of clusters.

Options

(a)

3, 5, 1, 4, 2

(b)

5, 3, 1, 2, 4

(c)

5, 3, 1, 4, 2

(d)

3, 5, 2, 4, 1

(e)

None of these

Answer

(c)

Solution

The correct sequence of steps used in k-means algorithm is:

- Specify the number of clusters.
- Assign cluster centres randomly.
- Assign each data point to the nearest cluster centres.
- Re-compute cluster centres.
- Re-assign each point to nearest cluster centres.

Question 2

Statement

If $F(z_1^t, z_2^t, \ldots, z_n^t)$ represents the value of objective function in iteration t of Lloyd's algorithm, then which of the following is true?

Options

(a)

$F(z_1^{t+1}, z_2^{t+1}, \ldots, z_n^{t+1}) > F(z_1^t, z_2^t, \ldots, z_n^t)$

(b)

$F(z_1^{t+1}, z_2^{t+1}, \ldots, z_n^{t+1}) < F(z_1^t, z_2^t, \ldots, z_n^t)$

(c)

$F(z_1^{t+1}, z_2^{t+1}, \ldots, z_n^{t+1}) = F(z_1^t, z_2^t, \ldots, z_n^t)$

Answer

(b)

Solution

In Lloyd's algorithm, in each iteration, the data points change their cluster only if they find a cluster center which is closer to them as compared to their existing cluster's center. Therefore, every re-assignment results in the reduction of the objective function value, which is represented by $F(z_1^t, z_2^t, \ldots, z_n^t)$for iteration t.

Question-3

Statement

If μ_1 and μ_2 are means of two clusters in k-means, then the boundary between the two clusters will be

Options

(a)

Perpendicular to the line joining μ_1 and μ_2 and at the point $\frac{(\mu_1 + \mu_2)^2}{2}$

(b)

Parallel to the line joining μ_1 and μ_2 and at the point $\frac{(\mu_1 + \mu_2)^2}{2}$

(c)

Perpendicular to the line joining μ_1 and μ_2 and at the point $\frac{\mu_1 + \mu_2}{2}$

(d)

Parallel to the line joining μ_1 and μ_2 and at the point $\frac{\mu_1 + \mu_2}{2}$

Answer

(c)

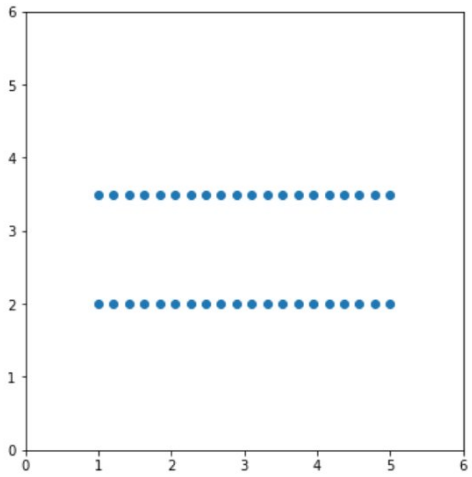
Solution

The clusters in k-means are separated by half-spaces. These half-spaces are formed by the perpendicular bisector of the line joining the clusters' centres. Therefore, if μ_1 and μ_2 are the centres of two clusters, the boundary between these clusters will be perpendicular to the line joining μ_1 and μ_2 and at the midpoint of this line, which is $\frac{\mu_1 + \mu_2}{2}$

Question-4

Statement

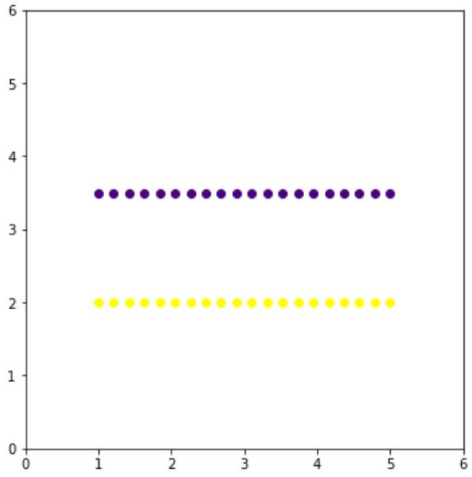
Consider the following data points:



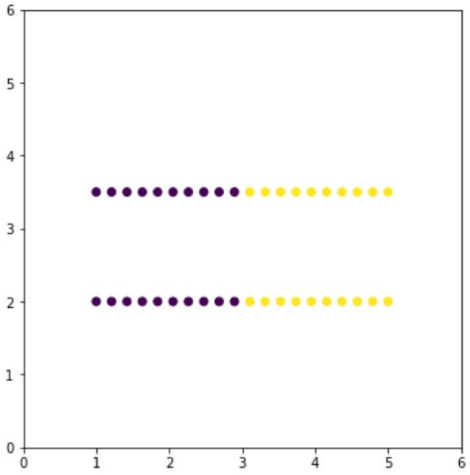
Assume K-means is run on these points with k = 2. Which of the following are expected to be the clusters formed out of K-means?

Options

(a)



(b)



(c)

Depends on cluster center initializations and the distance between the two lines.

Answer

(c)

Solution

The kind of clusters obtained will indeed depend on the cluster initializations. In the specific case of the mid-points of each line being chosen as the initial clusters, the clusters obtained will be of the type of Option (a). In all other cases, the clusters will be of the option (b) type.

Another point to note is that the clusters obtained will depend on the distance between the lines as well. In the case where the distance between the two lines is much larger than the gap between the points on the lines, the cluster center initializations will not matter and we will get clusters of type (a).

Question-5

Statement

In the initalization step of k-means++, the squared distances from the closest mean for 5 points x_1, x_2, x_3, x_4, x_5 are: 25, 67, 89, 24, 56. In this context, which of the following is true?

Options

(a)

Any point out of of x_1, x_2, x_3, x_4, x_5 may be chosen uniformly at random as next mean.

(b)

Certainly x_3 will be chosen as its distance from closest mean is largest.

(c)

x_3 will be chosen with the highest probability, but we are not sure whether this point will definitely be chosen.

Answer

(c)

Solution

K-means++ performs a smart initialization of cluster centers. The first cluster center is chosen uniformly at random out of all data points. To choose the next cluster center, the squared distances of all the remaining data points from the first cluster center are computed. These squared distances become the scores for these data points, which are further normalized and are treated as probabilities for being chosen as the next data point.

In the given question, although the score of x_3 is the maximum, it will result into the highest probability for x_3 to be chosen as the next cluster center. However, since it is probabilistic, we can not guarantee that x_3 will certainly be chosen as the next cluster center.

Question-6

Statement

With respect to Lloyd's algorithm, choose the correct statements:

Options

(a)

The partition configurations can not repeat themselves.

(b)

After doing the reassignments, we might get the same partition configuration again.

(c)

Objective function after making the re-assignments strictly reduces.

(d)

Objective function after making the re-assignments may increase.

(e)

Change of value of objective function indicates that the partition configuration has changed.

(f)

For partitioning n data points across k partitions, Lloyd's algorithm takes k^n iterations to converge.

Answer

(a), (c), (e)

Solution

(a) - (e) In Lloyd's algorithm, in each iteration, re-assignments happen only for those data points, which find a mean that is closer to them than their own mean. Hence, the value of the objective function strictly reduces, resulting in a new partition. Since the value of objective function can not be same for any two partitions, this means that the partitions can not repeat themselves.

(f) k^n is only the upper limit of number of possible partitions for k clusters and n data points.

Question-7

Statement

For 1000 data points, out of k = 1, 10 and 100, which value of k is likely to result in the maximum value of the objective function?

Options

(a)

1

(b)

10

(c)

100

(d)

Insufficient information. Depends on data.

Answer

(a)

Solution

If all data points are in the same cluster (i. e., $k = 1$), the value of the objective function will be high, as there will be only one mean, and every point's distance will be measured from this one mean. As the value of k is increased, due to the presence of more means, the distances of the data points from these means will reduce.

Therefore, $k = 1$ will result in the maximum value of objective function.

Question-8

Statement

For 100 data points, if k = 100, what will be the value of the objective function?

Options

(a)

100

(b)

0

(c)

100*100

Answer

(b)

Solution

For 100 data points, if $k = 100$, this means that every point is in their own cluster. In this case, since the point itself will represent the mean in each cluster, the distance of each point from its mean will be zero, resulting in zero value of the objective function.

Question-9

Statement

Choose the correct statements:

Options

(a)

In k-means algorithm, all cluster initializations lead to the same result.

(b)

One initialization might get stuck in local minima, while another may lead to global minima.

(c)

One initialization may converge while another may not.

(d)

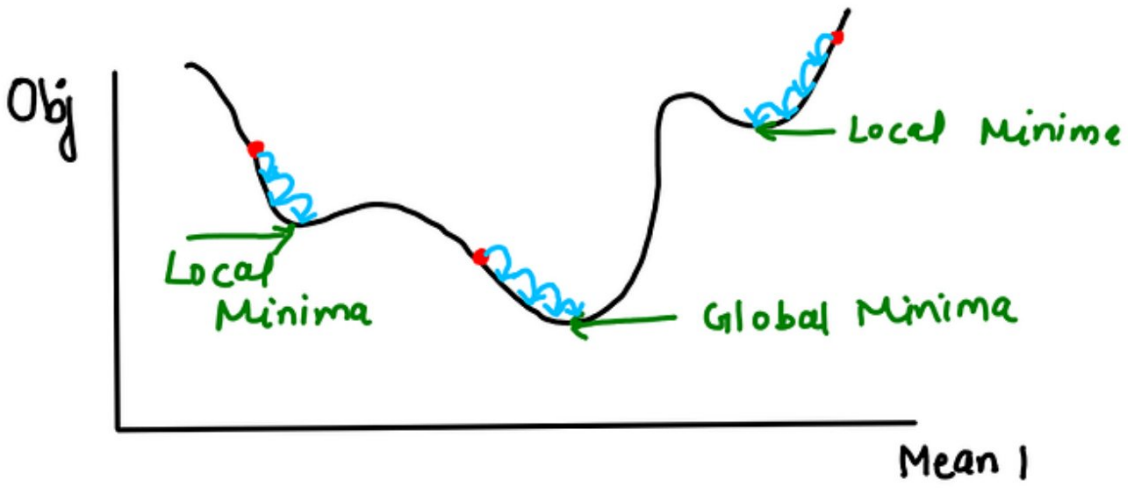
The initialization of cluster centres may affect the number of iterations K-means takes to converge.

Answer

(b), (d)

Solution

The following image shows how different initial cluster means may result into k-means converging in either local minima or global minima. Further, depending on the initial cluster means, the number of iterations required for K-means to converge might vary.



Question-10

Statement

Outliers are data points that deviate significantly from the rest of data points. Knowing the way Lloyd's algorithm works, do you think it is sensitive to outliers?

Options

(a)

Yes

(b)

No,

Answer

(a)

Solution

Since K-means is based on computing means and euclidean distances, the presence of outliers may affect the inherent clustering present in the 'non-outliers' .

Practice

This document has 9 questions.

Question-1

Statement

Consider a dataset that has only 100 points, out of which 20 points have the value 1, 50 have value 2 and 30 have value 3. We use a categorical distribution to model this data. The parameters of the distribution are:

$$p = P(x = 1), \quad q = P(x = 2), \quad r = P(x = 3)$$

If the distribution seems unfamiliar to you, think about an imaginary dice with three faces. What is the likelihood function for this data under this distribution?

Options

(a)

$$p^{30} \cdot q^{50} \cdot r^{20}$$

(b)

$$(p + q + r)^{100}$$

(c)

$$p^{20} \cdot q^{50} \cdot r^{30}$$

Answer

(c)

Solution

Let us use a, b, c to refer to the number of ones, twos and threes respectively. By the i.i.d assumption, we have:

$$L(p, q, r; D) = p^a \cdot q^b \cdot r^c$$

Question-2

Statement

What is the value of $p + q + r$? Enter your answer correct to two decimal places.

Answer

1.0

Range: $[0.99, 1.01]$

Solution

Since the sample space is $\{1, 2, 3\}$, the probabilities should sum to 1.

Question-3

Statement

What is the maximum likelihood estimate of p? Enter your answer correct to two decimal places.

Answer

0.2

Range: $[0.19, 0.21]$

Solution

The log-likelihood is:

$$l(a, b, c; D) = a \log p + b \log q + c \log r$$

If we want to maximize this likelihood, we can't just differentiate the function and set it to zero, as there is a constraint of $p + q + r = 1$ involved. One way to get around this is to substitute $r = 1 - p - q$ to get an unconstrained problem in two variables (p, q):

$$l(a, b, c; D) = a \log p + b \log q + c \log(1 - p - q)$$

We can now compute the partial derivatives with respect to p and q and set them to zero. A fair amount of algebra will convince us that:

$$\hat{p} = \frac{a}{a + b + c}, \quad \hat{q} = \frac{b}{a + b + c}, \quad \hat{r} = \frac{c}{a + b + c}$$

An interesting insight that this equation gives us is the case when $c = 0$. This reduces to the MLE for a Bernoulli random variable. We can also see how this equation could be extended to the case of a categorical distribution that has a support whose cardinality is k. This is left as an exercise to the learners.

Question-4

Statement

Consider a dataset of n data-points, $D = \{x_1, \cdots, x_n\}$. If we assume these points to have been generated from a Gaussian distribution, $\mathcal{N}(\mu, \sigma^2)$, what is the expression for the log-likelihood after removing constant terms?

(1) Constant terms are those that don't depend on either μ or σ^2

(2) $\log$ always means $\log_e$ unless otherwise specified.

Options

(a)

$$\frac{1}{2\sigma^2} \cdot \sum_{i=1}^{n} (x_i - \mu)^2$$

(b)

$$-\log \sigma - \frac{1}{2\sigma^2} \cdot \sum_{i=1}^{n} (x_i - \mu)^2$$

(c)

$$-n \log \sigma - \frac{1}{2\sigma^2} \cdot \sum_{i=1}^{n} (x_i - \mu)^2$$

(d)

$$n \log \sigma + \frac{1}{2\sigma^2} \cdot \sum_{i=1}^{n} (x_i - \mu)^2$$

Answer

(c)

Solution

First, we compute the likelihood. Using the i.i.d assumption:

$$L(\mu, \sigma^2; D) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[\frac{-(x_i - \mu)^2}{2\sigma^2}\right]$$

Next, the log-likelihood:

$$l(\mu, \sigma^2; D) = \sum_{i=1}^{n} \left[-\log(\sqrt{2\pi}\sigma) - \frac{(x_i - \mu)^2}{2\sigma^2}\right]$$

The first term inside the summation is independent of the x_is, hence it can be taken out after scaling it by a factor of n:

$$l(\mu, \sigma^2; D) = -n \log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2} \cdot \sum_{i=1}^{n} (x_i - \mu)^2$$

After removing the constants from the first term, we get:

$$l(\mu, \sigma^2; D) = -n \log \sigma - \frac{1}{2\sigma^2} \cdot \sum_{i=1}^{n} (x_i - \mu)^2$$

Question-5

Statement

Consider a dataset of heights of 300 individuals. The first 100 are drawn from the active pool of basketball players in the NBA. The next 100 are drawn from the list of chess grand masters. The last 100 are drawn randomly from the city of Chennai. All 300 individuals are in the age-group of 20 to 25. If we use a GMM to understand this data, what is a good choice of K, the number of mixtures?

Answer

2

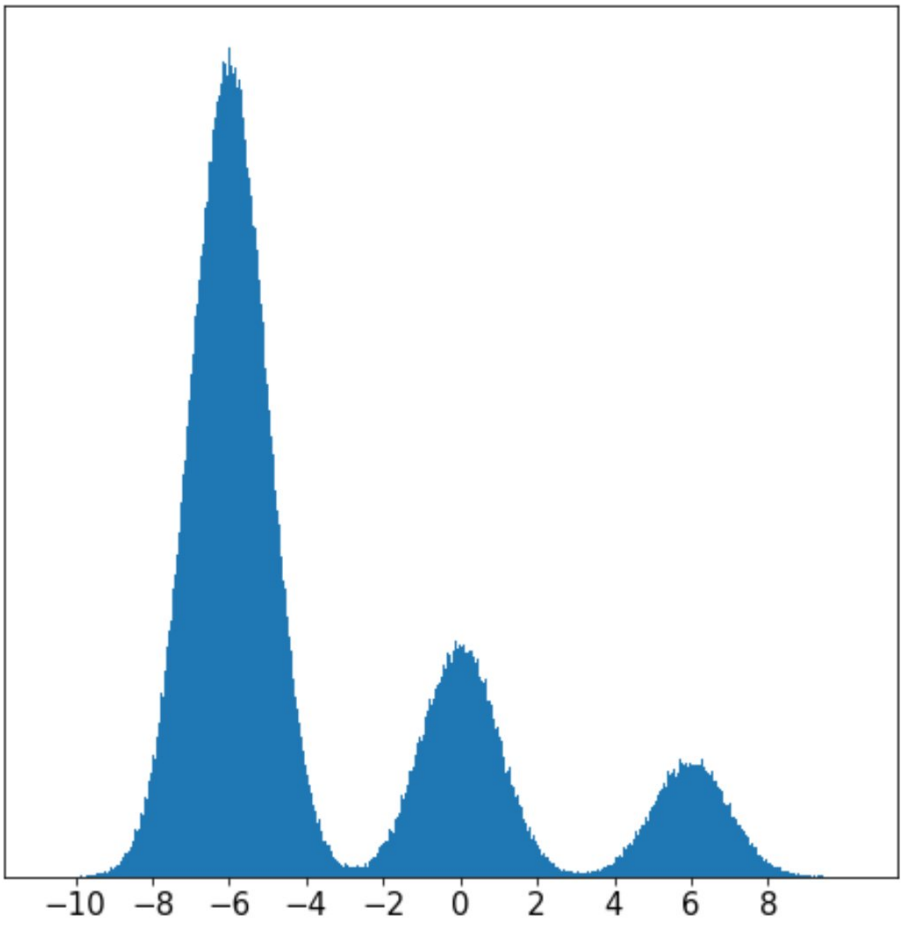
Solution

Though there are three different classes of people, as far as heights are concerned, basketball players certainly are in a different zone. Height is not correlated with chess. Hence we can reason that the average height of a chess player will not be too different from the average height of someone from Chennai. So, 2 mixtures would be sufficient.

Common Data for questions (6) to (8)

Statement

Consider the histogram of one million points sampled from a GMM with three mixtures as shown in the figure below. The mixtures are labeled from left to right as 1, 2 and 3. The mean for each mixture is one of the ticks displayed on the x-axis. All the mixtures have unit variance:



Question-6

Statement

What is the mean of mixture-3? Note that the mean is an integer here.

Answer

6

Solution

Visual inspection

Question-7

Statement

Which of the following could be the values of π_1, π_2 and π_3?

Options

(a)

$\pi_1 = \pi_2 = \pi_3 = \frac{1}{3}$

(b)

$\pi_1 = 0.4, \pi_2 = 0.1, \pi_3 = 0.4$

(c)

$\pi_1 = 0.7, \pi_2 = 0.2, \pi_1 = 0.1$

(d)

$\pi_1 = 0.4, \pi_2 = 0.4, \pi_3 = 0.1$

Answer

(c)

Solution

The heights of the mixtures give us an idea of their importance.

Question-8

Statement

If the point -3 is observed, what is the probability that it has come from mixture-2? Use the values of π_1, π_2, π_3 obtained from the previous question. Enter your answer correct to two decimal places.

Answer

0.22

Solution

We need to compute $P(z = 2 \mid x = -3)$. Using Bayes' rule:

$$P(z = 2 \mid x = -3) = \frac{P(z = 2) \cdot f(x = -3 \mid z = 2)}{f(x = -3)}$$

We have:

- $\mu_1 = -6, \mu_2 = 0, \mu_3 = 6.$
- $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 1$
- $\pi_1 = 0.7, \pi_2 = 0.2, \pi_3 = 0.1$

$$f(x \mid z = k) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left[\frac{-(x - \mu_k)^2}{2\sigma_k^2}\right]$$

We now have to compute the each of these quantities.

Question-9

Statement

Assume that you are given a set of one 10000 data-points in $\mathbb{R}$. You fit a GMM with $K = 2$ for this dataset using the EM algorithm to estimate the parameters. The EM algorithm was initialized as follows:

(1) $\mu_1 = -1, \mu_2 = 1$

(2) $\pi_1 = \pi_2 = 0.5$

(3) $\sigma_1^2 = \sigma_2^2 = 1$

The estimated means are $\hat{\mu}_1$ and $\hat{\mu}_2$ for the the two mixtures. A little while later, a domain expert comes and tells you that the dataset given to you was actually sampled from a Gaussian with mean 0 and variance 1. Which of the following options is true? Code the EM algorithm and observe what happens.

Options

(a)

$\hat{\mu}_1$ is very close to $\hat{\mu}_2$ but both are not close to 0

(b)

$\hat{\mu}_1$ is not close to $\hat{\mu}_2$ and neither of them is close to 0

(c)

$\hat{\mu}_1$ is very close to $\hat{\mu}_2$ and both are close to 0

Answer

(b)

Solution

The points that are in the interval around the mean, somewhere around $[-0.7, 0.7]$, form some sort of a barrier. The mixture on the left is unable to advance beyond a certain point to the right. Likewise, the mixture on the right is unable to advance beyond a certain point to the left. This is observed for initializations of the means that are significantly far away from the true mean. This problem will be discussed during the programming session.

Practice Assignment

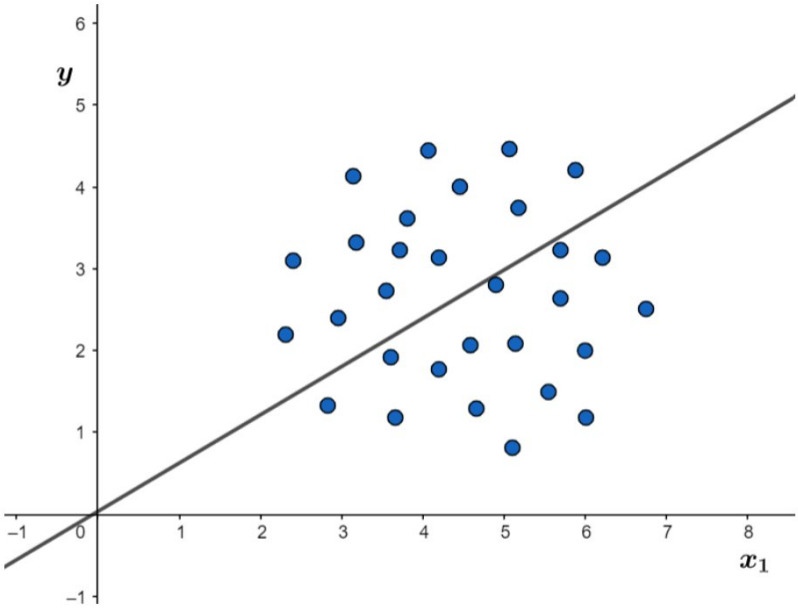
Note:

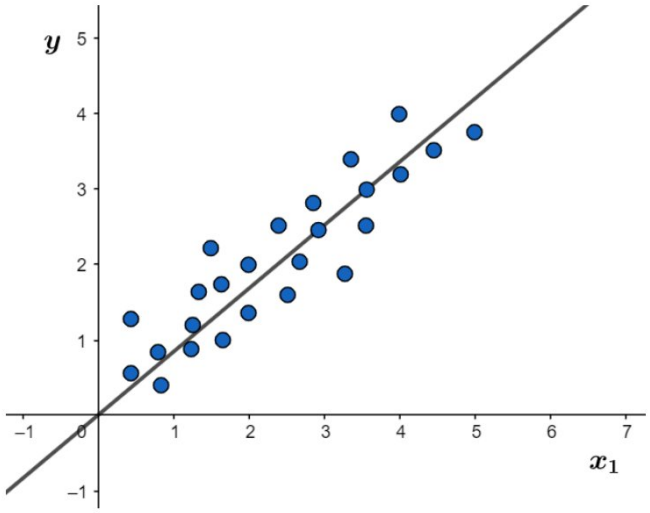
1. In the following assignment, X denotes the data matrix of shape (d, n) where d and n are the number of features and samples, respectively.
2. x_i denotes the i^{th} sample and y_i denotes the corresponding label.
3. w denotes the weights (parameter) in the linear regression model.

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

$$\begin{aligned} \text{RMSE} &= \left(\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \right)^{1/2} \\ &= \left(\frac{1}{4} ((2.2 - 1.8)^2 + (-0.1)^2 + (-2.5 + 2.7)^2 + (1 - 0.9)^2) \right)^{1/2} \\ &= 0.23 \end{aligned}$$

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

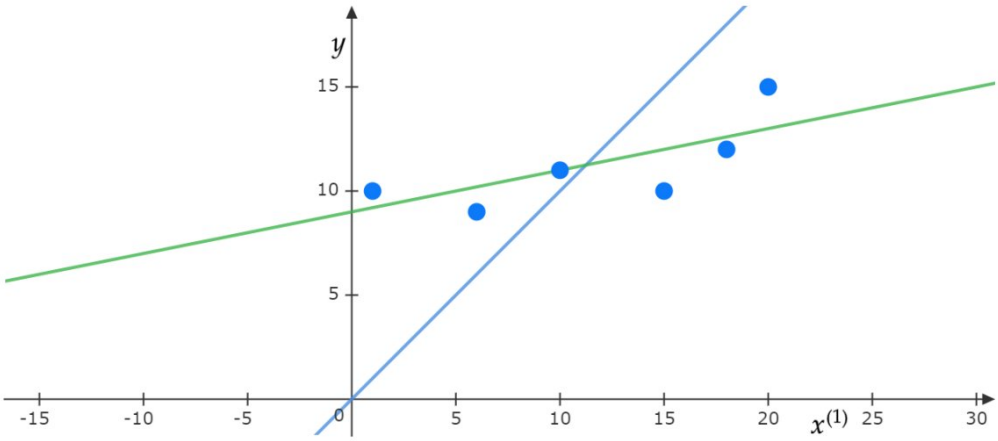
(b)

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

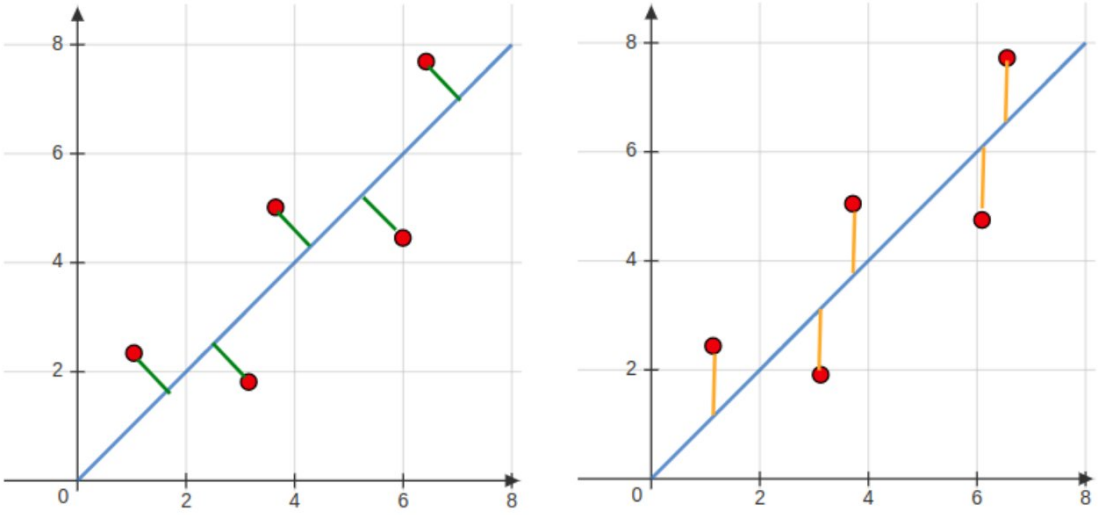
If we allow our model to have an intercept on the y-axis (on the label axis), our model need not always pass through origin, we can tackle the above problem.

Practice

Question-1

Statement

Consider the following image:



The images shows two types of lines: Green and Orange. Which of these lines are used for calculating the residuals in linear regression?

Options

(a)

Green

(b)

Orange

(c)

Does not matter; Any of these may be used.

(d)

None of these

Answer

(b)

Solution

In linear regression, the task is to predict the y-axis value given the x-axis value (for a one variable scenario). The regression line tends to fit the given data points as best as possible, such that the residual error may be minimized. Here the residual error is the distance between the actual y-value and the y-value that is predicted by the regression line. Hence, this distance line will be a vertical line. Hence the orange lines will represent these residuals.

The green lines on the other hand, represent the errors in PCA, as in PCA, the principal components tend to fit the data points such that the variance in the data is captured. The errors in PCA, thus are perpendicular lines.

Question-2

Statement

Assume that the eigenvalues of a 3×3 matrix A are 2, 3, 1. What will be the eigenvalues of the matrix $A + 5I$ where I is a 3×3 identity matrix.

Options

(a)

2, 3, 1

(b)

7, 8, 6

(c)

10, 15, 5

(d)

6, 6, 6

Answer

(b)

Solution

If a, b, c are eigenvalues of a matrix A, then then the eigenvalues of the matrix $A + \alpha I$ will be $a + \alpha, b + \alpha$ and $c + \alpha$.

Question-3

Statement

What will be the trace of $(XX^T + \lambda I)^{-1}$?

Note: λ_i are the eigenvalues of XX^T

Options

(a)

$\sum_{i=1}^{d} \lambda_i + \lambda$

(b)

$\sum_{i=1}^{d} 1 + \lambda_i$

(c)

$\sum_{i=1}^{d} \frac{1}{1 + \lambda_i}$

(d)

$\sum_{i=1}^{d} \frac{1}{\lambda_i + \lambda}$

Answer

(d)

Solution

The trace of a matrix is the sum of its eigenvalues.

If λ_i's are the eigenvalues of XX^T, then $\lambda_i + \lambda$ will be the eigenvalues of $XX^T + \lambda I$.

If a, b are eigenvalues of a matrix M, then 1/a and 1/b will be the eigenvalues of the matrix M^{-1}.

So, the eigenvalues of $(XX^T + \lambda I)^{-1}$ will be $\frac{1}{\lambda_i + \lambda}$.

Hence the trace will be $\sum_{i=1}^{d} \frac{1}{\lambda_i + \lambda}$

Question-4

Statement

Assume that as part of 3-fold cross-validation, the data was divided into parts f_1, f_2, f_3. Subsequently, cross-validation was performed. In this context, which of the following is True?

Options

(a)

In each iteration of cross-validation, one of f_1, f_2, f_3 will be used for training, and the remaining two parts will be used for testing.

(b)

In each iteration of cross-validation, two of f_1, f_2, f_3 will be used for training, and the remaining one part will be used for validation.

(c)

In each iteration of cross-validation, two of f_1, f_2, f_3 will be used for training, and the remaining one part will be used for testing.

(d)

In each iteration of cross-validation, one of f_1, f_2, f_3 will be used for training, and the remaining two parts will be used for validation.

Answer

(b)

Solution

In k-fold cross-validation, the training data is split into k parts.

K-fold cross validation involves k iterations.

In each iteration, one of the k parts is used for validation and the remaining k-1 parts are used for training.

Hence, in k-fold cross validation, one part will be used for validation and remaining two will be used for training.

Question-5

Statement

Assume that we obtain the weight vector $w = [4, 6, 8, 10]$ when λ= 3 is used in Ridge Regression.

If λ is increased to 6, which of the following is most likely to be the updated weight vector?

Options

(a)

$w = [7, 9, 11, 13]$

(b)

$w = [1, 3, 6, 9]$

(c)

$w = [10, 12, 13, 15]$

Answer

(b)

Solution

In regularization, as the value of λ (which represents penalty for higher weights) is increased, the weights are further reduced.

Therefore, in the given question, as the value of λ is increased from 3 to 6, the penalty for having higher weights will increase, thus shrinking the weights.

Question-6

Statement

The mean squared error of $\hat{w}_{ML}$ could come out to be large if

Options

(a)

Trace of XX^T is large.

(b)

Trace of $(XX^T)^{-1}$ is large.

(c)

σ^2 is large.

(d)

λ_i's are large.

Answer

(b), (c)

Solution

MSE = $\sigma^2 trace(XX^T)^{-1}$

Hence, MSE will come out to be large if σ^2 is large or trace of $(XX^T)^{-1}$ is large.

Question-7

Statement

Cross validation may be useful as it may help in

Options

(a)

Reducing the trace of the inverse of covariance matrix.

(B)

Increasing the trace of the inverse of covariance matrix.

Answer

(a)

Solution

The expression for w^* is given by $(XX^T)^{-1}Xy$

Cross validation results into a weight vector $(XX^T + \lambda I)^{-1}Xy$

This in turn reduces the trace of the inverse of covariance matrix, thus reducing the MSE.

Question-8

Statement

Assume that Lasso is applied to a data set containing only one feature. If λ is increased to a high value, which of the following is correct?

Options

(a)

The regression line may become horizontal.

(b)

The regression line may become horizontal.

(c)

The regression line will remain the same.

Answer

(a)

Solution

In lasso, as the value of λ is increased, due to a very high penalty, the coefficient of the only feature becomes zero.

The coefficient represents the slope. And when slope = 0, it results into a horizontal line.

Question-9

Statement

In Ridge Regression, if λ tends to zero, the solution approaches to

Options

(a)

Zero

(b)

One

(c)

Linear Regression

(d)

Lasso Regression

(e)

Infinity

Answer

(c)

Solution

The weight vector expression for linear regression is $(XX^T)^{-1}Xy$

The weight vector expression for ridge regression is $(XX^T + \lambda I)^{-1}Xy$

When λ tends to zero, the solution of ridge is same as that of linear regression.

Practice

This document has 12 questions.

Question-1

Statement

You are given a training dataset of 100 data-points for a classification task. You wish to use a k-NN classifier, with $k = 10$. $d(\mathbf{x_i}, \mathbf{x_j})$ is a function that returns the Euclidean distance between two points $\mathbf{x_i}$ and $\mathbf{x_j}$. Calling d an a pair of points corresponds to a single distance computation. If you want to predict the label of a new test point, how many distances would you have to compute?

Answer

100

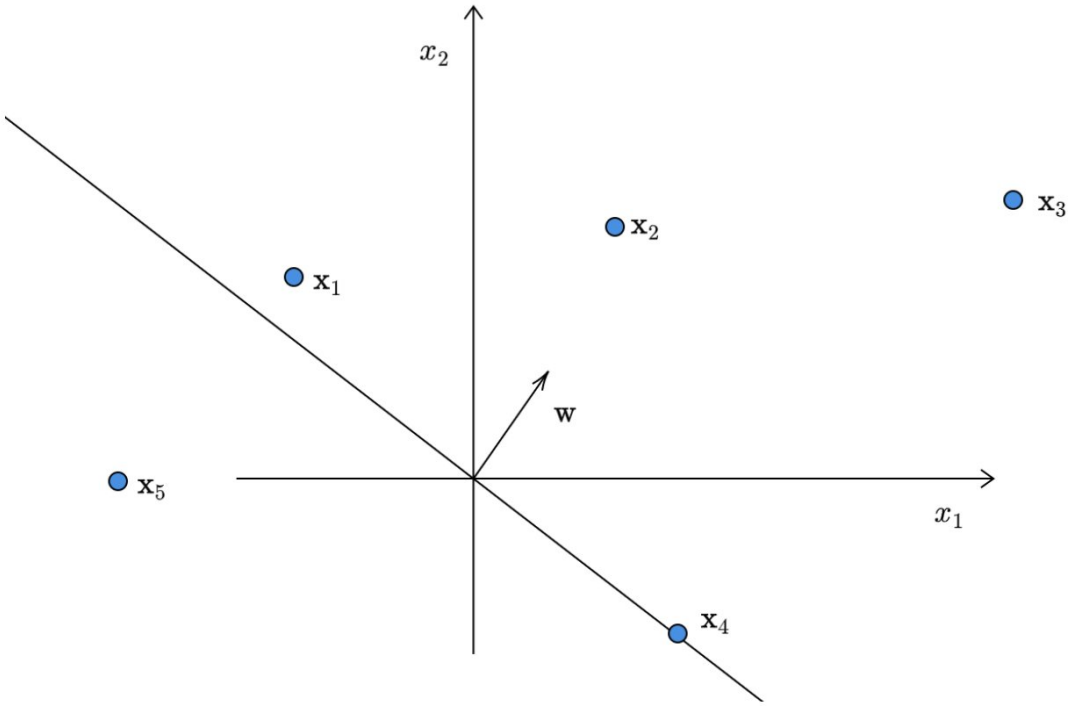
Solution

We would have to compute 100 distances: $d(\mathbf{x}_{\text{test}}, \mathbf{x}_i)$ for $1 \leq i \leq 100$.

Common Data for questions (2) to (3)

Statement

Consider the following data-points in a binary classification problem. $\mathbf{w}$ is the weight vector corresponding to a linear classifier. The labels are $+1$ and -1.



Question-2

Statement

What is the predicted label for these five points?

Options

(a)

All five points are predicted as $+1$.

(b)

All five points are predicted as -1.

(c)

$\mathbf{x}_1, \mathbf{x}_2$ and $\mathbf{x}_3$ are predicted as $+1$ and $\mathbf{x}_4, \mathbf{x}_5$ are predicted as -1.

(d)

$\mathbf{x}_1, \mathbf{x}_2$ and $\mathbf{x}_3$ are predicted as -1 and $\mathbf{x}_4, \mathbf{x}_5$ are predicted as $+1$.

Answer

(c)

Solution

For a linear classifier with weight vector $\mathbf{w}$, the prediction for a test-point $\mathbf{x}$ is given by:

$$y_{\text{pred}} = \text{sign}(\mathbf{w}^T\mathbf{x}) = \begin{cases} 1, & \mathbf{w}^T\mathbf{x} > 0 \\ -1, & \mathbf{w}^T\mathbf{x} \leq 0 \end{cases}$$

The geometric interpretation of this is as follows: a linear classifier divides the space into two half-spaces. This decision is made just by looking at the sign of the dot-product.

Half-space-1: The dot-product is positive.

$$\mathbf{w}^T\mathbf{x} > 0$$

This corresponds to those cases in which the data-point makes an acute angle with the weight vector.

Half-space-2: The dot-product is non-positive.

$$\mathbf{w}^T\mathbf{x} \leq 0$$

This corresponds to those cases in which the data-point makes either a right angle or an obtuse angle with the weight vector. When a point lies on the line (right angle with weight), it could be classified either way. As per convention, we choose -1.

Question-3

Statement

Which of the following statements are true?

Options

(a)

$$0 < \mathbf{w}^T\mathbf{x}_1 < \mathbf{w}^T\mathbf{x}_2 < \mathbf{w}^T\mathbf{x}_3$$

(b)

$$0 < \mathbf{w}^T\mathbf{x}_3 < \mathbf{w}^T\mathbf{x}_2 < \mathbf{w}^T\mathbf{x}_1$$

(c)

$$\mathbf{w}^T\mathbf{x}_4 = 0$$

(d)

$$\mathbf{w}^T x_5 > 0$$

(e)

$$\mathbf{w}^T x_5 < 0$$

Answer

(a), (c), (e)

Solution

The farther away a data-point is from the decision boundary, the larger the magnitude of its projection onto the weight vector. Hence:

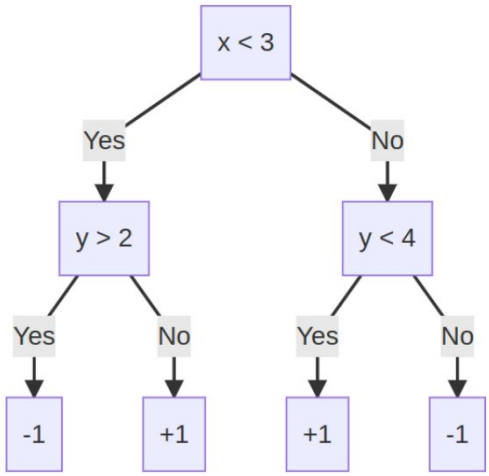
$$|\mathbf{w}^T\mathbf{x}_3| > |\mathbf{w}^T\mathbf{x}_2| > |\mathbf{w}^T\mathbf{x}_1|$$

Since all these three points lie on the positive half-plane, these dot products will be positive and we can remove the modulus sign. As for point $\mathbf{x}_4$, its projection on $\mathbf{w}$ will be zero as it is orthogonal to it. Finally, $\mathbf{w}^T\mathbf{x}_5$ will be negative as $\mathbf{x}_5$ lies in the negative half-plane.

Comprehension Type (4 - 5)

Statement

Consider the following decision tree for a binary classification problem that has two features: (x, y).



Question-4

Statement

Which of the following statements are true?

Options

(a)

$(1, 0)$ is classified as -1

(b)

$(0, 1)$ is classified as $+1$

(c)

$(4, 2)$ is classified as -1

(d)

$(5, 7)$ is classified as -1

Answer

(b), (d)

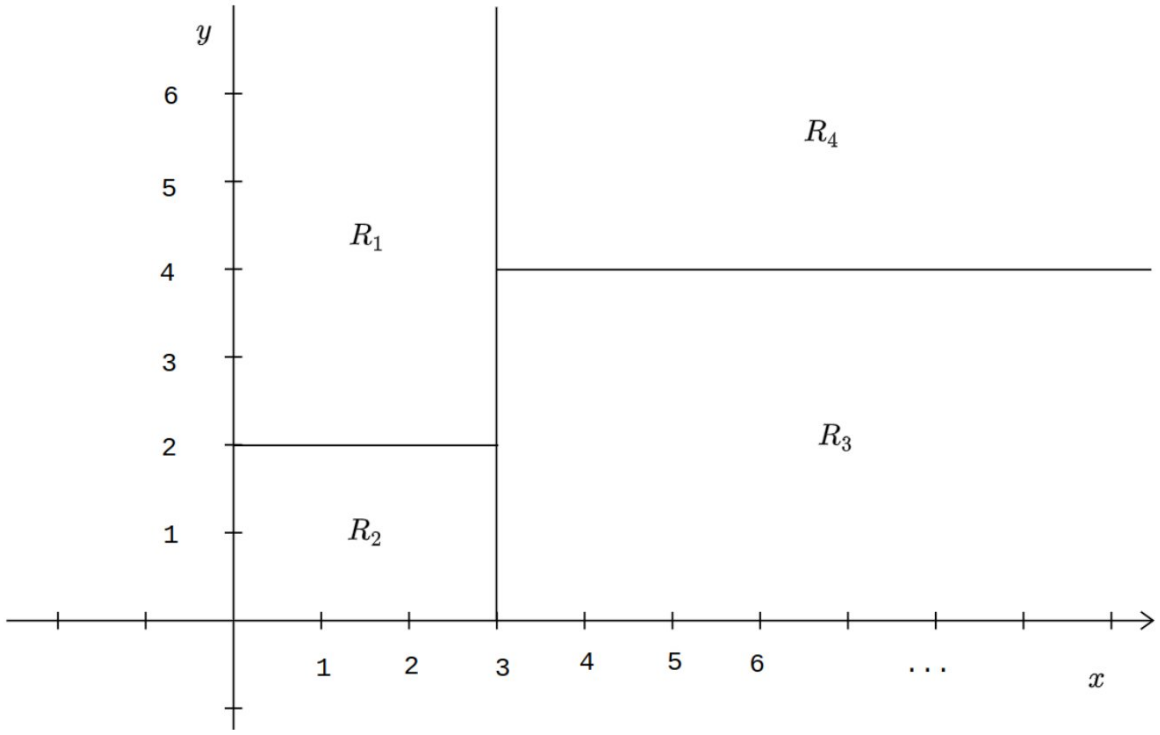
Solution

Start at the root of the decision tree and keep traversing the internal nodes until you end up with a prediction/leaf node.

Question-5

Statement

This decision tree partitions the feature space into four regions as shown below:



Assume that $x, y \geq 0$ for all points. What can you say about the predicted labels of points that fall in the four regions?

Options

(a)

$$R_1 : +1$$
$$R_2 : -1$$
$$R_3 : +1$$
$$R_4 : -1$$

(b)

$$R_1 : -1$$
$$R_2 : +1$$
$$R_3 : +1$$
$$R_4 : -1$$

(c)

$$R_1 : -1$$
$$R_2 : +1$$
$$R_3 : -1$$
$$R_4 : +1$$

(d)

$$R_1 : +1$$
$$R_2 : +1$$
$$R_3 : -1$$
$$R_4 : -1$$

Answer

(b)

Solution

The tree partitions the space into four regions: R_1, R_2, R_3, R_4. These four regions correspond to the four leaves from left to right.

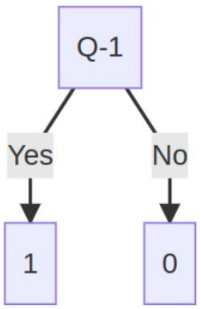
Question-6

Statement

Consider the following training dataset for a binary classification task:

x	y
2	1
10	0
8	0
-4	1
0	1
9	0

The following decision tree cleanly separates the two classes, such that the resulting leaves are pure.



Q-1 is of the form $x < v$. How many possible integer values can v take?

Answer

6

Solution

v can take any one of the values from this set: $\{3, 4, 5, 6, 7, 8\}$. It can also take the value 8 as $x < 8$ will go to the right branch for the data-point 8.

Question-7

Statement

p is the proportion of points with label $+1$ in some leaf in a decision tree. For what values of p will this node be assigned a label of $+1$? Select the most appropriate option.

Options

(a)

$0 \leq p < 0.5$

(b)

$0.5 < p \leq 1$

(c)

$0 \leq p \leq 1$

Answer

(b)

Solution

In a vote, we need more points to belong to class 1 than class 0. Therefore, $p > 0.5$

Question-8

Statement

Q is some internal node (question node) of a decision tree that splits into two leaf nodes (prediction nodes) L and R. The proportion of data-points belonging to class 1 in each of the three nodes is the same and is equal to 0.7. What is the information gain for the question corresponding to node Q?

Answer

0

Solution

Let p be the proportion of ones in each node. Then, all three nodes have the same entropy as entropy only depends on p. Call this E. If γ is the ratio into which the original dataset is partitioned by this question, then:

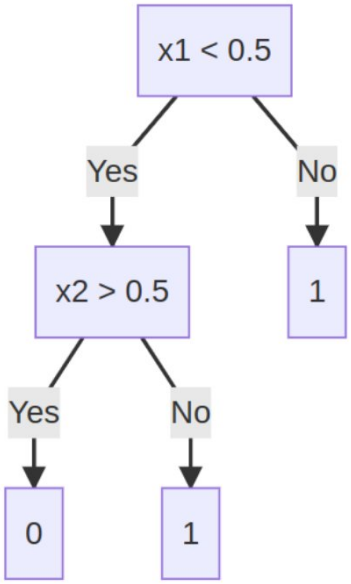
$$IG = E - [\gamma E + (1 - \gamma)E] = 0$$

We gain nothing because of this split. Intuitively this makes sense because in both leaf nodes the status quo prevails. To put it in another way, none of the child nodes have become purer than the parent node. They have the same level of impurity as their parent.

Question-9

Statement

Consider the following decision tree for a classification problem in which all the data-points are constrained to lie in the unit square in the first quadrant. That is, $0 \leq x_1, x_2 \leq 1$.



If a point is picked at random from the unit square, what is the probability that the decision tree predicts this point as belonging to class 1?

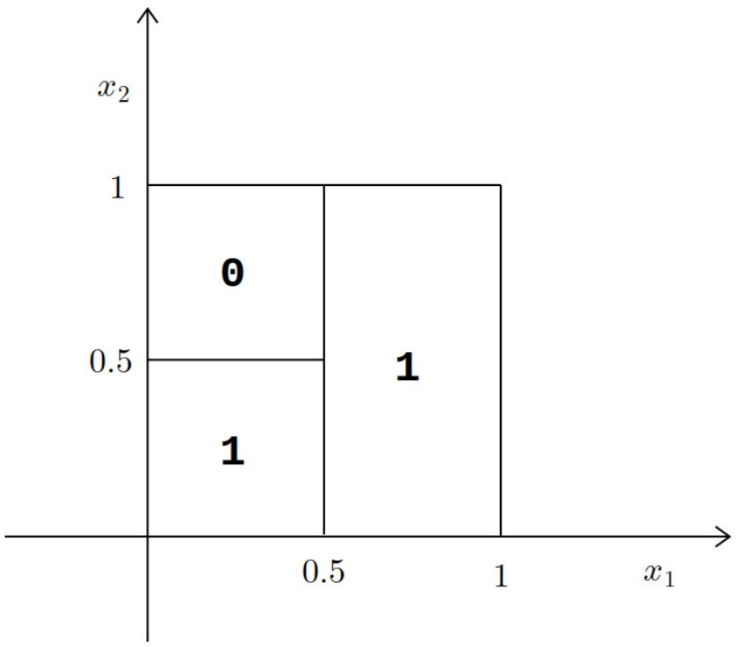
Answer

0.75

Range: [0.74, 0.76]

Solution

The decision regions will look as follows:

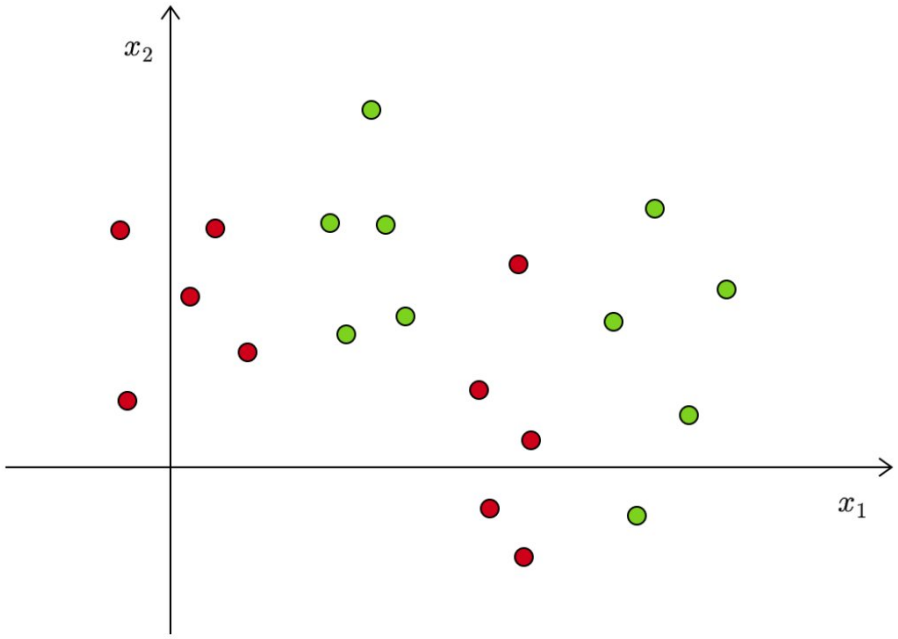


We see that 75% of the region belongs to class 1, hence the required probability is 0.75.

Question-10

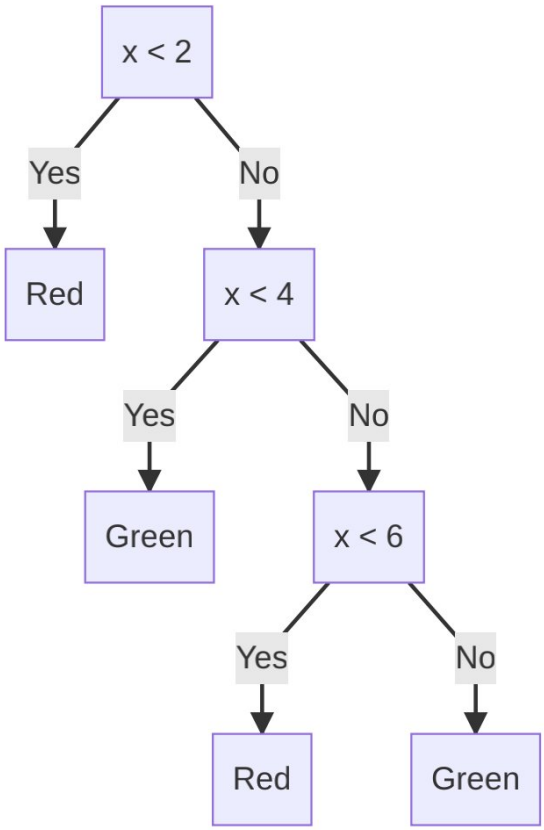
Statement

Consider the following dataset. By visually inspecting the data-points, construct a decision tree.



Solution

The structure of the data seems clear. There are four blobs of red and green points. From left to right, we have blob-1, blob-2, blob-3 and blob-4. They take the colors red, green, red and green respectively. Assume that the lines parallel to the y-axis that separate consecutive blobs are $x = 2$, $x = 4$ and $x = 6$. Think about why we need only three lines. One tree that we could construct has the following form:



Question-11

Statement

$H(p)$ is the entropy of a node. p is the proportion of points that belong to class 1. q is the proportion of points that belong to class 0. Which of the following statements is true?

Options

(a)

$H(p) > H(q)$

(b)

$H(p) < H(q)$

(c)

$H(p) = H(q)$

(d)

Insufficient information. We need the number of data-points that are in this node.

Answer

(c)

Solution

We notice the symmetry in the entropy function about the axis $p = 0.5$. Therefore, $H(p) = H(1-p)$:

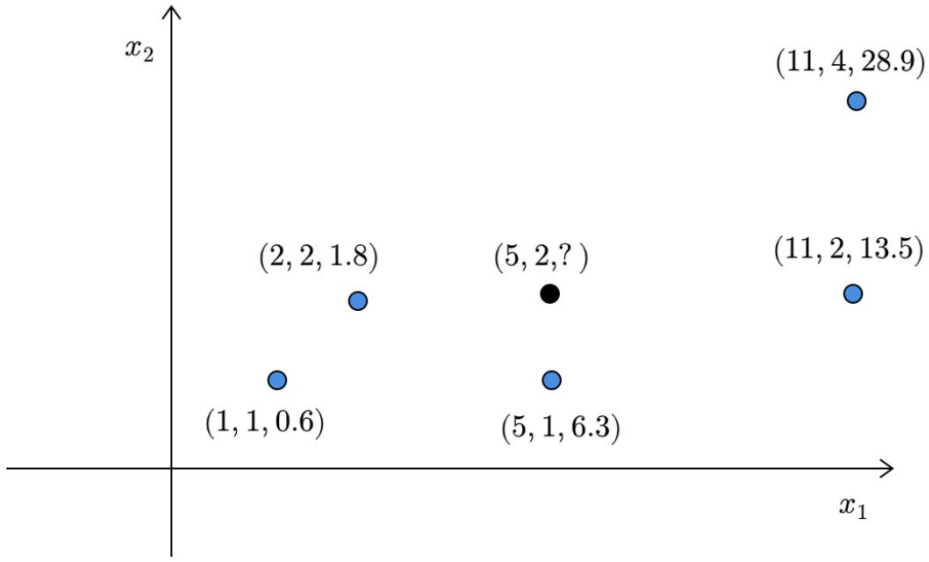
$$H(1-p) = -(1-p)\log(1-p) - p\log p = H(p)$$

Since $q = 1 - p$, we have $H(p) = H(q)$. So, for determining the impurity of a node, we could either use the proportion of ones or the proportion of zeros.

Question-12

Statement

The k-NN algorithm can be adapted to perform regression giving us what is called the k-NN regressor. Rather than looking at the majority vote among the k nearest neighbors, a k-NN regressor computes the mean of the labels of the k nearest neighbors and returns this as the prediction. Consider the following dataset:



All blue points belong to the training dataset. The annotation for each point should be understood as (x_1, x_2, y). What is the predicted label for the black test-point? Use $k = 3$.

Answer

2.9

Range: [2.8, 3]

Solution

The three nearest neighbors are:

- $(1, 1)$
- $(5, 1)$
- $(2, 2)$

The prediction for the test-point is the average of the labels corresponding to these three points:

$$\frac{1.8 + 0.6 + 6.3}{3} = 2.9$$

Practice assignment

Question 1

Statement

Consider a 3-class classification dataset with labels $0, 1,$ and 2. The data points belong to $\{0, 1, 2\}^3$. If we apply the generative model-based algorithm on the same dataset, how many features need to be estimated? Assume that the features given the label are not independent.

Answer

80

Solution

We need to estimate the parameters for $P(y = 0), P(y = 1)$. Since $P(y = 0) + P(y = 1) + P(y = 2) = 1$, we need to estimate two parameters for the distribution of y.

For the distribution of $x|y = 0$, we need to estimate the parameters for $P(x|y = 0)$.

Since $x \in \{0, 1, 2\}^3$, we can have $3^3 = 27$ possible data points and we need to estimate the probability for 26 such points as the sum will be one.

Similarly, for $x|y = 1$ and $x|y = 2$, we need 26 parameters each.

Therefore, total parameters to estimate = $2 + 3(26) = 80$

Question 2

In question 1, if the features are conditionally independent given the labels, how many parameters need to be estimated?

Answer

20

Solution

We need to estimate the parameters for $P(y = 0), P(y = 1)$. Since $P(y = 0) + P(y = 1) + P(y = 2) = 1$, we need to estimate two parameters for the distribution of y.

For a given label (say $y = 0$), we need to estimate
$P(f_1 = 0|y = 0), P(f_1 = 1|y = 0), P(f_2 = 0|y = 0), P(f_2 = 1|y = 0), P(f_3 = 0|y = 0), P(f_3 = 1|y = 0)$

Similarly, for labels $y = 1$ and $y = 2$.

Therefore, total parameters to estimate = $2 + 3(6) = 20$

Common data for questions 3, 4, and 5

Statement

Consider a naive Bayes model is trained on the following data matrix X of shape (d, n) and corresponding label vector y:

$$X = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix} \quad y = [1, 1, 0]^T$$

Assume that $\hat{p}$ and $\hat{p}_j^{y_i}$ are estimates for $P(y = 1)$ and $P(f_j = 1|y = y_i)$, respectively. Here, $f_i;\ i = 1, 2$ is the i^{th} feature. These parameters are estimated using MLE.

Question 3

Statement

If a test point has label 0, what will be the probability that the point is $[0, 0]^T$?

Options

(a)

$\hat{p}_1^0 \times \hat{p}_2^0 \times (1 - \hat{p})$

(b)

$(1 - \hat{p}_1^0) \times (1 - \hat{p}_2^0) \times (1 - \hat{p})$

(c)

$\hat{p}_1^0 \times \hat{p}_2^0$

(d)

$(1 - \hat{p}_1^0) \times (1 - \hat{p}_2^0)$

Answer

(d)

Solution

We know that $\hat{p}_j^{y_i}$ is the estimate for $P(f_j = 1|y = y_i)$. It implies that $(1 - \hat{p}_j^{y_i})$ is the estimate for $P(f_j = 0|y = y_i)$

Therefore,

$$\begin{aligned} P(x = [0, 0]^T|y = 0) &= P(f_1 = 0|y = 0).P(f_2 = 0|y = 0) \\ &= (1 - \hat{p}_1^0) \times (1 - \hat{p}_2^0) \end{aligned}$$

Question 4

Statement

What is the value of p_1^1?

Answer

1

Solution

$\hat{p}_1^1$ is the estimate for $P(f_1 = 1|y = 1)$

$$\hat{p}_1^1 = \frac{\sum_{i=1}^{n} \mathbb{1}(f_1 = 1, y = 1)}{\sum_{i=1}^{n} \mathbb{1}(y = 1)}$$

Here, the first two examples belong to label 1 and the first feature value for both examples is 1, therefore

$\hat{p}_1^1 = 1$

Question 5

Statement

What will be the probability that a test data point $[0, 1]$ is labeled as 0? Assume no smoothing of data is done.

Answer

0

Solution

$$P(y=0|x=[0,1]) = \frac{P(x=[0,1]|y=0).P(y=0)}{P(x=[0,1])}$$
$$= \frac{(1-\hat{p}_1^0)\hat{p}_2^0\hat{p}}{P(x=[0,1])}$$

Here

$\hat{p}_2^0 = 0$ since, label zero example takes only zeros for all the features

Therefore,

$P(y=0|x=[0,1]) = 0$

Question 6

Statement

Consider a spam classification problem that was modeled using naive Bayes. The features take a value of 1 or 0 depending on whether a word is present in the email or not. Assume that the probability of a mail being spam is 0.2. The following table gives the estimation for conditional probabilities for some of the words:

word	label	$P(\text{word}\|\text{label})$
Hurray!	spam	0.7
win	spam	0.2
exciting	spam	0.01
prizes	spam	0.3
Hurray!	Non-spam	0.01
win	Non-spam	0.02
exciting	Non-spam	0.01
prizes	Non-spam	0.1

Consider a mail with the following sentence: "Hurray! win exciting prizes"

With what probability the mail will be predicted spam? Assume that these are the only possible words (that is there are four features) in a mail. Write your answer correct to two decimal places. A

Answer

0.99 Range: $[0.98.\ 1]$

Solution

$$P(y=\text{spam}|\text{mail}) = \frac{P(\text{mail}|\text{spam}).P(\text{spam})}{P(\text{mail}|\text{spam}).P(\text{spam}) + P(\text{mail}|\text{non-spam}).P(\text{non-spam})}$$

Here,

$P(\text{spam}) = 0.2, \quad P(\text{non-spam}) = 0.8$

Denote spam as 0 and non-spam as 1.

Therefore,

$$\begin{aligned}
P(y=0|\text{mail}) &= \frac{P(\text{mail}|0)(0.2)}{P(\text{mail}|0)(0.2) + P(\text{mail}|1)(0.8)} \\
&= \frac{P(\text{Hurray!}|0)P(\text{win}|0)P(\text{exciting}|0)P(\text{prizes}|0)(0.2)}{P(\text{Hurray!}|0)P(\text{win}|0)P(\text{exciting}|0)P(\text{prizes}|0)(0.2) + P(\text{Hurray!}|1)P(\text{win}|1)P(\text{exciting}|1)P(\text{prizes}|1)(0.8)} \\
&= \frac{0.7(0.2)(0.01)(0.3)(0.2)}{0.7(0.2)(0.01)(0.3)(0.2) + 0.01(0.02)(0.01)(0.1)(0.8)} \\
&= 0.99
\end{aligned}$$

Question 7

Statement

A binary classification dataset contains only one feature and the data points given the label follow the gaussian distributions whose means and variances are already estimated as:

$$x|(y=0) \sim \mathrm{N}(0,1)$$
$$x|(y=1) \sim \mathrm{N}(2,2)$$

What will be the decision boundary learned using the naive Bayes algorithm? Assume that $\hat{p}$, an estimate for $P(y=1)$, is 0.5.

Hint: Solve $P(y=1|x) = P(y=0|x)$

Options

(a)

$$\frac{x^2}{2} - \frac{(x-2)^2}{4} = \frac{1}{2}\ln 2$$

(b)

$$\frac{x^2}{4} = \frac{1}{2}\ln 2$$

(c)

$$4x = 2\ln 2 + 4$$

(d)

$$\frac{x^2}{4} - \frac{(x-2)^2}{2} = \ln 2$$

Answer

(a)

Solution

The decision boundary is given by

$$\{x : P(y=1|x) = P(y=0|x)\}$$

$$\begin{aligned}
&P(y=1|x) = P(y=0|x) \\
\Rightarrow &\frac{P(x|y=1).P(y=1)}{P(x)} = \frac{P(x|y=0).P(y=0)}{P(x)} \\
\Rightarrow &P(x|y=1) = P(x|y=0) \quad (\because P(y=0) = P(y=1) = 0.5) \\
\Rightarrow &\frac{1}{\sqrt{2\pi}\sqrt{2}}\exp(-(x-2)^2/4) = \frac{1}{\sqrt{2\pi}}\exp(-(x)^2/2) \\
\Rightarrow &\exp(-(x-2)^2/4) = \sqrt{2}\exp(-(x)^2/2) \\
\Rightarrow &\ln(\exp(-(x-2)^2/4)) = \ln(\sqrt{2}\exp(-(x)^2/2)) \\
\Rightarrow &\frac{-(x-2)^2}{4} = \frac{1}{2}\ln 2 + \frac{-x^2}{2} \\
\Rightarrow &\frac{-x^2}{2} - \frac{-(x-2)^2}{4} = \frac{1}{2}\ln 2
\end{aligned}$$

Common data for questions 8, 9 and 10

Statement

Consider the gaussian naive Bayes algorithm was run on the following dataset:

feature 1 (f_1)	**feature 2** (f_2)	**Label**
1.5	1.6	0
2.1	2.4	1

feature 1 (f_1)	feature 2 (f_2)	Label
2.9	1.5	1
1.7	0.8	1

Question 8

Statement

What will be the value of $\hat{p}$?

Answer

0.75 Range; [0.74, 0.76]

Solution

$$\hat{p} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{3}{4}$$

Question 9

Statement

What will be the value of $\hat{\mu}_0$?

Options

(a)

1.5

(b)

1.55

(c)

$(1.5, 1.6)$

(d)

$(2.23, 1.56)$

Answer

(c)

Solution

$$\begin{aligned}\hat{\mu}_0 &= \frac{\sum_{i=1}^{n} \mathbb{1}(y_i = 0) x_i}{\sum_{i=1}^{n} \mathbb{1}(y_i = 0)} \\ &= \frac{(1.5, 1.6)}{1}\end{aligned}$$

Question 10

Statement

What will be the value of $\hat{\mu}_1$?

Options

(a)

2.23

(b)

1.56

(c)

$(1.5, 1.6)$

(d)

$(2.23, 1.56)$

Answer

(d)

Solution

$$\begin{aligned}\hat{\mu}_1 &= \frac{\sum_{i=1}^{n} \mathbb{1}(y_i = 1) x_i}{\sum_{i=1}^{n} \mathbb{1}(y_i = 1)} \\ &= \frac{[2.1, 2.4] + [2.9, 1.5] + [1.7, 0.8]}{3} \\ &= (2.23, 1.56)\end{aligned}$$

Practice

Question-1

Statement

Which of the following data sets is/are linearly separable?

D1: $\left(\begin{bmatrix}1\\2\end{bmatrix}, +1\right), \left(\begin{bmatrix}1\\-2\end{bmatrix}, +1\right), \left(\begin{bmatrix}-2\\1\end{bmatrix}, -1\right), \left(\begin{bmatrix}-3\\2\end{bmatrix}, -1\right), \left(\begin{bmatrix}3\\2\end{bmatrix}, +1\right) \left(\begin{bmatrix}-3\\3\end{bmatrix}, -1\right),$

D2: $\left(\begin{bmatrix}2\\0\end{bmatrix}, +1\right), \left(\begin{bmatrix}2\\1\end{bmatrix}, +1\right), \left(\begin{bmatrix}3\\1\end{bmatrix}, -1\right), \left(\begin{bmatrix}-1\\0\end{bmatrix}, -1\right), \left(\begin{bmatrix}-1\\-1\end{bmatrix}, -1\right) \left(\begin{bmatrix}-2\\0\end{bmatrix}, -1\right),$

Options

(a)

D1

(b)

D2

(c)

Both D1 and D2

(d)

Neither D1 nor D2

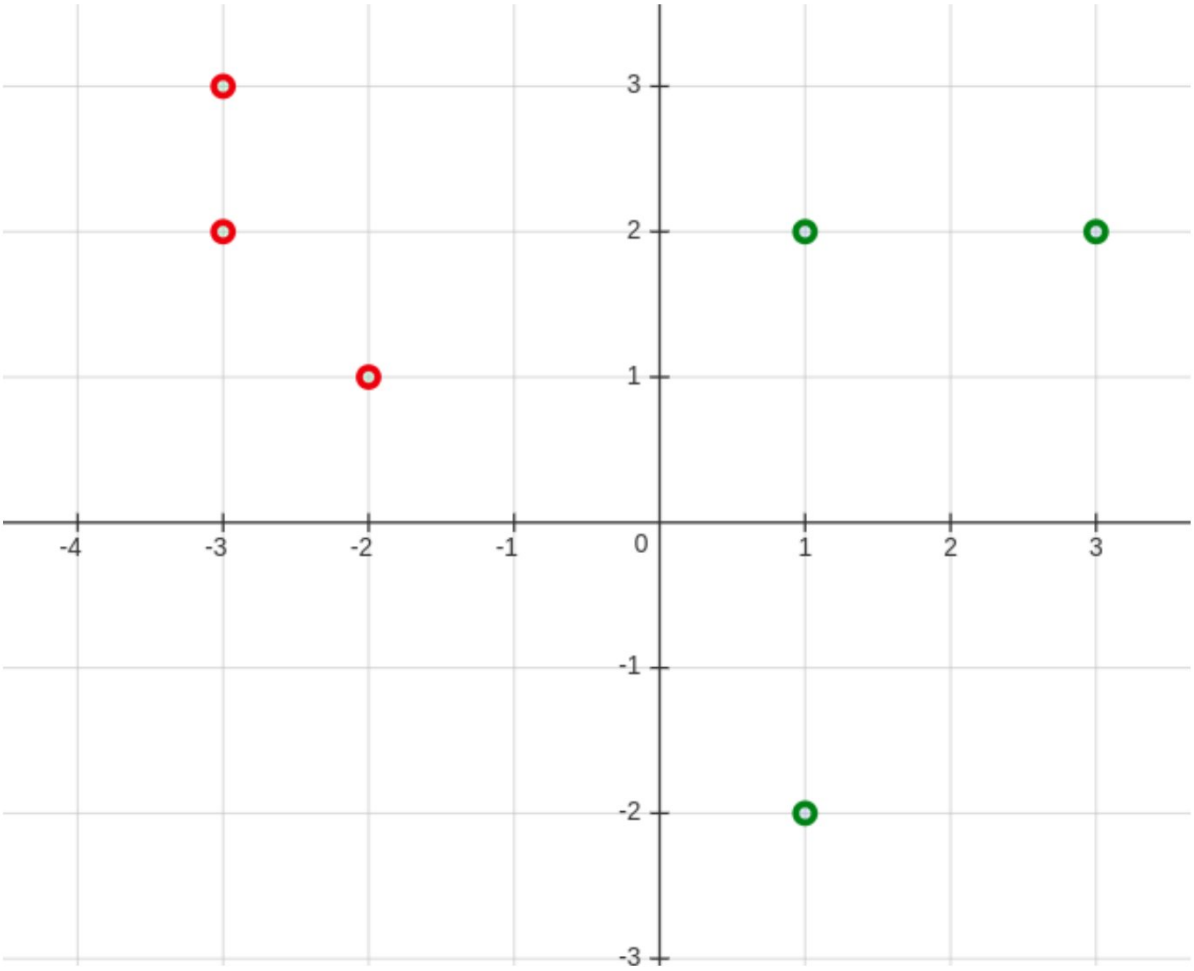
Answer

(a)

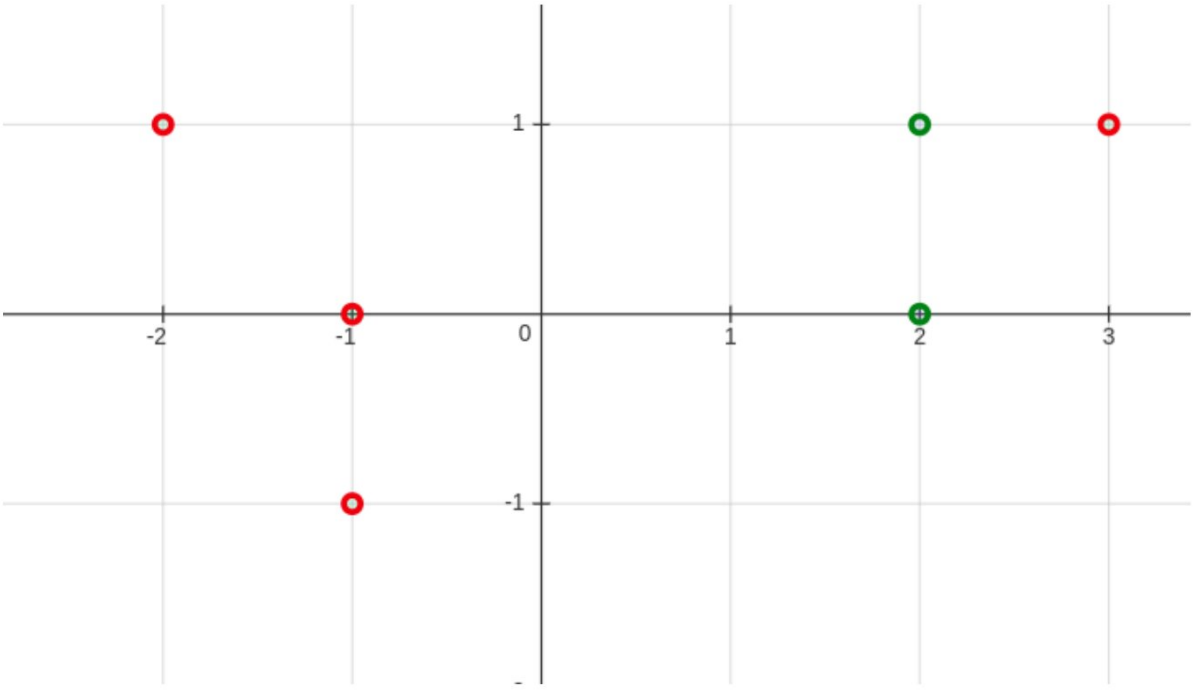
Solution

Below, red points belong to -ve class, green points belong to +ve class.

D1:



D2:



In D2, there is an intermixing of data points belonging to green and red class and there is no linear separator that can separate the data points beloning to the two classes.

Question-2

Statement

Consider the following data set with three data points:

$$\left(\begin{bmatrix}0\\2\end{bmatrix}, +1\right), \left(\begin{bmatrix}0\\-2\end{bmatrix}, +1\right), \left(\begin{bmatrix}-2\\1\end{bmatrix}, -1\right)$$

If the Perceptron algorithm is applied to this data with the initial weight vector w^0 to be a zero vector, what will be the outcome?

Options

(a)

The algorithm will converge with $w = \begin{bmatrix}2\\-1\end{bmatrix}$

(b)

The algorithm will converge with $w = \begin{bmatrix}2\\1\end{bmatrix}$

(c)

The algorithm will converge with $w = \begin{bmatrix}-2\\1\end{bmatrix}$

(d)

The algorithm will never converge.

Answer

(d)

Solution

$w^0 = [0, 0]$

Iteration 1:

${w^0}^T x_1 = 0$, hence $y_{1_{pred}} = +1$

${w^0}^T x_2 = 0$, hence $y_{2_{pred}} = +1$

${w^0}^T x_3 = 0$, hence $y_{3_{pred}} = +1$

Mistake is for x_3

Hence, $w_1 = w_0 + x_3 y_3 = [2, -1]$

Iteration 2:

When we compute ${w}^T x_i$'s' we find a mistake for x_1, hence we update the weight vector to:

$w_2 = w_1 + x_1 y_1 = [2, 1]$

Iteration 3:

The mistake is found to be for x_2.

$w_3 = w_2 + x_2 y_2 = [2, -1]$

w_3 is same as w_1.

Hence, perceptron will keep oscillating ans repeating these weights, and will never converge.

Question-3

Statement

Assume that Perceptron algorithm is applied to a data set in which the maximum of the lengths of the data points is 4. Assume that the squared length of the weight vector in an iteration of the algorithm is 36. As per the given information, which of the following can be a valid squared length of the new weight vector obtained in the next iteration?

Options

(a)

45

(b)

55

(c)

60

(d)

65

Answer

(a)

Solution

R = 4

$||w^l||^2 = 36$

$||w^{l+1}||^2 \leq ||w^l||^2 + R^2 \leq 36 + 16 \leq 52$

Question-4

Statement

Assume that Perceptron algorithm is applied to a data set in which the maximum of the lengths of the data points is 4 and the value of margin (γ) is 1. What is the maximum number of mistakes that can be made by the algorithm on this data?

Options

(a)

10

(b)

13

(c)

16

(d)

19

Answer

(c)

Solution

$R = 4$

$\gamma = 1$

#mistakes $\leq \frac{R^2}{\gamma^2} \leq 16/1 \leq 16$

Question-5

Statement

Assume that each of the four corners of a unit square represents a data point. Each of these data points can either be assigned a positive class or a negative class.

(a) How many different data sets (as per different assignments of positive and negative labels) will be possible using these data points?

(b) Out of those data sets, how many will the Perceptron algorithm be able to correctly classify?

Options

(a)

(a): 8, (b): 4

(b)

(a):8, (b): 6

(c)

(a): 16, (b): 14

(d)

(a):16, (b):12

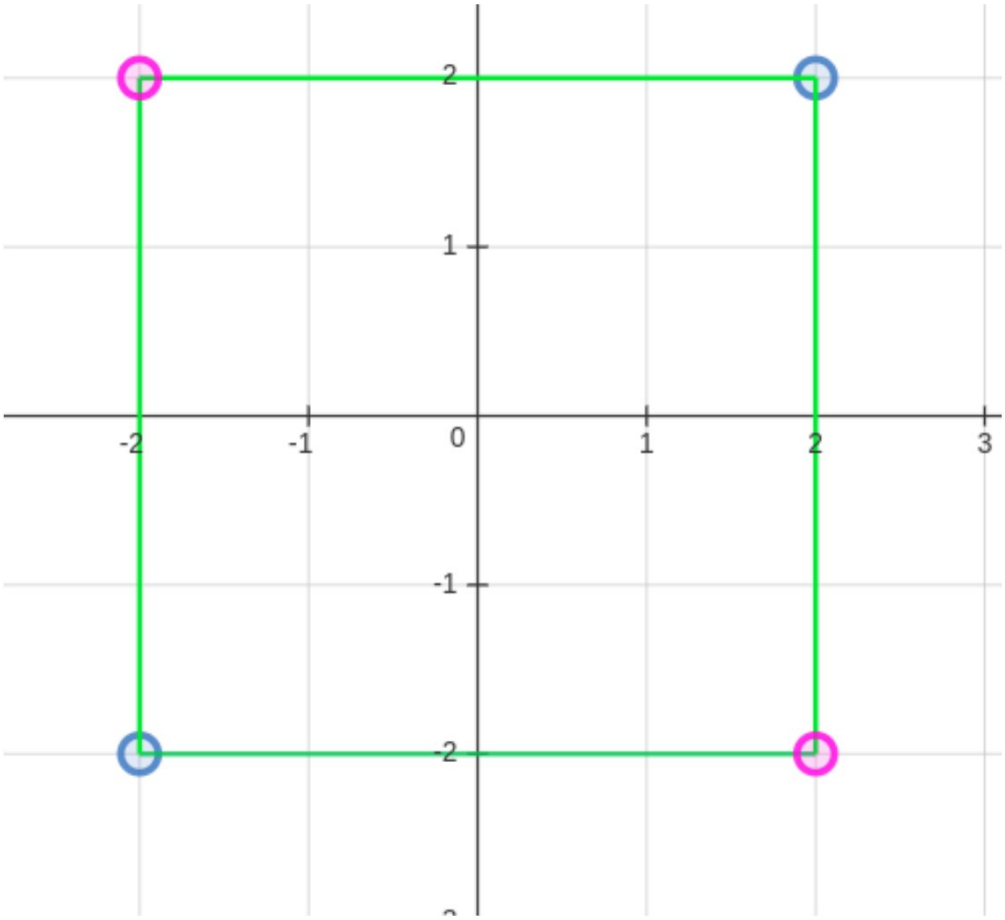
Answer

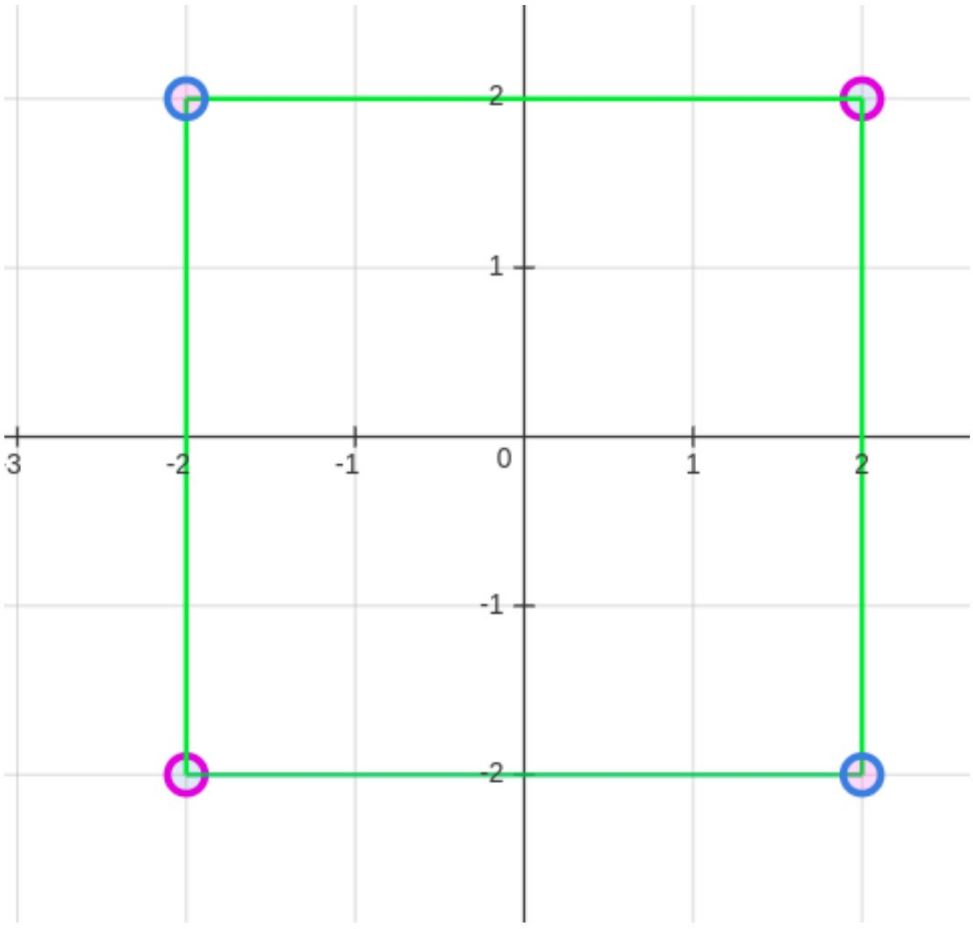
(c)

Solution

(a) Each corner can be assigned either a positive class or a negative class. Hence there are two possibilities for each corner, resulting into a total of 2^4 = 16 possibilities.

(b) The following two data sets will not be linearly separable.


2
1
-2
-1
0
1
2
3
-1
-2


2
1
3
-2
-1
0
1
2
-1
-2

Question-6

Statement

Assume that you trained a Perceptron and after training got finished, you found that the data that was used for training had been accidentally labeled opposite to what it should have been, i.e., every example that was labelled +1 should have been a -1, and vice versa.

Assume that you no longer have the data and have no ability to change the code that uses the perceptron to flip its answers. All you have access to is the weight vector of the perceptron ($w_{incorrect}$). How would you change the weights to obtain $w_{updated}$ in order to flip all the answers? You may assume that there are no data points that fall exactly on the boundary between the two classes.

Options

(a)

$w_{updated} = -1 * w_{incorrect}$

(b)

$w_{updated} = 1 - w_{incorrect}$

(c)

$w_{updated} = 1 + w_{incorrect}$

(d)

$w_{updated} = -1/w_{incorrect}$

Answer

(a)

Solution

$W_{incorrect}^T x \geq 0$ will predict a positive class and $W_{incorrect}^T x < 0$ will predict a negative class.

If we only have access to the weights, we can negate the weight vector, i.e.,
$w_{updated} = -1 * w_{incorrect}$

So that wherever $W_{incorrect}^T x \geq 0$ and we get a positive class as a prediction, $W_{updated}^T x \leq 0$ and hence will predict a negative class, and vice versa.

Question-7

Statement

Assume that in the truth tables of logic gates OR, AND and XOR, every occurrence of 0 is replaced by -1 and every occurrence of 1 is replaced by +1. In this way, three different data sets are generated respectively.

Perceptron algorithm will not be able to correctly classify the data set produced by which of the three gates?

Options

(a)

OR

(b)

AND

(c)

XOR

(d)

None of these

(e)

All of these

Answer

(c)

Solution

Following will be the datat sets generated:

f_1	f_2	OR	AND	XOR
-1	-1	-1	-1	-1
-1	1	1	-1	1
1	-1	1	-1	1
1	1	1	1	-1

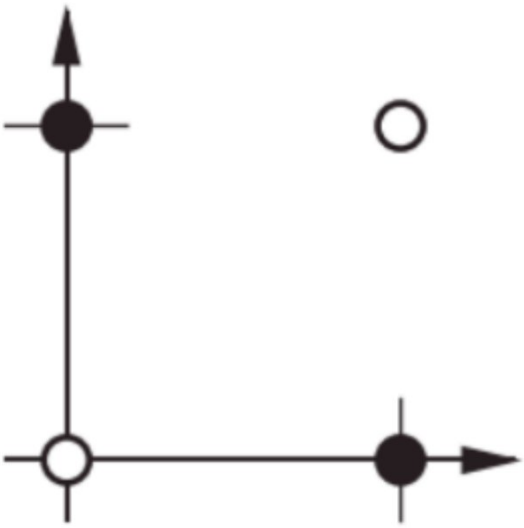
XOR data set is similar to what we saw in Q5, and is not linearly separable.

Question-8

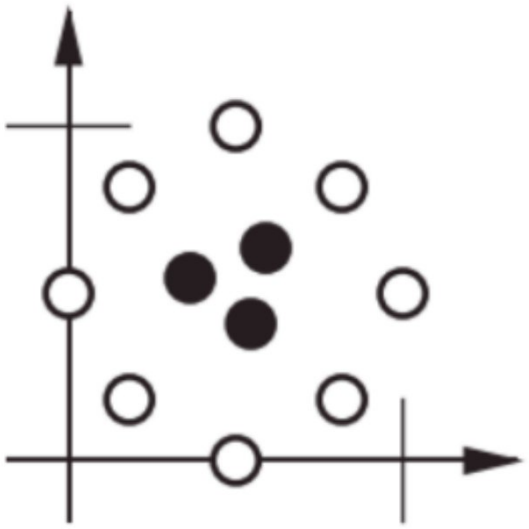
Statement

Consider the following three data sets, where a white circle represents a negative class and a black circle represents a positive class:

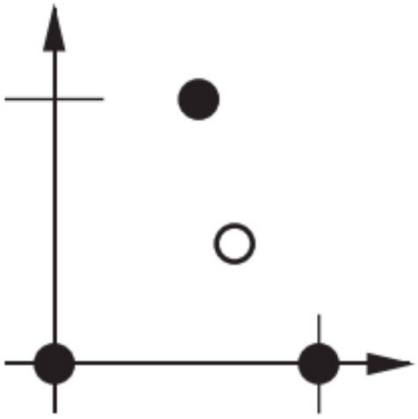
1.



2.



3.



Which of these data sets will the Perceptron Algorithm be able to correctly classify?

Options

(a)

1 only

(b)

1 and 3

(c)

2 only

(d)

3 only

(e)

All of these

(f)

None of these

Answer

(f)

Solution

None of the data sets is linearly separable.

Question-9

Statement

Consider two points x_1 and x_2 with $w^T x$ value of 10 and -10 respectively. Let p_{x_1} and p_{x_2} be the probabilities returned by Logistic Regression for these two data points. Which of the following is correct?

Options

(a)

p_{x_1} will be same as p_{x_2}.

(b)

p_{x_2} will be much higher than p_{x_1}.

(c)

p_{x_1} will be much higher than p_{x_2}.

Answer

(c)

Solution

$$p_{x_1} = \frac{1}{1 + e^{-10}}$$

$$p_{x_2} = \frac{1}{1 + e^{10}}$$

Hence, p_{x_1} will be much higher than p_{x_2}.

Week 11

Question 1

Statement

In a random forest model, let $p < d$ be the number of randomly selected features that are used to identify the best split at any node of a tree. Which of the following is true? (d is the total number of features)

Options

(a)

Increasing p reduces the correlation between any two trees in the forest.

(b)

Decreasing p reduces the correlation between any two trees in the forest.

(c)

Increasing p increases the performance of individual trees in the forest.

(d)

Decreasing p increases the performance of individual trees in the forest.

Answer

(b), (c)

Solution

If we increase the number of randomly selected features in a random forest model, classifiers tend to have a similar structure and the correlation between them will be higher. But decreasing the number of features may lead to a wide variety of trees and the correlation between them will decrease.

Increasing the number of features increases the performance of individual trees as it considers the most number of features to split the node.

Question 2

Statement

A dataset was generated as per the following process:

$$y = a + bx + cx^2 + dx^3 + ex^4 + \varepsilon$$

where a, b, c, d, and e are the constants and the error ε follows Gaussian distribution with mean 0 and a fixed variance. The following two models were fit on the same dataset:

- model 1: $y = w_0 + w_1 x_1$
- model 2: $y = w_0 + w_1 x + w_2 x^2 + \ldots + w_8 x^8 + w_9 x^9$

Select the correct options.

Options

(a)

Model 1 has a higher bias than model 2.

(b)

Model 1 has a higher variance than model 2.

(c)

Model 1 is more likely to overfit.

(d)

Model 2 is more likely to overfit.

(e)

Model 1 is more likely to underfit.

(f)

Model 2 is more likely to underfit.

Answer

(a), (d), (e)

Solution

Model 1 has a less complex structure to capture the real structure of the dataset and therefore, it has a higher bias than model 2 as model 2 will try to fit each and every point in the dataset.

If we make a slight change in the dataset, model 2 may deviate a lot as it always tries to capture noise in the dataset, and therefore model 2 has a higher variance than model 1.

Model 1 has high bias and low variance and it tends to underfit.

Model 2 has low bias and high variance and therefore it tends to overfit.

Question 3

Statement

How does bagging help in improving the classification performance?

Options

(a)

It helps in reducing bias

(b)

It helps in reducing variance

(c)

Bagging is inefficient if the classifiers are fully uncorrelated (independent).

(d)

Bagging is inefficient if the classifiers are fully correlated.

Answer

(b), (d)

Solution

In bagging, classifiers having high variance are trained, and their aggregate outputs are considered as the final model. The aggregation results in reducing the variances of each classifier.

If the classifiers are fully correlated, it means on average their output is the same and, therefore aggregating them does not make much difference.

Question 4

Statement

Is the following statement true or false?

If $\alpha_i^* = 0$ in the soft-margin SVM problem, then the i^{th} data point is correctly classified by the optimal w^*.

Options

(a)

True

(b)

False

Answer

(a)

Solution

Given that $\alpha_i^* = 0$.

Since $\alpha_i^* + \beta_i^* = C, \quad \Rightarrow \beta_i^* = C$

Using 2nd CS condition, we have

$$\begin{aligned} &\beta_i^* \xi_i^* = 0 \\ &\Rightarrow \xi_i^* = 0 \qquad (\because \beta_i^* = C) \end{aligned} \tag{1}$$

Now the first constraint, we have

$$\begin{aligned} 1 - {w^*}^T x_i y_i - \xi_i^* &\leq 0 \\ \Rightarrow 1 - {w^*}^T x_i y_i &\leq 0 \\ \Rightarrow {w^*}^T x_i y_i &\geq 1 \end{aligned}$$

It implies that i^{th} data point is correctly classified by w^*.

Question 5

Statement

Assume that the first decision stump of the AdaBoost algorithm misclassifies 30 data points out of 100 data points. What will be the weights assigned to the incorrectly classified points for sampling the data points to train the second decision stump? Assume that error is defined as the proportion of incorrectly classified examples by the decision stump.

Options

(a)

$$\frac{\sqrt{7/3}}{100}$$

(b)

$$\frac{\sqrt{3/7}}{100}$$

(c)

$$\frac{\sqrt{7/3}}{30\sqrt{7/3} + 70\sqrt{3/7}}$$

(d)

$$\frac{\sqrt{3/7}}{30\sqrt{3/7} + 70\sqrt{7/3}}$$

Answer

(c)

Solution

Weights assigned for creating the 1st bag are given by

$$D_0(i) = \frac{1}{n} = \frac{1}{100}$$

the error by the first decision stump is

$$\text{error} = 0.3$$

Therefore

$$\alpha = \ln\left(\sqrt{\frac{1-0.3}{0.3}}\right) = \ln\left(\sqrt{\frac{7}{3}}\right)$$

The weight for the incorrectly classified points will be increased by

$$w_1(i) = w_0(i)e^{\alpha} \quad \text{the i for which } h_i(x) \neq y_i$$
$$= \frac{\sqrt{7/3}}{100}$$

The weight for the correctly classified points will be decreased by

$$w_1(i) = w_0(i)e^{-\alpha} \quad \text{the i for which } h_i(x) = y_i$$
$$= \frac{\sqrt{3/7}}{100}$$

There are 30 incorrectly classified points and 70 correctly classified points. We will normalize these weights to sum over all points equal to 1. For that, we will divide each weight with the sum of all weights.

That is

For incorrectly classified points weights will be

$$\frac{\frac{\sqrt{7/3}}{100}}{\frac{30\sqrt{7/3}}{100} + \frac{70\sqrt{3/7}}{100}}$$
$$= \frac{\sqrt{7/3}}{30\sqrt{7/3} + 70\sqrt{3/7}}$$

Common data for questions 6 and 7

Statement

You have been given a dataset in 1-d space, which consists of 4 positive data points $\{1, 2, 3, 4\}$ and 3 negative data points $\{-3, -2, -1\}$. We want to learn a soft-margin SVM (though the dataset is linearly separable) for this dataset. Think those points on the real line.

Question 6

Statement

If $C \to \infty$, how many support vectors do we have?

Answer

2

Solution

When $C \to \infty$, the soft margin problem turns out to be the hard margin and it is clear that the dataset is linearly separable. Therefore, the decision boundary will be the origin and the support vectors will be -1 and 1. So, there will be two support vectors.

Question 7

Statement

If $C = 0$, how many support vectors do we have?

Answer

0

Solution

If $C = 0$, then the problem collapses as $w = 0$ and the concept of support vectors becomes moot. So, there will be no support vectors as for support vectors, $w^T x_i = \pm 1$ but here, $w = 0$.

Question 8

Statement

Is the following statement true or false?

For a fixed size of the training and test set, increasing the complexity of the model always leads to a reduction of the test error.

Options

(a)

True

(b)

False

Answer

(b)

Solution

If we increase the complexity of the model, the model tends to overfit as it tries to fit each and every data point including noise in the training dataset but may fail to classify the test data point. Therefore, the model will tend to overfit, and test errors may go up.

Question 9

Statement

If we remove all the non-support vectors from the dataset, what will be the effect on the model?

Options

(a)

Model will overfit

(b)

Model will underfit

(c)

The model will not be changed

(d)

Can not say

Answer

(c)

Solution

Remember that the weight vector in the SVM problem comes out to be

$$w^* = \sum_{i=1}^{n} \alpha_i^* x_i y_i$$

That is only the points for which $\alpha_i > 0$ (the support vectors) decide on the weight vector.

Therefore, if we remove the non-support vectors from the dataset, the model will not change.

Practice

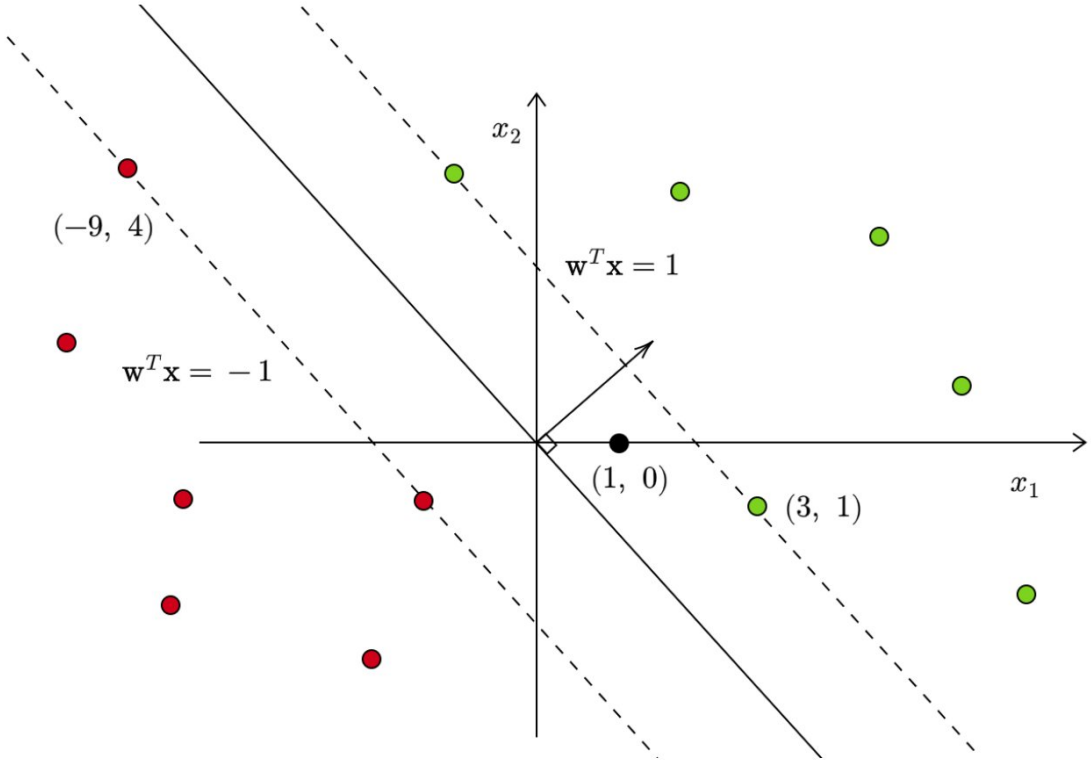
This document has 13 questions.

Common Data for Questions (1) to (6)

Statement

Consider a hard-margin SVM trained on a dataset in $\mathbb{R}^2$ for a binary classification task. Red and green points belong to the training dataset. Red points belong to class -1 and green points belong to class $+1$. The black-point is a test data-point. The dotted lines are the supporting hyperplanes for the SVM.

Note: We don't have access to the test data-point during training; it is given to us *after* the model has been learned on the training dataset.



Question-1

Statement

What is the maximum number of support vectors that the model could have?

Answer

4

Solution

The number of support vectors is upper bounded by the total number of points that lie on the supporting hyperplanes. Recall that support vectors are those points for which $\alpha_i^* > 0$. By complementary slackness, if $\alpha_i^* > 0$ then $(\mathbf{w}^T\mathbf{x}_i)y_i = 1$. So, we can rightfully claim that if $\alpha_i^* > 0$, then that point lies on one of the two supporting hyperplanes.

Now, can we claim that the number of support vectors is exactly equal to 4? This is a bit tricky. The claim we are trying to make here is stronger:

> If a point lies on either of the two supporting hyperplane, then it is a support vector.

Mathematically, this means the following. If $(\mathbf{w}^T\mathbf{x}_i)y_i = 1$, then $\alpha_i^* > 0$. This is something that is not guaranteed by complementary slackness. All that complementary slackness tell us is that for every point in the dataset, $\alpha_i^*[1 - (\mathbf{w}^T\mathbf{x}_i)y_i] = 0$. It could be the case that both $\alpha_i^* = 0$ and $(\mathbf{w}^T\mathbf{x}_i)y_i = 1$. In such a case, the i^{th} point wouldn't qualify as a support vector. To summarize, every support vector lies on one of the two supporting hyperplanes, but every point on the supporting hyperplanes need not be a support vector.

Question-2

Statement

What is the value of the weight vector $\mathbf{w}$? Select all options that are correct.

Options

(a)

$$\mathbf{w} = \begin{bmatrix} 5/21 \\ 2/7 \end{bmatrix}$$

(b)

$$\mathbf{w} = \begin{bmatrix} 5/3 \\ 2 \end{bmatrix}$$

(c)

$$\mathbf{w} = \begin{bmatrix} 7/2 \\ 21/5 \end{bmatrix}$$

(d)

$$\mathbf{w} = \begin{bmatrix} 1/2 \\ 3/5 \end{bmatrix}$$

Answer

(a)

Solution

Let the weight vector be $\mathbf{w} = \begin{bmatrix} w_1 & w_2 \end{bmatrix}^T$. Then, we have:

$$3w_1 + w_2 = 1$$

$$-9w_1 + 4w_2 = -1$$

Solving this system of equations, we get:

$$\mathbf{w} = \begin{bmatrix} 5/21 \\ 2/7 \end{bmatrix}$$

Question-3

Statement

What is the equation of the decision boundary? Select all options that are correct.

Options

(a)

$$\frac{5}{21}x_1 + \frac{2}{7}x_2 = 0$$

(b)

$$\frac{5}{3}x_1 + 2x_2 = 0$$

(c)

$$7x_1 + 5x_2 = 0$$

(d)

$$\frac{1}{2}x_1 + 5x_2 = 0$$

Answer

(a), (b)

Solution

The equation of the solid line (decision boundary) is given by:

$$\frac{5}{21}x_1 + \frac{2}{7}x_2 = 0$$

We can now multiply both sides by 7 to get:

$$\frac{5}{3}x_1 + 2x_2 = 0$$

Note that this kind of scaling cannot be done with the weight vector! Scaling the weight vector alone would result in a different set of supporting hyperplanes. Moreover, the scaled weight vector would not be a solution to the primal problem.

Question-4

Statement

What is the width of the separation between the two supporting hyperplanes? Enter your answer correct to three decimal places.

Note: the exact value of the width is different from the solution to the optimization problem that is discussed in the lecture.

Answer

5.378

Range: [5.36, 5.39]

Solution

The width is given by:

$$\frac{2}{||\mathbf{w}||} = \frac{2}{\sqrt{(5/21)^2 + (2/7)^2}} \approx 5.378$$

Question-5

Statement

What is the predicted label of the black test-point?

Answer

1

Solution

We can infer this either from the graph or we can use:

$$\text{sign}(\mathbf{w}^T\mathbf{x}) = \text{sign}(5/21 \cdot 1 + 2/7 \cdot 0) = 1$$

Question-6

Statement

Is the following statement true or false?

For any test-point that falls within the region bounded by the supporting hyperplanes, no label can be assigned, as it doesn't satisfy the constraints in the optimization problem.

Options

(a)

True

(b)

False

Answer

(b)

Solution

The decision boundary is given by $\mathbf{w}^T\mathbf{x} = 0$. Though the points on the supporting hyperplanes help in determining the value of $\mathbf{w}$, they don't meddle with the prediction of points that fall within them. This depends purely on the sign of $\mathbf{w}^T\mathbf{x}$.

Question-7

Statement

SVM is a ———

Options

(a)

generative model

(b)

discriminative model

Answer

(b)

Solution

SVM is a discriminative model. There are no explicit probabilities involved, but we can present it this way:

$$P(y = 1 \mid \mathbf{x}) = \begin{cases} 1, & \mathbf{w}^T\mathbf{x} \geq 0 \\ 0, & \mathbf{w}^T\mathbf{x} < 0 \end{cases}$$

We aren't really concerned about how the point $\mathbf{x}$ was generated. Given a point, we use the line $\mathbf{w}$ to determine its label. Note that this kind of presentation is similar to what we saw for perceptrons.

Question-8

Statement

Study the similarities and differences between the following models:

(1) perceptron

(2) logistic regression

(3) SVM

Solution

Similarities:

- All three are linear models
- As a result, the decision boundary is given by $\mathbf{w}^T\mathbf{x} = 0$ for all three models
- All three are discriminative models.

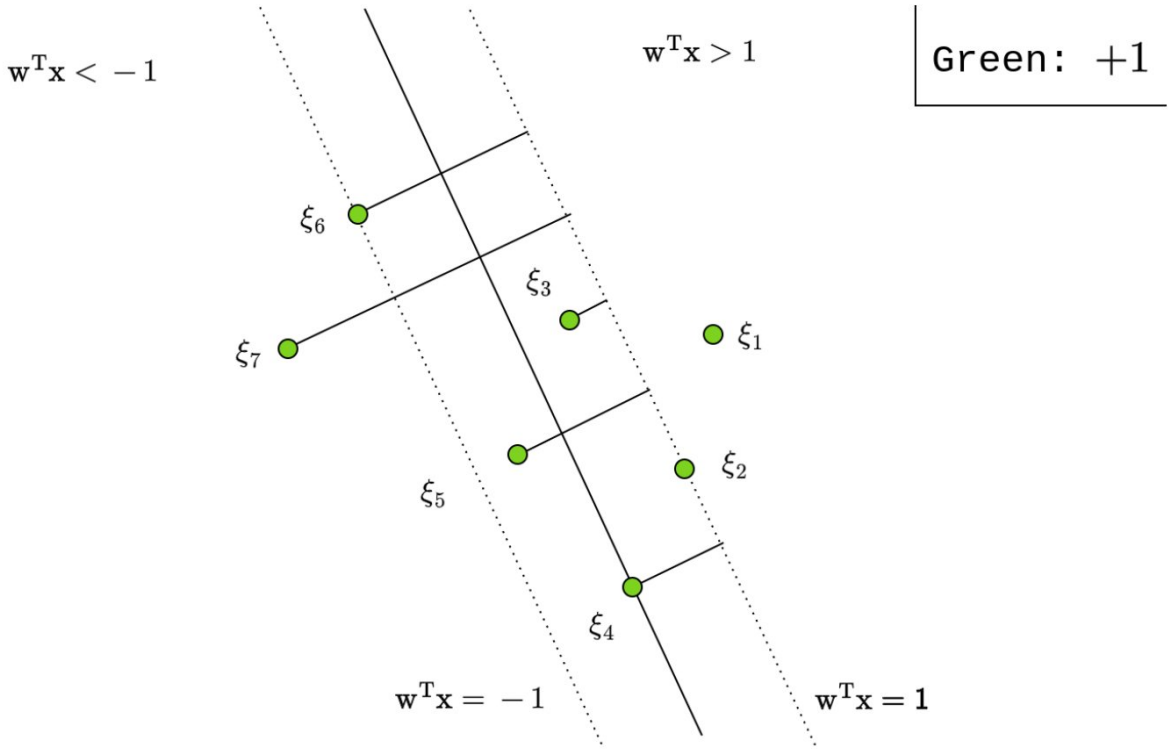
Differences:

- Logistic regression associates an explicit probability for each data-point. Farther apart a point is from the decision boundary, greater is the confidence with which LR predicts its label. This is not the case with perceptron and SVM.
- SVM uses a more principled approach compared to perceptrons. Max-margin classifiers will generalize better than perceptrons.
- The soft-margin formulation is robust to outliers, this is not the case with perceptrons which is guaranteed to converge only in the linearly separable case. Logistic regression is also more forgiving with outliers compared to perceptrons.

Common Data for Questions (9) and (13)

Statement

Consider a soft-margin SVM that has been trained on a dataset in $\mathbb{R}^2$. A subset of the data-points and the decision boundary (solid line) is shown below:



For each point, consider ξ_i to be the minimum bribe that has to be paid to take it to the correct supporting hyperplane.

Solution

For all the problems here, the basic idea is to start with the inequalities for the slack-variables:

$$\xi_i \geq 0$$
$$\xi_i \geq 1 - (\mathbf{w}^T \mathbf{x}_i) y_i$$

These correspond to the two inequalities in the soft-margin formulation for each point. Combining these two inequalities, we get:

$$\xi_i \geq \max(0, 1 - (\mathbf{w}^T \mathbf{x}_i) y_i)$$

Since we have been asked for the minimum bribe, it is going to be:

$$\xi_i = \max(0, 1 - (\mathbf{w}^T \mathbf{x}_i) y_i)$$

For points which are either on the right supporting hyperplane or beyond it, we have $(\mathbf{w}^T \mathbf{x}_i) y_i \geq 1$. As a result, we see that $\xi_i = 0$. That is, these points do not have to pay any bribe as they are already on the right side. Other points have to pay a non-zero, positive bribe. The value of the bribe can be computed by studying the geometry of the three parallel lines: $\mathbf{w}^T \mathbf{x} = 0$, $\mathbf{w}^T \mathbf{x} = -1$ and $\mathbf{w}^T \mathbf{x} = 1$.

Question-9

Statement

What is the value of ξ_1?

Answer

0

Question-10

Statement

What is the value of ξ_2?

Answer

0

Question-11

Statement

What is the value of ξ_4?

Answer

1

Question-12

Statement

What is the value of ξ_6?

Answer

2

Question-13

Statement

Select all true statements.

Options

(a)

$\xi_3 < 0$

(b)

$0 < \xi_3 < 1$

(c)

$0 < \xi_5 < 1$

(d)

$1 < \xi_5 < 2$

(e)

$\xi_7 > 2$

(f)

$\xi_7 < 0$

Answer

(b), (d), (e)

Practice

Common Instructions for questions 1-3

Consider the following data set containing one feature:

x	y
1	+1
-1	+1
-3	-1
-2	-1

Consider $w_0 = 1$ and $w_1 = 1$.

Question-1

Statement

If $\hat{y} = sign(w^T x)$, what will be the value of 0-1 loss for this data?

Answer

0 (No range required)

Solution

Zero-one loss: if $y = sign(w^T x)$, then loss = 0, else 1

Here, $w^T x = w_0 + w_1 x$

x	$w^T x$	$sign(w^T x)$	y
1	1+1 = 2	+1	+1
-1	-1+1 = 0	+1	+1
-3	-3+1 = -2	-1	-1
-2	-2+1 = -1	-1	-1

Since $sign(w^T x) = y$ for all data points, zero-one loss = 0

Question-2

Statement

What will be the value of squared loss, i.e., $(\hat{y} - y)^2$ where $\hat{y} = w^T x$?

Note: For this loss, use $y = 0$ wherever $y = -1$.

Answer

7 (No range required)

Solution

x	$\hat{y} = w^T x$	y	$(\hat{y} - y)^2$
1	1+1 = 2	1	1
-1	-1+1 = 0	1	1
-3	-3+1 = -2	0	4
-2	-2+1 = -1	0	1

Hence squared loss = 1+1+4+1 = 7

Question-3

Statement

What will be the value of hinge loss, i.e., $max(0, 1 - w^T xy)$?

Answer

1 (No range required)

Solution

x	$w^T x$	y	$w^T xy$	$1 - w^T xy$	**Hinge loss**
1	1+1 = 2	+1	2	-1	0
-1	-1+1 = 0	+1	0	1	1
-3	-3+1 = -2	-1	2	-1	0
-2	-2+1 = -1	-1	1	0	0

Total error = 0+1+0+0 = 1

Question-4

Statement

Given a cat image, you want to classify which of the 10 cat breeds it belongs to, using a neural network.

Which loss function will be appropriate?

Options

(a)

0-1 Loss

(b)

Cross entropy Loss

(c)

Squared Loss

(d)

Hinge Loss

Answer

(b)

Solution

It's a multi-class classification problem. Hence, cross entropy loss will be correct.

(0-1 loss is non-continuous and non-differentiable

Squared loss is not an appropriate loss for classification.

Hinge loss is used for binary classification.)

Question-5

Statement

Assume that $y_i = [1, 0, 0]$ and $\hat{y}_i = [0.7, 0.2, 0.1]$ for a data point x_i. What will be the value of the cross-entropy loss?

Answer

0.5146 (Range: 0.50 to 0.55)

Solution

Cross- entropy loss = $-\sum_i y_i log_2 \hat{y}_i$

$= -1 * log_2(0.7) - 0 * log_2(0.2) - 0 * log_2(0.1)$

$= -log_2(0.7)$

$= 0.5146$

Question-6

Statement

Following is the output produced by an activation function at some hidden layer of a neural network:

[0, 4.9, 0, 5.2, 7.4, 0]

Which of the following could possibly be the activation function?

Options

(a)

Sigmoid

(b)

ReLU

(c)

Tanh

Answer

(b)

Solution

Sigmoid transforms values in the range -1 to 0. So, it may not be correct.

Tanh transforms values between -1 and 1, so, it may not be correct.

ReLU transforms negative values to zero, and keeps the positive values as it is, so it may be the activation function used.

Question-7

Statement

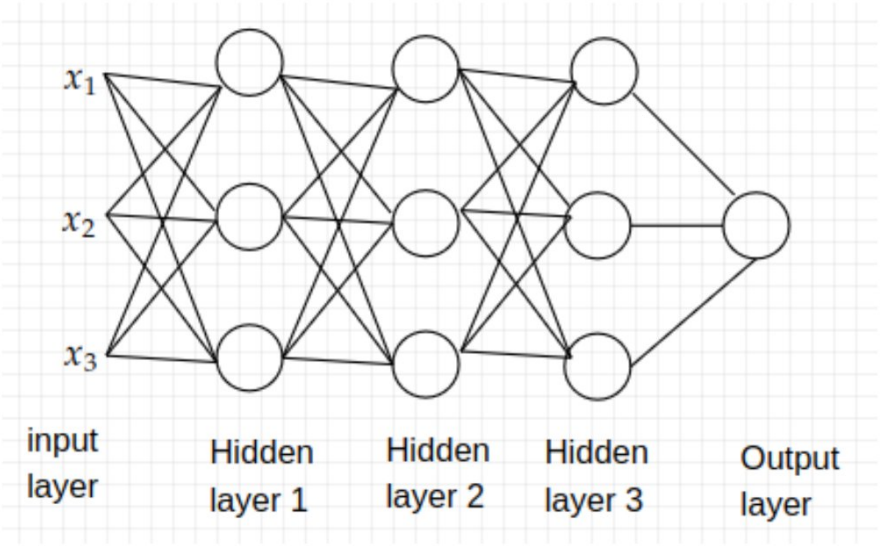
Consider a neural network with 3 inputs and one output. If there are 3 hidden layers each with 3 neurons, how many parameters need to be learnt by the back-propagation algorithm?

Note: Assume that each hidden and output layer neuron also contains a bias.

Answer

40 (No range required)

Solution



#weights from input to hidden layer 1= 3*3 = 9

#weights from hidden layer 1 to hidden layer 2= 3*3 = 9

#weights from hidden layer 2 to hidden layer 3= 3*3 = 9

#weights from hidden layer 3 to output layer= 3*1 = 3

Total number of neurons = 10

Hence, number of bias terms = 10

Therefore, total number of parameters to be computed = 9+9+9+3+10 = 40

Question-8

Statement

Which of the following is/are true?

Options

(a)

Both Sigmoid and Softmax are activation functions.

(b)

Sigmoid is used for binary classification tasks, while SoftMax applies to multiclass problems.

(c)

SoftMax function is an extension of the Sigmoid function.

(d)

Sigmoid function is also called Logistic function.

(e)

Both functions transform a real value to a number between 0 and 1.

Answer

(a), (b), (c), (d), (e)

Solution

Softmax function is a generalization of sigmoid function, to be used in multi-class classification, such that of there are k classes, the sum of probability values returned for each of these classes is equal to 1.

Graded

This document has 11 questions.

Note to Learners

Statement

The projection is treated as a vector in all the questions. If we wish to talk about the length of the projection, then that would be mentioned explicitly. Likewise, the residue after the projection is also treated as a vector. If we wish to talk about the length of the residue, that would be mentioned explicitly.

Question-1 [0.5 point]

Statement

Consider a point $\mathbf{x} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$ and a line passing through the origin which is represented by the vector $\mathbf{w} = \begin{bmatrix} 3 \\ 3 \end{bmatrix}$. What can you say about the following quantities? (MSQ)

(1) the projection of $\mathbf{x}$ onto the line

(2) the residue

Options

(a)

The residue is equal to the zero vector.

(b)

The residue is equal to the vector $\mathbf{x}$.

(c)

The projection is the zero vector.

(d)

The projection is equal to the vector $\mathbf{x}$.

Answer

(b), (c)

Solution

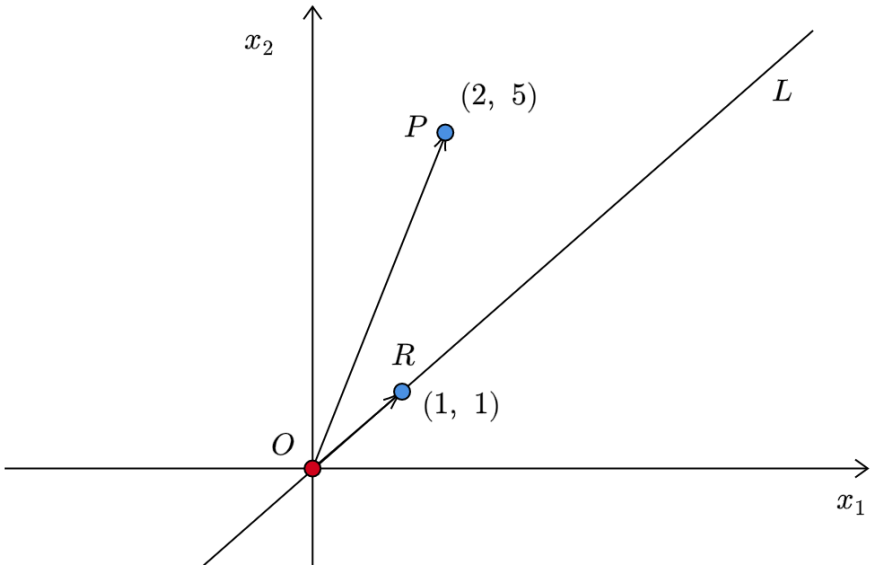
We have $\mathbf{x}^T\mathbf{w} = 0$. So, the projection is the zero vector. The residue is given by:

$$\mathbf{x} - (\mathbf{x}^T\mathbf{w})\mathbf{w} = \mathbf{x}$$

Common data for questions (2) to (5)

Statement

Consider a point P and a line L that passes through the origin O. The point R lies on the line.



We use the following notation:

$$\mathbf{w} = \overrightarrow{OR}$$

$$\mathbf{x} = \overrightarrow{OP}$$

Question-2 [1 point]

Statement

Consider the following statements:

Statement-1: The projection of $\mathbf{x}$ on the line L is given by $(\mathbf{x}^T\mathbf{w})\mathbf{w}$

Statement-2: The projection of $\mathbf{x}$ on the line L is given by $(\mathbf{x}^T\mathbf{w})\mathbf{x}$

Statement-3: The projection of $\mathbf{x}$ on the line L is given by $(\mathbf{x}^T\mathbf{x})\mathbf{w}$

Statement-4: The projection of $\mathbf{x}$ on the line L is given by $\mathbf{w}^T\mathbf{x}$

Which of the above statements is true?

Options

(a)

Statement-1

(b)

Statement-2

(c)

Statement-3

(d)

Statement-4

(e)

None of these statements are true.

Answer

(e)

Solution

The projection of a point $\mathbf{x}$ on a line $\mathbf{w}$ is given by:

$$\frac{\mathbf{x}^T\mathbf{w}}{\mathbf{w}^T\mathbf{w}}\mathbf{w}$$

This is the expression when $\mathbf{w}$ does not have unit length. In this problem, $\mathbf{w}$ does not have unit length. If $||\mathbf{w}|| = 1$, then the expression becomes:

$$(\mathbf{x}^T\mathbf{w})\mathbf{w}$$

Question-3 [1 point]

Statement

Find the length of the projection of $\mathbf{x}$ on the line L. Enter your answer correct to two decimal places.

Answer

4.95

Range: $[4.9, 5.0]$

Solution

The length of the projection is given by:

$$\frac{|\mathbf{x}^T\mathbf{w}|}{||\mathbf{w}||} = \frac{2+5}{\sqrt{2}} \approx 4.95$$

Question-4 [1 point]

Statement

Find the residue after projecting $\mathbf{x}$ on the line L.

Options

(a)

$$\begin{bmatrix} 3.5 \\ 3.5 \end{bmatrix}$$

(b)

$$\begin{bmatrix} -1.5 \\ 1.5 \end{bmatrix}$$

(c)

$$\begin{bmatrix} 2 \\ 5 \end{bmatrix}$$

(d)

$$\begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Answer

(b)

Solution

The residue is given by:

$$\mathbf{x} - \frac{\mathbf{x}^T\mathbf{w}}{\mathbf{w}^T\mathbf{w}}\mathbf{w} = \begin{bmatrix} 2 \\ 5 \end{bmatrix} - \frac{7}{2} \cdot \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -1.5 \\ 1.5 \end{bmatrix}$$

Question-5 [1 point]

Statement

Find the reconstruction error for this point. Enter your answer correct to two decimal places.

Answer

4.5

Range: $[4.4, 4.6]$

Solution

The reconstruction error is given by the square of the length of the residue. If the residue is $\mathbf{x}'$, then:

$$(\mathbf{x}')^T\mathbf{x}' = (-1.5)^2 + 1.5^2 = 4.5$$

Programming based solution. This is to be used only to verify the correctness of the calculations. The added benefit is that you get used to `NumPy`.

```
import numpy as np

x = np.array([2, 5])
w = np.array([1, 1])
w = w / np.linalg.norm(w)

# Projection
proj = (x @ w) * w
print(f'Projection = {np.linalg.norm(proj)}')
# Residue
res = x - proj
print(f'Residue = {res}')
# Reconstruction error
recon = res @ res
print(f'Reconstruction error = {recon}')
```

Question-6 [0.5 point]

Statement

Consider the following images of points in 2D space. The red line segments in one of the images represent the lengths of the residues after projecting the points on the line L. Which image is it?

Image-1

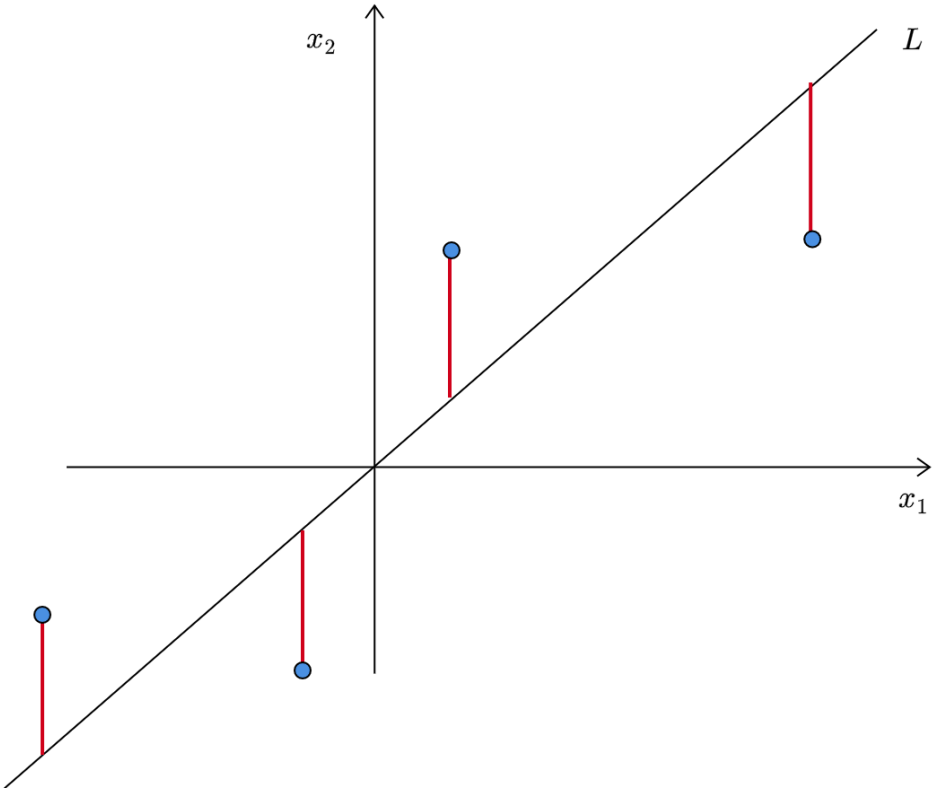
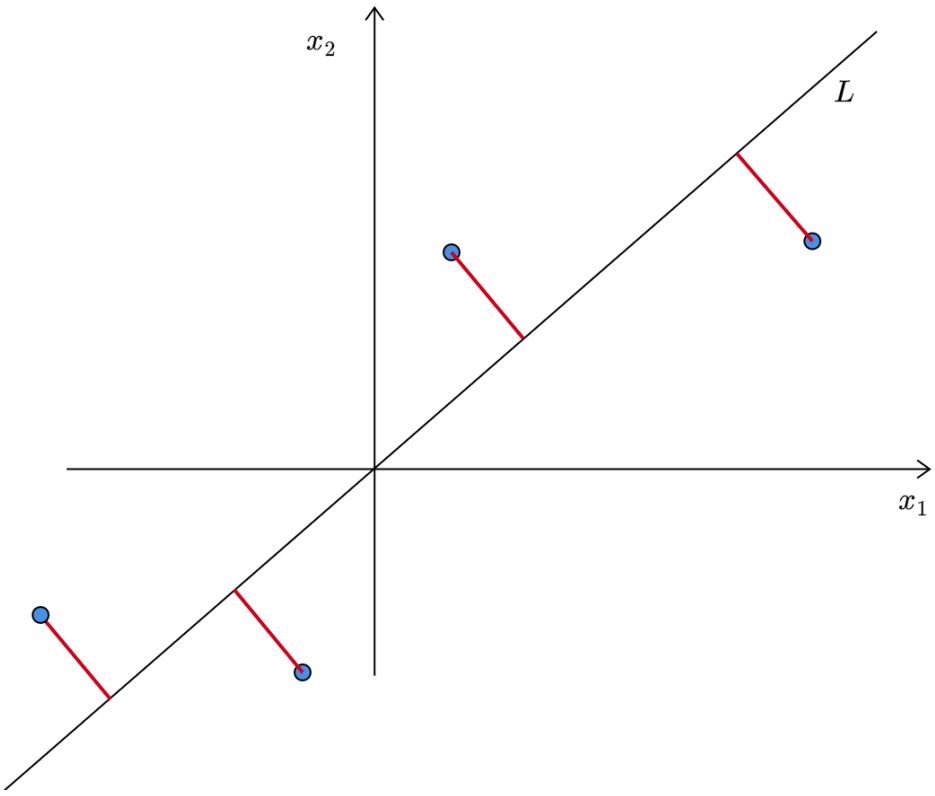


Image-2



Options

(a)

Image-1

(b)

Image-2

Answer

(b)

Solution

The residue after the projection should be perpendicular to the line. Note that by projection we mean the orthogonal projection of a point on a line. The projection of a point on a line is one of the proxies for that point on the line, in fact it is the "best" possible proxy. But every proxy does not become a projection. The projection of a point on a line is unique.

Question-7 [1 point]

Statement

Consider a dataset that has 1000 samples, where each sample belongs to $\mathbb{R}^{30}$. PCA is run on this dataset and the top 4 principal components are retained, the rest being discarded. If it takes one unit of memory to store a real number, find the percentage decrease in storage space of the dataset by moving to its compressed representation. Enter your answer correct to two decimal places; it should lie in the range $[0, 100]$.

Answer

86.27

Range: $[86.2, 86.3]$

Solution

Original space = $1000 \times 30 = 30000$

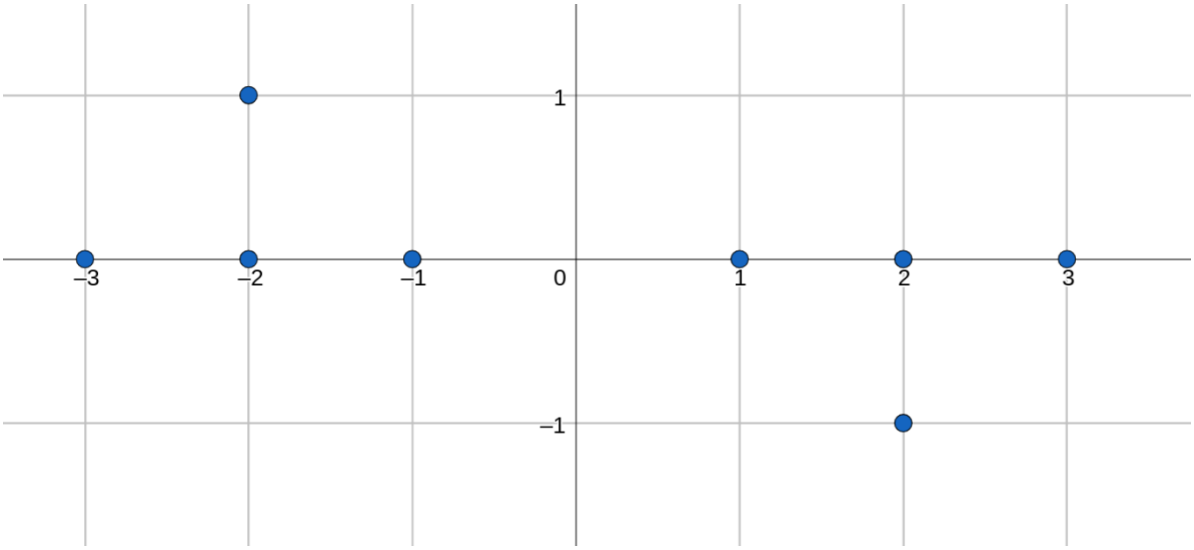
Compressed space = $1000 \times 4 + 4 \times 30 = 4120$

Percentage decrease in space = $\frac{30000 - 4120}{30000} \times 100 \approx 86.27$

Common Data for questions (8) to (9)

Statement

Consider a dataset that has 8 points all of which belong to $\mathbb{R}^2$:



Question-8 [1 point]

Statement

Find the covariance matrix of this dataset.

Options

(a)

$$\begin{bmatrix} 4.5 & -0.5 \\ -0.5 & 0.25 \end{bmatrix}$$

(b)

$$\begin{bmatrix} 36 & -4 \\ -4 & 2 \end{bmatrix}$$

(c)

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

(d)

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

Answer

(a)

Solution

Let us first arrange the data in the form of a $n \times d$ matrix. Here, $n = 8$ and $d = 2$:

$$\mathbf{X} = \begin{bmatrix} -3 & 0 \\ -2 & 0 \\ -2 & 1 \\ -1 & 0 \\ 1 & 0 \\ 2 & 0 \\ 2 & -1 \\ 3 & 0 \end{bmatrix}$$

The covariance matrix is therefore:

$$\frac{1}{n} \cdot \mathbf{X}^T \mathbf{X} = \begin{bmatrix} 4.5 & -0.5 \\ -0.5 & 0.25 \end{bmatrix}$$

Question-9 [1 point]

Statement

If PCA is run on this dataset, find the variance of the dataset along the first principal component. The eigenvectors of the covariance matrix are given below:

$$\begin{bmatrix} -0.993 \\ 0.115 \end{bmatrix}, \quad \begin{bmatrix} -0.115 \\ -0.993 \end{bmatrix}$$

Recall that the first principal component is the most important. Enter your answer correct to two decimal places.

Answer

4.55

Range: $(4.5, 4.6)$

Solution

If $(\lambda_k, \mathbf{w}_k)$ is the k^{th} eigenpair for $\mathbf{C}$, we have:

$$\lambda_k = \mathbf{w}_k^T \mathbf{C} \mathbf{w}_k$$

Of the two eigenvalues, the larger one is the answer.

Verification for these two problems:

```
import numpy as np

X = np.array([[-3, 0],
              [-2, 0],
              [-2, 1],
              [-1, 0],
              [1, 0],
              [2, 0],
              [2, -1],
              [3, 0]])

C = X.T @ X / X.shape[0]
print(f'Covariance matrix = {C}')
eigval, eigvec = np.linalg.eigh(C)
print(f'Variance = {eigval[-1]}')
```

A more detailed version. The variance of the dataset along the j^{th} principal component is σ_j^2 and is given by:

$$\sigma_j^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (\mathbf{x}_i^T \mathbf{w}_j)^2$$

$$= \mathbf{w}_j^T \left(\frac{1}{n} \cdot \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \right) \mathbf{w}_j$$

$$= \mathbf{w}_j^T \mathbf{C} \mathbf{w}_j$$

$$= \lambda_j$$

So, the variance along the j^{th} principal component is the j^{th} largest eigenvalue of the covariance matrix.

Question-10 [1 point]

Statement

Consider a dataset of 100 points all of which lie in $\mathbb{R}^5$. The eigenvalues of the covariance matrix are given below:

$$3.4, \quad 2.8, \quad 0.5, \quad 0.4, \quad 0.01$$

If we run the PCA algorithm on this dataset and retain the top-k principal components, what is a good choice of k? Use the heuristic that was discussed in the lectures.

Answer

4

Solution

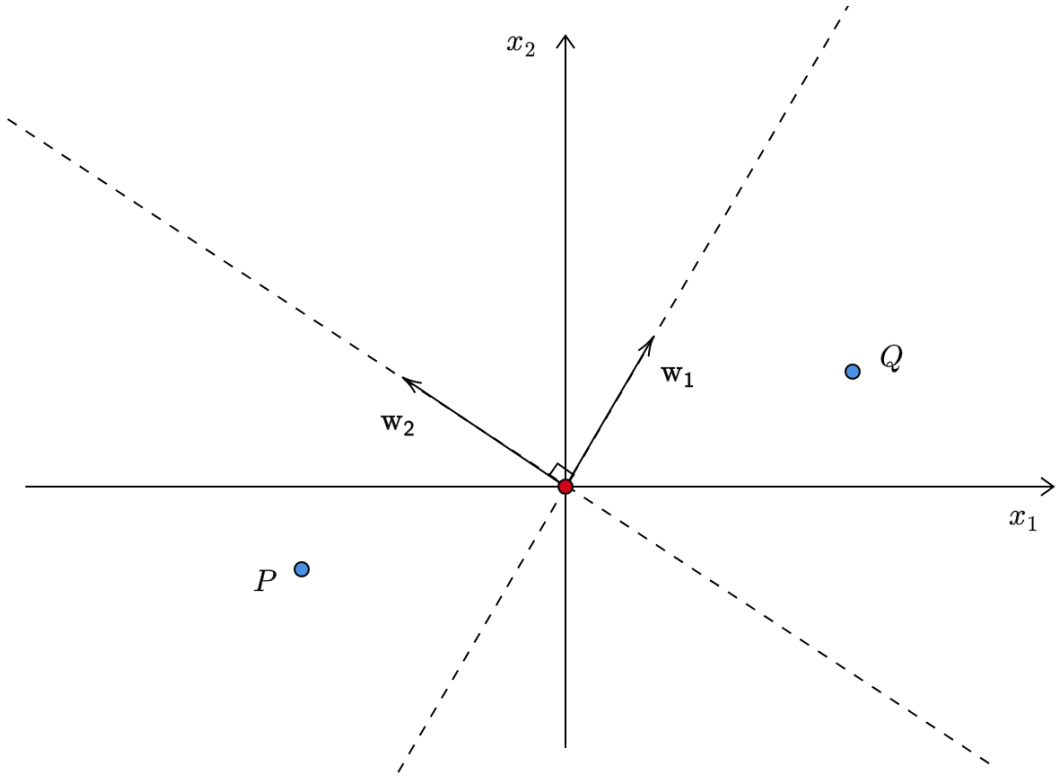
The top-k principal components should capture 95% of the variance. Here is a code snippet to answer this question:

```
L = [3.4, 2.8, 0.5, 0.4, 0.01]

den = sum(L)
for k in range(1, len(L) + 1):
    num = sum(L[: k])
    if num / den >= 0.95:
        break
print(k)
```

Question-11 [1 point]

Statement

PCA is run on a dataset that has 2 features. The resulting principal components are $\mathbf{w}_1$ and $\mathbf{w}_2$. We represent the points in 2D space in terms of this new coordinate system made up of the principal components. The first coordinate corresponds to $\mathbf{w}_1$ and the second to $\mathbf{w}_2$. In such a scenario, what would be the sign of the coordinates for the points P and Q?



Options

(a)

$P : (-ve, -ve)$

(b)

$P : (-ve, +ve)$

(c)

$Q : (+ve, +ve)$

(d)

$Q : (+ve, -ve)$

Answer

(b), (d)

Solution

Each vector $\mathbf{w}$ is associated with a line perpendicular to it. This line divides the space into two halves. The basic idea is to identify the sign of the half-planes into which the line perpendicular to the vector $\mathbf{w}$ divides the space.

Week 2 Graded assignment

Common data for Questions 1 and 2

A function k is defined as follows.

$$k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$$
$$k(x_1, x_2) = x_1^T x_2$$

Question 1

Statement

Is k a valid kernel?

Options

(a)

Yes

(b)

No

Answer

(a)

Solution

Let ϕ be an identity transformation that is

$$\phi : \mathbb{R}^d \to \mathbb{R}^d$$
$$\phi(x) = x$$

It is clear by the definition of given kernel that

$$k(x_1, x_2) = \phi(x_1)^T \phi(x_2)$$

It implies that k is a valid kernel.

Question 2

If k is the valid kernel, we apply it to the three-dimensional dataset to run the kernel PCA. Select the correct options.

Options

(a)

We cannot run the PCA as k is not a valid kernel.

(b)

It will be the same as PCA with no kernel.

(c)

It will be the same as the polynomial transformation of degree 2 and then run the PCA.

(d)

It will be the same as PCA with a third-degree polynomial kernel.

Answer

(b)

Solution

We have seen (in question 1) that k corresponds to the identity transformation. It implies that applying kernel and running PCA is same as standard PCA on the given dataset.

Question 3

Statement

Consider ten data points lying on a curve of degree two in a two-dimensional space. We run a kernel PCA with a polynomial kernel of degree two on the same data points. Choose the correct options.

Options

(a)

The transformed data points will lie on a 5-dimensional subspace of $\mathbb{R}^6$.

(b)

The transformed data points will lie on a 6-dimensional subspace of $\mathbb{R}^{10}$

(c)

There will be some $w \in \mathbb{R}^6$ that all of the data points are orthogonal to.

(d)

There will be some $w \in \mathbb{R}^{10}$ that all of the data points are orthogonal to.

Answer

(a), (c)

Solution

Since we are applying the polynomial kernel of degree two on the 2D dataset, the dataset will be transformed into a 6D feature space. (verify)

And the dataset is given to lying on a curve of degree two, the transformed dataset will live in the linear subspace of $\mathbb{R}^6$. and therefore, there will be some $w \in \mathbb{R}^6$ that all of the data points are orthogonal to.

Question 4

Statement

Which of the following matrices can not be appropriate matrix $K = X^T X$ for some data matrix X?

Options

(a)

$$\begin{bmatrix} 1 & 8 \\ 8 & -1 \end{bmatrix}$$

(b)

$$\begin{bmatrix} 1 & 8 \\ 8 & 1 \end{bmatrix}$$

(c)

$$\begin{bmatrix} 1 & 8 \\ -8 & 1 \end{bmatrix}$$

(d)

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Answer

(a), (b) and (c)

Solution

We know that K matrix must be symmetric and positive semi-definite.

All the given matrices are symmetric.

We need to check whether they are positive semi-definite or not.

For that, we will check the eigenvalues of the matrices and if all the eigenvalues are non-negative, then the matrix is positive semi-definite.

Option (a)

$$A = \begin{bmatrix} 1 & 8 \\ 8 & -1 \end{bmatrix}$$

Let λ be the eigenvalue of A, then

$$|A - \lambda I| = 0$$
$$\begin{vmatrix} 1-\lambda & 8 \\ 8 & -1-\lambda \end{vmatrix} = 0$$
$$(1-\lambda)(-1-\lambda) = 64$$
$$\lambda^2 - 1 = 64$$
$$\lambda = \pm\sqrt{65}$$

Since A has a non-negative eigenvalue, A is not a positive semi-definite matrix.

Similarly, check for all the options.

Question 5

Statement

A function k is defined as

$$k : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$$
$$k(x_1, x_2) = (x_1^T x_2)^2$$

Is k a valid kernel?

Options

(a)

Yes

(b)

No

Answer

(a)

Solution

The given function is

$$k : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$$
$$k(x_1, x_2) = (x_1^T x_2)^2$$

Let $x_1 = [a_1, a_2]^T$ and $x_2 = [b_1, b_2]^T$ then

$$
\begin{aligned}
k(x_1, x_2) &= ([a_1, a_2][b_1, b_2]^T)^2 \\
&= (a_1 b_1 + a_2 b_2)^2 \\
&= a_1^2 b_1^2 + 2 a_1 b_1 a_2 b_2 + a_2^2 b_2^2 \\
&= [a_1^2, \sqrt{2} a_1 a_2, a_2^2][b_1^2, \sqrt{2} b_1 b_2, b_2^2]^T \\
&= \phi(x_1)^T \phi(x_2)
\end{aligned}
$$

where $\phi : \mathbb{R}^2 \to \mathbb{R}$ such that $\phi([a_1, a_2]^T) = [a_1^2, \sqrt{2} a_1 a_2, a_2^2]^T$

It means there exists a transformation mapping ϕ such that $k(x_1, x_2) = \phi(x_1)^T \phi(x_2)$. Therefore, k is a valid kernel.

Question 6

Statement

Kernel PCA was run on the four data points $[1, 2]^T$, $[2, 3]^T$, $[2, -3]^T$, and $[4, 4]^T$ with the polynomial kernel of degree 2. What will be the shape of the matrix K? Notations are used as per lectures.

Options

(a)

2×2

(b)

4×4

(c)

6×6

(d)

None of the above

Answer

(b)

Solution

The K matrix is defined as $X^T X$ where X is a data matrix of shape (d, n). That is K matrix is of shape (n, n) where n is a number of examples.

It is given that $n = 4$. Therefore, shape of K matrix is $(4, 4)$

Question 7

Statement

Find the element at the index $(2, 3)$ of the matrix K defined in Question 6. Take the points in the same order.

Options

(a)

-4

(b)

16

(c)

13

(d)

196

Answer

(b)

Solution

The polynomial kernel of degree 2 is given by

$$k(x_1, x_2) = (x_1^T x_2 + 1)^2$$

The $(2, 3)th$ element of K matrix will be $k(x_2, x_3)$.

$$k(x_2, x_3) = ([2, 3][2, -3]^T + 1)^2 = (-5 + 1)^2 = 16$$

Question 8

Statement

A dataset containing 200 examples in four-dimensional space has been transformed into higher dimensional space using the polynomial kernel of degree two. What will be the dimension of transformed feature space?

Answer

15 (No range required)

Solution

Let the features be x_1, x_2, x_3, and x_4. After the transformation of degree two, features will be

$1, x_1, x_2, x_3, x_4, x_1x_2, x_1x_3, x_1x_4, x_2x_3, x_2x_4, x_3x_4, x_1^2, x_2^2, x_3^2$, and x_4^2. So, the dimension of transformed feature space will be 15.

Question 9

Statement

Let $x_1, x_2, \ldots, x_n$ be d-dimensional data points $(d > n)$ and X be the matrix of shape $d \times n$ containing the data points. The k^{th} largest eigenvalue and corresponding unit eigenvector of $X^T X$ is λ and α_k, respectively. What will be the projection of x_i on the k^{th} principal component?

Options

(a)

$x_i^T \alpha_k$

(b)

$\frac{x_i^T \alpha_k}{\lambda}$

(c)

$\frac{x_i^T X \alpha_k}{\sqrt{\lambda}}$

(d)

$\frac{x_i^T X \alpha_k}{\sqrt{n\lambda}}$

Answer

(c)

Solution

If the k^{th} largest eigenvalue and corresponding unit eigenvector of $X^T X$ is λ and α_k, respectively, the k^{th} principal component will be $\frac{X\alpha_k}{\sqrt{\lambda}}$.

Therefore, the projection of the point x_i on the k^{th} principal component will be $\frac{x_i^T X \alpha_k}{\sqrt{\lambda}}$.

Question 10

Statement

Let k_1 and k_2 be two valid kernels. Is $3k_1 + 5k_2$ a valid kernel?

Options

(a)

Yes

(b)

No

Answer

(a)

Solution

Let k_1 and k_2 be two valid kernels defined on $\mathbb{R}^d \times \mathbb{R}^d$, it implies that they satisfies the following two properties

(1) k_1 and k_2 are symmetric functions that is

$$k_1(x_1, x_2) = k_1(x_2, x_1) \quad \forall x_1, x_2 \in \mathbb{R}^d \quad \text{and}$$
$$k_2(x_1, x_2) = k_2(x_2, x_1) \quad \forall x_1, x_2 \in \mathbb{R}^d$$

(2) For any $x \in \mathbb{R}^n$, we have

$$x^T K_1 x \geq 0 \quad \text{and}$$
$$x^T K_2 x \geq 0$$

where,

$K_1 = [k_1(x_i, x_j)]_{n \times n}$ and $K_2 = [k_2(x_i, x_j)]_{n \times n}$ are K matrices corresponding to kernels k_1 and k_2, respectively.

Assume that $k = 3k_1 + 5k_2$

To show: $k(x_1, x_2) = k(x_2, x_1) \quad \forall x_1, x_2 \in \mathbb{R}^d$

and

$x^T K x \geq 0 \quad \forall x \in \mathbb{R}^n$, where $K = [k(x_i, x_j)]_{n \times n}$ is K matrices corresponding to kernel k

Now,

$$\begin{aligned} k(x_1, x_2) &= 3k_1(x_1, x_2) + 5k_2(x_1, x_2) \\ &= 3k_1(x_2, x_1) + 5k_2(x_2, x_1) \\ &= k(x_2, x_1) \end{aligned}$$

and

$$\begin{aligned} x^T K x &= x^T (3K_1 + 5K_2) x \\ &= 3x^T K_1 x + 5x^T K_2 x \geq 0 \end{aligned}$$

It implies that $3k_1 + 5k_2$ is a valid kernel.

Graded

This document has 10 questions.

Question-1

Statement

What would be the correct relationship among the following three quantities?:

- (1) $\sum_{i=1}^{n} ||x_i - \mu_{z_i^t}^t||^2$,
- (2) $\sum_{i=1}^{n} ||x_i - \mu_{z_i^{t+1}}^t||^2$ and
- (3) $\sum_{i=1}^{n} ||x_i - \mu_{z_i^{t+1}}^{t+1}||^2$

  where $\mu_{z_i^t}^t$ and $\mu_{z_i^{t+1}}^{t+1}$ refer to means of cluster z_i in iterations t and $t+1$ respectively. And $\mu_{z_i^{t+1}}^t$ is the mean of the cluster z_i where x_i is going to move in the next (i. e., $(t+1)^{th}$) iteration.

Options

(a)

(1) > (2) < (3)

(b)

(1) < (2) < (3)

(c)

(1) > (2) > (3)

(d)

(1) < (2) > (3)

Answer

(c)

Solution

The first quantity represents the value of objective function in iteration t. the third quantity represents the value of objective function in iteration $t+1$. The second quality represents an intermediate quantity which captures the distance of each data point from the mean that they will be moving towards, in the t+1 iteration. Since in every iteration, the reassignment happens only if a data point has found a closer mean, (3) will be lesser than (1). Further, since every point will want to move towards a closer cluster center in the subsequent iteration, the value of (2) will be between (1) and (3).

Question-2

Statement

Consider that in an iteration t of Lloyd's algorithm, the partition configuration (P^t) is $z_1^t, z_2^t, \ldots, z_n^t$ where each $z_i^t \in \{1, 2, \ldots, k\}$. Assume that the algorithm does not converge in iteration t, and hence some re-assignment happens, thus updating the partition configuration in the next iteration (P^{t+1}) to $z_1^{t+1}, z_2^{t+1}, \ldots, z_n^{t+1}$. How can we say that partition configuration P^{t+1} is better than P^t?

Options

(a)

The value of the objective function for P^{t+1} should be more than that for P^t

(b)

The value of the objective function for P^{t+1} should be lesser than that for P^t

(c)

The value of the objective function for P^{t+1} and P^t should be same.

Answer

(b)

Solution

Since in every iteration, the reassignment happens only if a data point has found a closer mean, P^{t+1} will be lesser than P^t.

Question-3

Statement

With respect to Lloyd's algorithm, choose the correct statements:

Options

(a)

At the end of k-means, the objective function settles in a local minima and reaching global minima may not be guaranteed.

(b)

At the end of k-means, the objective function always settles in the global minima.

(c)

The clusters produced by K-means are optimal.

(d)

If the resources are limited and the data set is huge, it will be good to prefer K-means over K-means ++.

(e)

In practice, k should be as large as possible.

Answer

(a), (d)

Solution

(a), (b) K-means may not always settle in a global minima.

(c) Finding optimal clusters is an NP-hard problem. K-means provides approximate clusters.

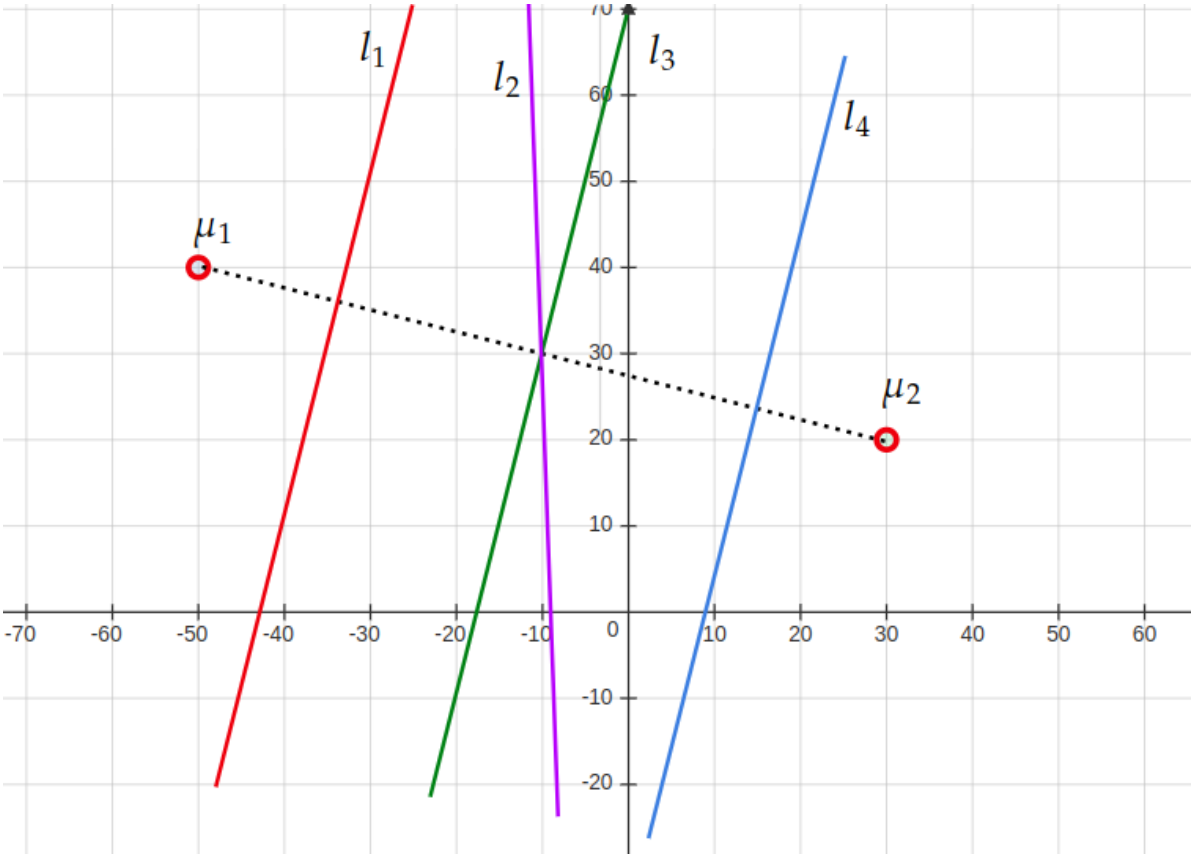
(d) If the dataset is huge. the elaborate initialization step in K-means++ will take a lot of time.

(e) In practice, k should neither be very small nor very large, because in both these cases, we may not be able to uncover groupings present in the data.

Question-4

Statement

Consider two cluster centres μ_1 and μ_2 corresponding to two clusters C_1 and C_2 as shown in the below image. Consider four half spaces represented by lines l_1, l_2, l_3 and l_4. Where would the data points falling in cluster C_1 lie?



Options

(a)

To the left of l_1

(b)

Between l_1 and l_2

(c)

Between l_3 and l_4

(d)

To the left of l_3

(e)

To the left of l_2

Answer

(d)

Solution

Half-spaces are perpendicular bisectors of the line joining the cluster centers.

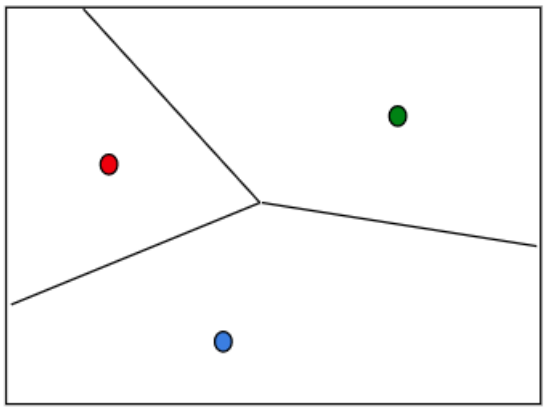
Question-5

Statement

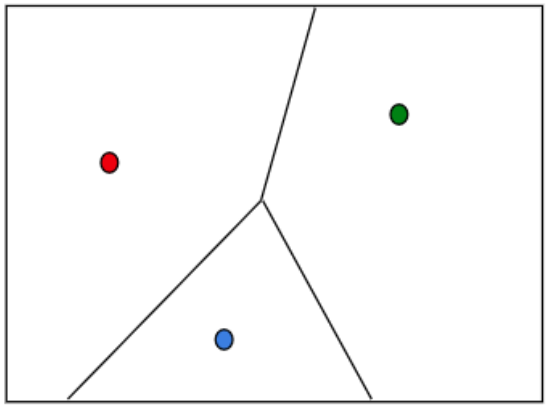
Which of the following best represents a valid voronoi diagram for K-means algorithm with K = 3? (The dots represent the cluster centres of respective clusters.)

Options

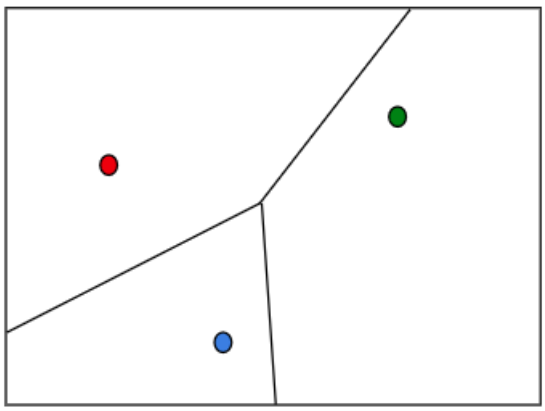
(a)



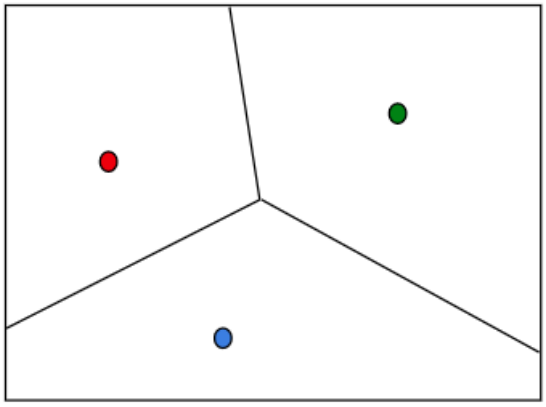
(b)



(c)



(d)



Answer

(d)

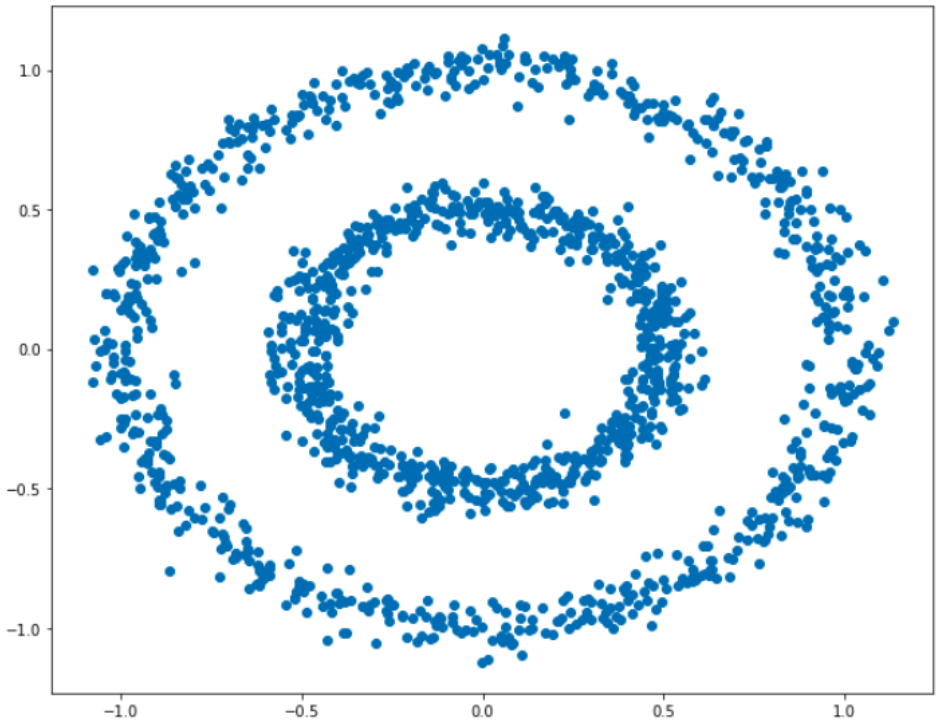
Solution

Half-spaces are perpendicular bisectors of the line joining the cluster centers.

Question-6

Statement

Consider the following data points:

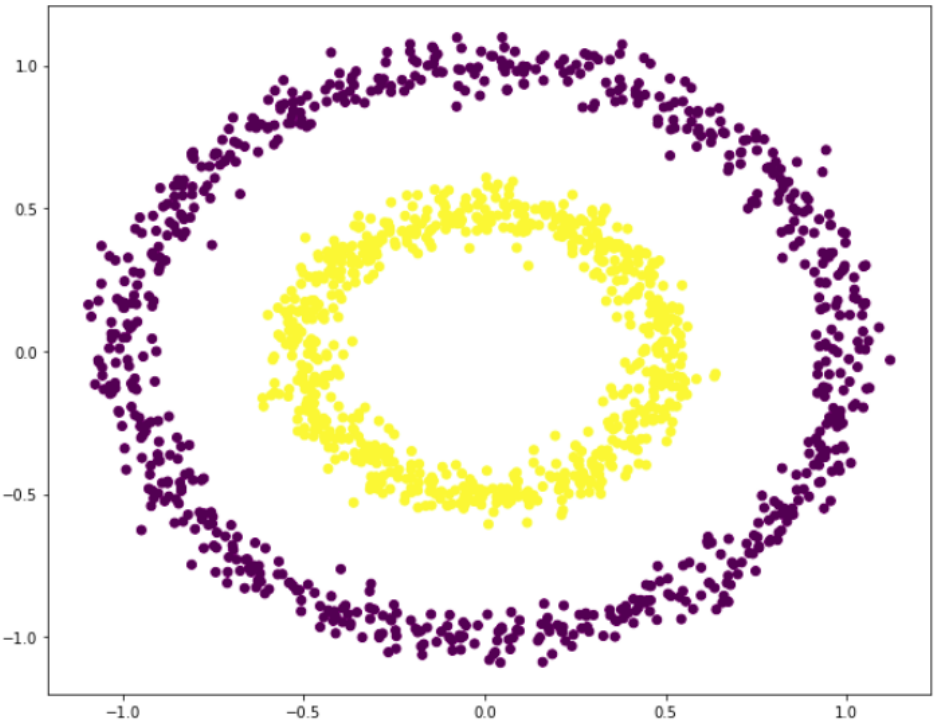


Assume that K-means is applied on this data with k = 2. Which of the following are expected to be the clusters produced?

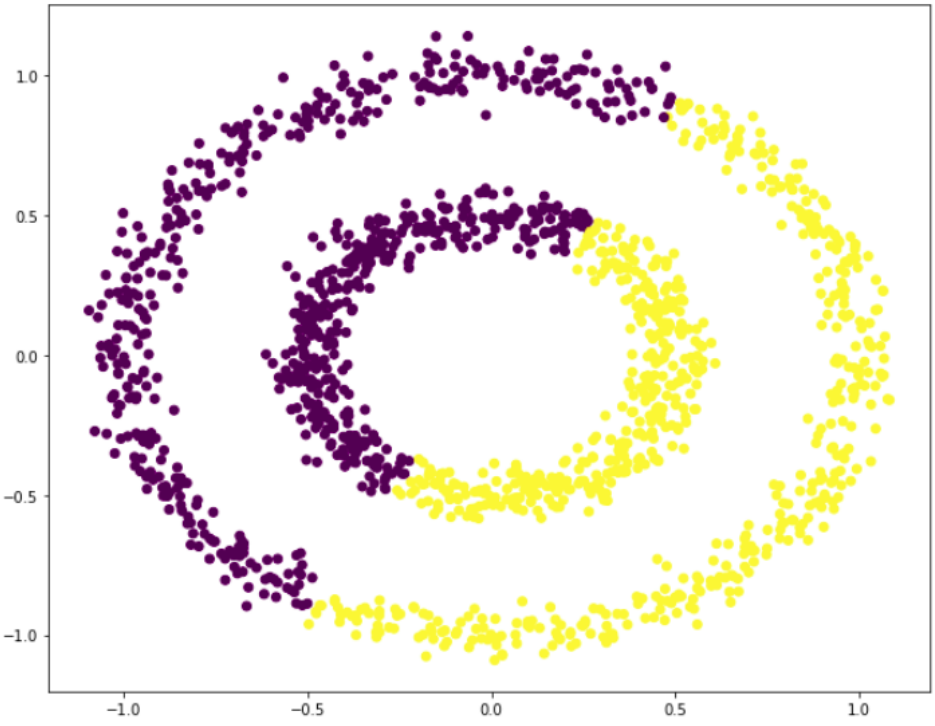
Note: Different colors represent different clusters.

Options

(a)



(b)



Answer

(b)

Solution

Half-spaces are perpendicular bisectors of the line joining the cluster centers.

In the given data, in case of option (a), the cluster centers will coincide, which is something does not happen as a result of applying k-means.

Question-7

Statement

Assume that in the initialization step of k-means++, the squared distances from the closest mean for 10 points $x_1, x_2, \ldots, x_{10}$ are: 25, 67, 89, 24, 56, 78, 90, 85, 35, 95. Which point has the highest probability of getting chosen as the next mean and how much will that probability be?

Options

(a)

x_4, 0.24

(b)

x_4, 0.037

(c)

x_{10}, 0.95

(d)

x_{10}, 0.1475

Answer

(d)

Solution

25+ 67+ 89+ 24+ 56+ 78+ 90+ 85+ 35+ 95 = 644

Probability for $x_{10} = 95/644 = 0.1475$

Probability for $x_4 = 24/644 = 0.037$

Question-8

Statement

Consider 7 data points $x_1, x_2, \ldots, x_7$: {(0, 4), (4, 0), (2, 2), (4, 4), (6, 6), (5, 5), (9, 9)}. Assume that we want to form 3 clusters from these points using K-Means algorithm. Assume that after first iteration, clusters $C1, C2, C3$ have the following data points:

$C1$: {(2,2), (4,4), (6,6)}

$C2$: {(0,4), (4,0)}

$C3$: {(5,5), (9,9)}

After second iteration, which of the clusters is the data point (2, 2) expected to move to?

Options

(a)

C_1

(b)

$C2$

(c)

C_3

(d)

Can't say, it is not deterministic.

Answer

(b)

Solution

C1: (4,4), C2: (2,2), C3: (7,7)

C2 mean is the closest to (2,2) with distance 0.

Question-9

Statement

Which of the following statements are True?

1. K-means is extremely sensitive to cluster center initializations.
2. Bad initialization can lead to poor convergence speed.
3. Bad initialization can lead to bad overall clustering.

Options

(a)

1 and 3

(b)

1 and 2

(c)

2 and 3

(d)

1, 2, and 3

Answer

(d)

Solution

1. Different cluster center initializations may result in different clusters produced by k-means.
2. Some initializations may take more time to converge.
3. Some initializations may converge either in a local minima rather than global minima.

Question-10

Statement

If the data set has two features x_1 and x_2, which of the following are true for K means clustering with k = 3?

1. If x_1 and x_2 have a correlation of 1, the cluster centres will be in a straight line.
2. If x_1 and x_2 have a correlation of 0, the cluster centres will be in straight line.

Options

(a)

1

(b)

2

(c)

None of these. Correlation does not affect cluster centres' position.

Answer

(a)

Solution

If x_1 and x_2 have a correlation of 1, all data points will lie along a line.

Hence the cluster centers will also lie along the same line.

Graded

This document has 11 questions.

Note to Learners

Statement

For all questions involving the Bernoulli distribution, the parameter p is $P(x = 1)$.

Question-1

Statement

Consider a dataset that has 10 zeros and 5 ones. What is the likelihood function if we assume a Bernoulli distribution with parameter p as the probabilistic model?

Options

(a)

$$p^{15}$$

(b)

$$(1-p)^{15}$$

(c)

$$p^{10} \cdot (1-p)^{5}$$

(d)

$$p^{5} \cdot (1-p)^{10}$$

Answer

(d)

Solution

We shall use the i.i.d. assumption. If x_i is the random variable corresponding to the i^{th} data-point, we have:

$$P(x_i = 1) = p$$

x_i and x_j are independent for $i \neq j$ and they are identically distributed. The likelihood is therefore the product of 15 terms, five of which correspond to ones and the rest to zeros:

$$L(p; D) = p^{5} \cdot (1-p)^{10}$$

Question-2

Statement

In the previous question, what is the estimate of $\hat{p}_{ML}$? Enter your answer correct to two decimal places.

Answer

0.33

Range: $[0.32, 0.34]$

Solution

The estimate is the fraction of ones:

$$\hat{p} = \frac{5}{15} = \frac{1}{3} \approx 0.33$$

Question-3

Statement

Consider a dataset that has a single feature (x). The first column in the table below represents the value of the feature, the second column represents the number of times it occurs in the dataset.

x	**Frequency**
-1	1
0	1
2	4
4	2
5	2

If we use a Gaussian distribution to model this data, find the maximum likelihood estimate of the mean.

Options

(a)

2

(b)

0

(c)

2.5

(d)

The mean cannot be computed as the variance of the Gaussian is not explicitly specified.

Answer

(c)

Solution

$$\hat{\mu}_{ML} = \frac{-1 + 0 + 4 \times 2 + 2 \times 4 + 2 \times 5}{10} = 2.5$$

Question-4

Statement

Consider a beta prior for the parameter p of a Bernoulli distribution:

$$p \sim \text{Beta}(3, 2)$$

The dataset has 15 ones and 10 zeros. What is the posterior?

Options

(a)

$\text{Beta}(3, 2)$

(b)

$\text{Beta}(30, 17)$

(c)

$\text{Beta}(18, 12)$

(d)

$\text{Beta}(17, 11)$

Answer

(c)

Solution

Since the beta distribution is a conjugate prior of the Bernoulli distribution, the posterior is also a beta distribution. Specifically, if the prior is $\text{B}(\alpha, \beta)$ and the dataset has n_1 ones and n_0 zeros, then the posterior:

$$\text{Beta}(\alpha + n_1, \beta + n_0)$$

In this problem, $\alpha = 3, \beta = 2, n_1 = 15, n_0 = 10$.

Question-5

Statement

In the previous question, we use the expected value of the posterior as a point-estimate for the parameter of the Bernoulli distribution. What is $\hat{p}$? Enter your answer correct to two decimal places.

Answer

0.6

Range: $[0.59, 0.61]$

Solution

The expected value of a beta distribution with parameters α and β is:

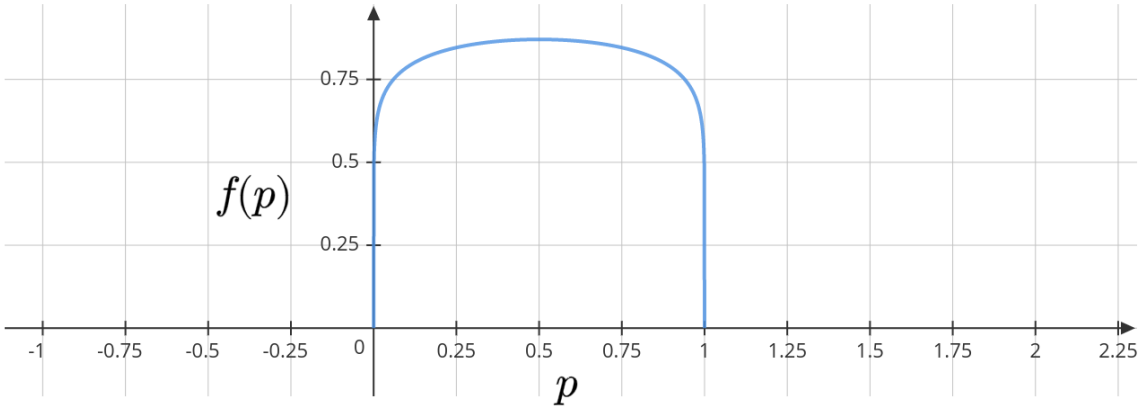
$$\frac{\alpha}{\alpha + \beta}$$

For the posterior, we have $\alpha = 18, \beta = 12$.

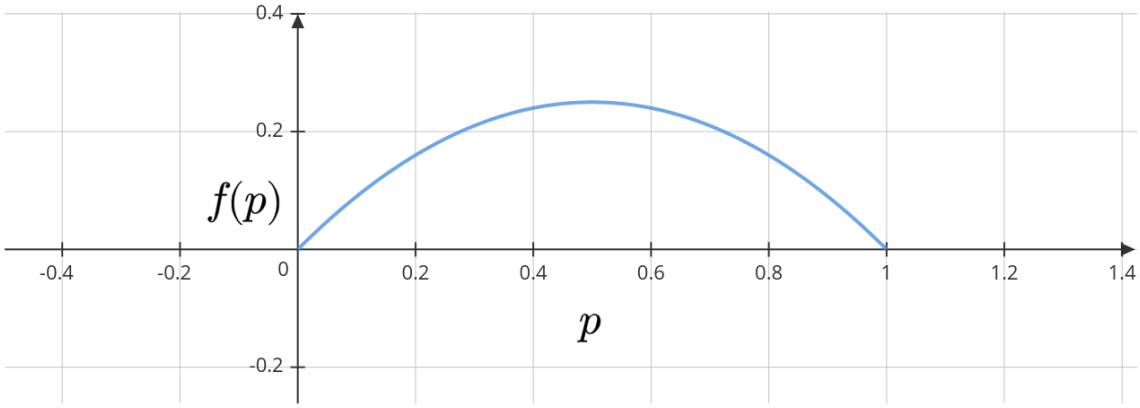
Question-6

Statement

Consider the following prior distribution (Beta) of the parameter p of a Bernoulli distribution:



After observing 10 data-points, the following is the posterior distribution:



Ignore the values on the Y-axis and just focus on the shapes of the distributions. Which of the following could correspond to the observed data?

Options

(a)

$$\{1, \quad 1, \quad 1, \quad 0, \quad 1, \quad 1, \quad 0, \quad 1, \quad 1, \quad 1\}$$

(b)

$$\{0, \quad 1, \quad 0, \quad 0, \quad 0, \quad 1, \quad 0, \quad 0, \quad 0, \quad 0\}$$

(c)

$$\{1, \quad 1, \quad 0, \quad 1, \quad 0, \quad 0, \quad 0, \quad 1, \quad 1, \quad 0\}$$

Answer

(c)

Solution

The prior encodes the belief that coin is somewhat unbiased. The posterior seems to have made that belief stronger. So, the data should have been something that strengthens the belief in the prior, meaning, an equal number of ones and zeros.

Common Data for questions (7) to (9)

Statement

We wish to fit a GMM with $K = 2$ for a dataset having 4 points. At the beginning of the t^{th} time step of the EM algorithm, we have $\theta^{(t)}$ as follows:

$$\pi_1 = 0.3, \quad \pi_2 = 0.7$$
$$\mu_1 = 2, \quad \sigma_1^2 = 1$$
$$\mu_2 = 3, \quad \sigma_2^2 = 1$$

The density of the points given a particular mixture is given to you for all four points. f is the density of a Gaussian.

x_i	$f(x_i \mid z_i = 1)$	$f(x_i \mid z_i = 2)$
1	0.242	0.054
2	0.399	0.242
3	0.242	0.399
4	0.054	0.242

Use three decimal places for all quantities throughout the questions.

Question-7

Statement

What is the value of λ_k^i for $i = 1$ and $k = 2$ after the E-step? Enter your answer correct to three decimal places.

Answer

0.342

Range: $[0.33, 35]$

Solution

From Bayes rule, we have:

$$\lambda_k^i = P(z_i = k \mid x_i) = \frac{P(z_i = k) \cdot f(x_i \mid z_i = k)}{f(x_i)}$$

So:

$$\lambda_2^1 = \frac{0.7 \times 0.054}{0.7 \times 0.054 + 0.3 \times 0.242} \approx 0.342$$

Question-8

Statement

If we pause the algorithm at this stage (after the E-step) and use the λ_k^i values to do a hard-clustering, what would be the cluster assignment? We use the following rule to come up with cluster assignments:

$$z_i = \underset{k}{\operatorname{argmax}}\, \lambda_k^i$$

The answer is in the form of a vector: $\mathbf{z} = \begin{bmatrix} z_1 & z_2 & z_3 & z_4 \end{bmatrix}^T$.

Options

(a)

$$\begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}^T$$

(b)

$$\begin{bmatrix} 2 & 2 & 2 & 2 \end{bmatrix}^T$$

(c)

$$\begin{bmatrix} 1 & 1 & 2 & 2 \end{bmatrix}^T$$

(d)

$$\begin{bmatrix} 1 & 2 & 2 & 2 \end{bmatrix}^T$$

Answer

(d)

Solution

We need to compute the table of λ_k^i values from which we can read of the cluster assignments.

x_i	λ_1^i	λ_2^i	z_i
1	0.658	0.342	1
2	0.414	0.586	2
3	0.206	0.794	2
4	0.087	0.912	2

Question-9

Statement

What is the value of μ_1 after the M-step? Enter your answer correct to three decimal places.

Answer

1.796

Range: $[1.7, 1.9]$

Solution

Now for μ_1:

$$\mu_1 = \frac{\sum_{i=1}^{4} \lambda_1^i x_i}{\sum_{i=1}^{4} \lambda_1^i}$$

Which is:

$$\mu_1 = \frac{0.658 \times 1 + 0.414 \times 2 + 0.206 \times 3 + 0.087 \times 4}{0.658 + 0.414 + 0.206 + 0.087} \approx 1.796$$

Question-10

Statement

A GMM is fit for a dataset with 5 points. At some time-step in the EM algorithm, the following are the values of λ_k^i for all points in the dataset for the k^{th} mixture after the E-step:

$$\lambda_k^1 = 0.3$$
$$\lambda_k^2 = 0.1$$
$$\lambda_k^3 = 0.4$$
$$\lambda_k^4 = 0.8$$
$$\lambda_k^5 = 0.2$$

What is the estimate of π_k after the M-step? Enter your answer correct to two decimal places.

Answer

0.36

Range: $[0.35, 0.37]$

Solution

$$\pi_k = \frac{1}{5} \cdot \sum_{i=1}^{5} \lambda_k^i = \frac{0.3 + 0.1 + 0.4 + 0.8 + 0.2}{5} = \frac{1.8}{5} = 0.36$$

Question-11

Statement

What is the value of the following expression after the E-step at time-step t in the EM algorithm? There are 100 data-points and 3 mixtures.

$$\sum_{i=1}^{100}\sum_{k=1}^{3}\lambda_k^i$$

Options

(a)

3

(b)

100

(c)

103

(d)

300

(e)

1

(f)

The answer depends on the time-step t we are at

Answer

(b)

Solution

We know that the λ_k^i values should sum to 1 for each data-point. Since there are 100 data-points, the expression should sum to 100.

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.

Graded

This document has 10 questions.

Question-1

Statement

We have a dataset of 1000 points for a classification problem using k-NN algorithm. Now consider the following statements:

S1: If $k = 10$, it is enough if we store any 10 points in the training dataset.

S2: If $k = 10$, we need to store the entire dataset.

S3: The number of data-points that we have to store increases as the size of k increases.

S4: The number of data-points that we have to store is independent of the value of k.

Options

(a)

S1 and S3 are true statements

(b)

S2 and S4 are true statements

(c)

S1 alone is a true statement

(d)

S3 alone is a true statement

(e)

S4 alone is a true statement

Answer

(b)

Solution

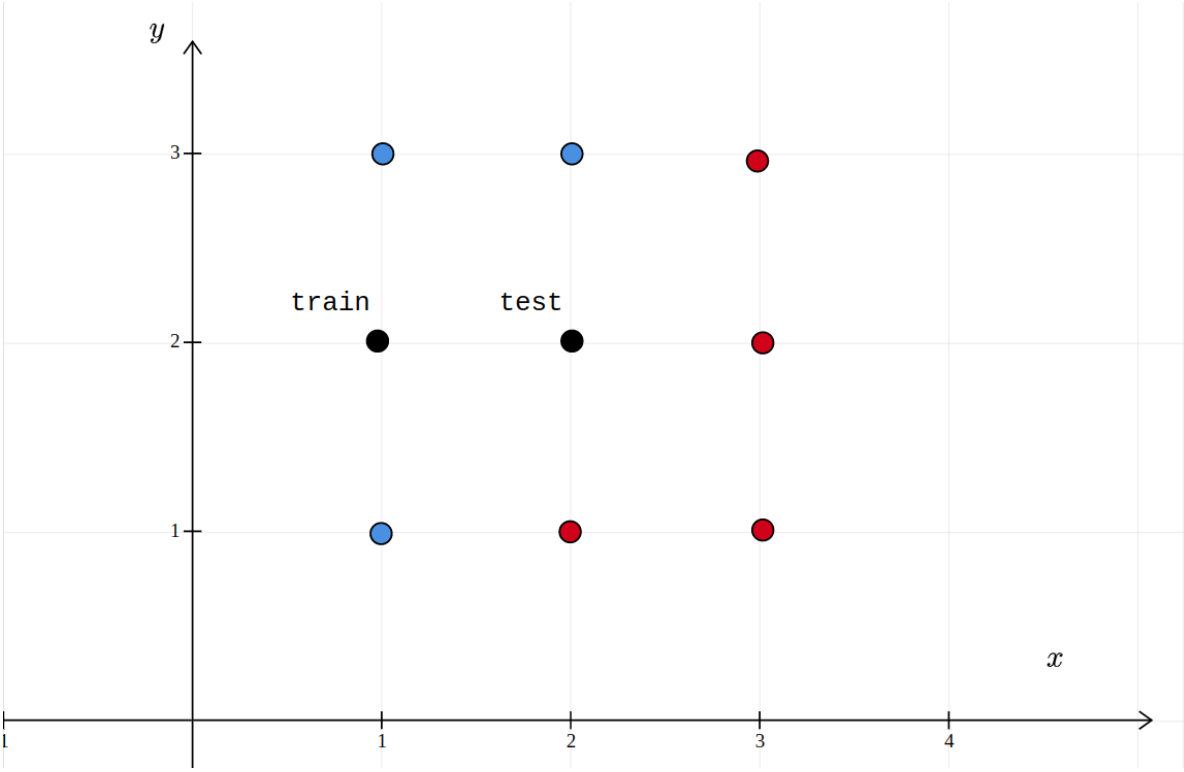
The entire training dataset has to be stored in memory. For predicting the label of a test-point, we have to perform the following steps:

- Compute distance of the test-point from each training point.
- Sort the training data points in ascending order of distance.
- Choose the first k points in this sequence.
- Return that label which garners the maximum vote among these k neighbors.

Question-2

Statement

The blue and the red points belong to two different classes. Both of them are a part of the training dataset. The black point at $(1, 2)$ also belongs to the training dataset, but its true color is hidden form our view. The black point at $(2, 2)$ is a test-point.



How should we recolor the black train point if the test point is classified as "red" without any uncertainty by a k-NN classifier, with $k = 4$? Use the Euclidean distance metric for computing distances.

Options

(a)

blue

(b)

red

(c)

Insufficient information

Answer

(b)

Solution

Since we are looking at the k-NN algorithm with $k = 4$, we need to look at the four nearest neighbors of the test data-point. The four points from the training dataset that are closest to the test data-point are the following:

- $(1, 2)$: black
- $(2, 3)$: blue
- $(3, 2)$: red
- $(2, 1)$: red

Each of them is at unit distance from the test data-point. From the problem statement, it is given that the test data-point is classified as "red" without any uncertainty. Let us now consider two scenarios that concern the black training data-point at $(1, 2)$:

Black training data-point is colored red

There are three red neighbors and one blue neighbor. Therefore, the test-data point will be classified as red. There is no uncertainty in the classification. This is what we want. However, for the sake of completeness, let us look at the alternative possibility.

Black training data-point is colored blue

There will be exactly two neighbors that are blue and two that are red. In such a scenario, we can't classify the black test-point without any uncertainty. That is, we could call it either red or blue. This is one of the reasons why we choose an odd value of k for the k-NN algorithm. If k is odd, then this kind of a tie between the two classes can be avoided.

Question-3

Statement

Consider the following feature vectors:

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 2 \\ 1 \\ -1 \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} 5 \\ -3 \\ -5 \\ 10 \end{bmatrix}, \mathbf{x}_3 = \begin{bmatrix} 3 \\ 1 \\ 2 \\ 4 \end{bmatrix}, \mathbf{x}_4 = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}, \mathbf{x}_5 = \begin{bmatrix} 10 \\ 7 \\ -3 \\ 2 \end{bmatrix}$$

The labels of these four points are:

$$y_1 = 1,\ y_2 = 0,\ y_3 = 1,\ y_4 = 0,\ y_5 = 0$$

If use a k-NN algorithm with $k = 3$, what would be the predicted label for the following test point:

$$\mathbf{x}_{\text{test}} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

Answer

1

Solution

The distances are:

- $d(\mathbf{x}_{\text{test}}, \mathbf{x}_1)^2 = 5$
- $d(\mathbf{x}_{\text{test}}, \mathbf{x}_2)^2 = 149$
- $d(\mathbf{x}_{\text{test}}, \mathbf{x}_3)^2 = 14$
- $d(\mathbf{x}_{\text{test}}, \mathbf{x}_4)^2 = 2$
- $d(\mathbf{x}_{\text{test}}, \mathbf{x}_5)^2 = 134$

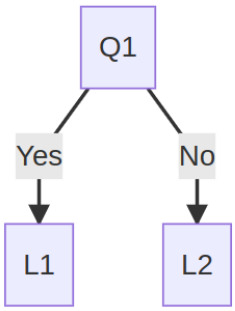
We see that among the three nearest neighbors, two have label 1 and one has label 0. Hence the predicted label is 1. For those interested in a code for the same:

```python
import numpy as np

x_1 = np.array([1, 2, 1, -1])
x_2 = np.array([5, -3, -5, 10])
x_3 = np.array([3, 1, 2, 4])
x_4 = np.array([0, 1, 1, 0])
x_5 = np.array([10, 7, -3, 2])

x_test = np.array([1, 1, 1, 1])

for x in [x_1, x_2, x_3, x_4, x_5]:
    print(round(np.linalg.norm(x_test - x) ** 2))
```

Comprehension Type (4 to 6)

Statement

Consider the following split at some node in a decision tree:



The following is the distribution of data-points and their labels:

Node	Num of points	Labels
Q1	100	0
Q1	100	1
L1	50	0
L1	30	1
L2	50	0
L2	70	1

For example, L1 has 80 points of which 50 belong to class 0 and 30 belong to class 1. Use $\log_2$ for all calculations that involve logarithms.

Question-4

Statement

If the algorithm is terminated at this level, then what are the labels associated with L1 and L2?

Options

(a)

L1 : 0

(b)

L1 : 1

(c)

L2 : 0

(d)

L2 : 1

Answer

(a), (d)

Solution

- L_1 has 80 data-points out of which 50 belong to class-0 and 30 belong to class-1. Since the majority of the points belong to class-0, this node will have 0 as the predicted label.
- L_2 has 120 data-points out of which 50 belong to class-0 and 70 belong to class-1. Since the majority of the points belong to class 1, this node will have 1 as the predicted label.

Question-5

Statement

What is the impurity in L1 if we use entropy as a measure of impurity? Report your answer correct to three decimal places.

Answer

0.954

Range: [0.94, 0.96]

Solution

If p represents the proportion of the samples that belong to class-1 in a node, then the impurity of this node using entropy as a measure is:

$$-p \log p - (1-p) \log(1-p)$$

For L_1, $p = \frac{30}{30+50} = \frac{3}{8}$. So, the impurity for L_1 turns out to be:

$$-\frac{3}{8}\log\left(\frac{3}{8}\right) - \frac{5}{8}\log\left(\frac{5}{8}\right) \approx 0.954$$

Code for reference:

```
import math
imp = lambda p: -p * math.log2(p) - (1 - p) * math.log2(1 - p)
print(imp(3 / 8))
```

Question-6

Statement

What is the information gain for this split? Report your answer correct to three decimal places. Use at least three decimal places in all intermediate computations.

Answer

0.030

Range: [0.025, 0.035]

Solution

The information gain because of this split is equal to the decrease in impurity. Here, $|L_1|$ and $|L_2|$ denote the cardinality of the leaves. N is the total number of points before the split at node Q.

$$IG = E(Q) - \left(\frac{|L_1|}{N} E(L_1) + \frac{|L_2|}{N} E(L_2) \right)$$

For this problem, the variables take on the following values:

- $N = 200$, there are 200 points in the node Q.
- $|L_1| = 80$, there are 80 points in the node L_1.
- $|L_2| = 120$, there are 120 points in the node L_2.

To calculate the entropy of the three nodes, we need the proportion of points that belong to class-1 in each of the three nodes. Let us call them p for node Q, p_1 for node L_1 and p_2 for node L_2:

- $p = \frac{100}{100 + 100} = \frac{1}{2}$
- $p_1 = \frac{30}{30 + 50} = \frac{3}{8}$
- $p_2 = \frac{70}{70 + 50} = \frac{7}{12}$

Now, we have all the data that we need to compute $E(Q)$, $E(L_1)$ and $E(L_2)$:

- $E(Q) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) = 1$
- $E(L_1) = -\frac{3}{8}\log\left(\frac{3}{8}\right) - \frac{5}{8}\log\left(\frac{5}{8}\right) \approx 0.954$
- $E(L_2) = -\frac{7}{12}\log\left(\frac{7}{12}\right) - \frac{5}{12}\log\left(\frac{5}{12}\right) \approx 0.980$

Now, we have all the values to compute the information gain:

$$IG = 1 - \left(\frac{80}{200} 0.954 + \frac{120}{200} 0.980 \right) \approx 0.030$$

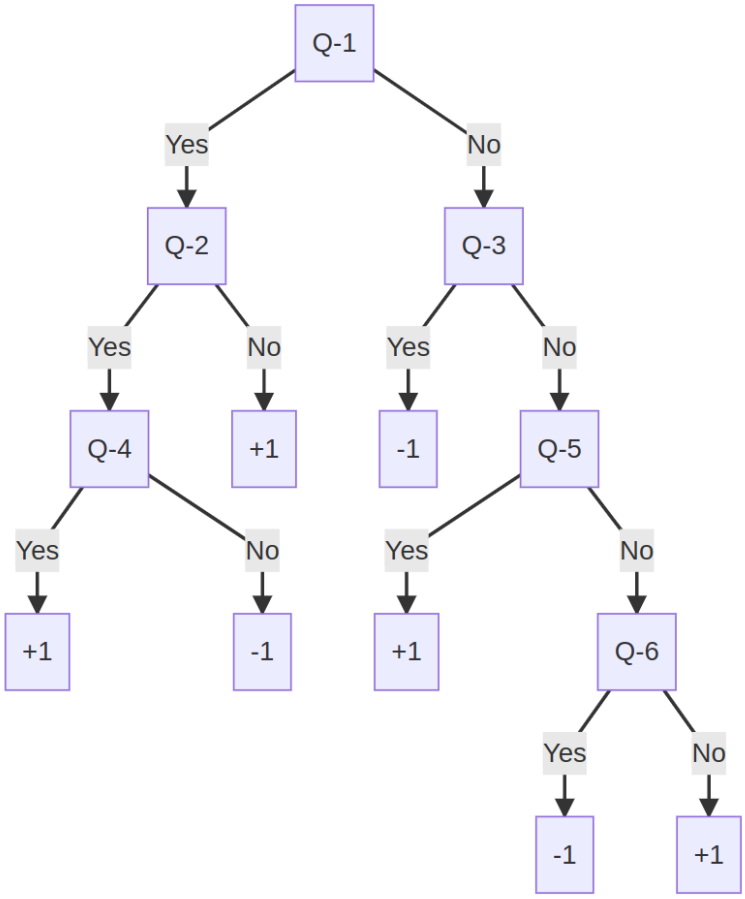
Code for reference:

```python
import math
imp = lambda p: -p * math.log2(p) - (1 - p) * math.log2(1 - p)

p_0 = 1 / 2
p_1 = 3 / 8
p_2 = 7 / 12

n = 200
l_1 = 80
l_2 = 120

ig = imp(p_0) - ((l_1 / n) * imp(p_1) + (l_2 / n) * imp(p_2))
print(ig)
```

Question-7

Statement

Consider the following decision tree. Q-i corresponds to a question. The labels are $+1$ and -1.



If a test-point comes up for prediction, what is the minimum and maximum number of questions that it would have to pass through before being assigned a label?

Options

(a)

$\min = 1$

(b)

$\min = 2$

(c)

$\min = 3$

(d)

$\max = 3$

(e)

$\max = 4$

Answer

(b), (e)

Solution

Look at all paths from the root to the leaves. Find the shortest and longest path.

Question-8

Statement

p is the proportion of points with label 1 in some node in a decision tree. Which of the following statements are true? [MSQ]

Options

(a)

As the value of p increases from 0 to 1, the impurity of the node increases

(b)

As the value of p increases from 0 to 1, the impurity of the node decreases

(c)

The impurity of the node does not depend on p

(d)

$p = 0.5$ correspond to the case of maximum impurity

Answer

(d)

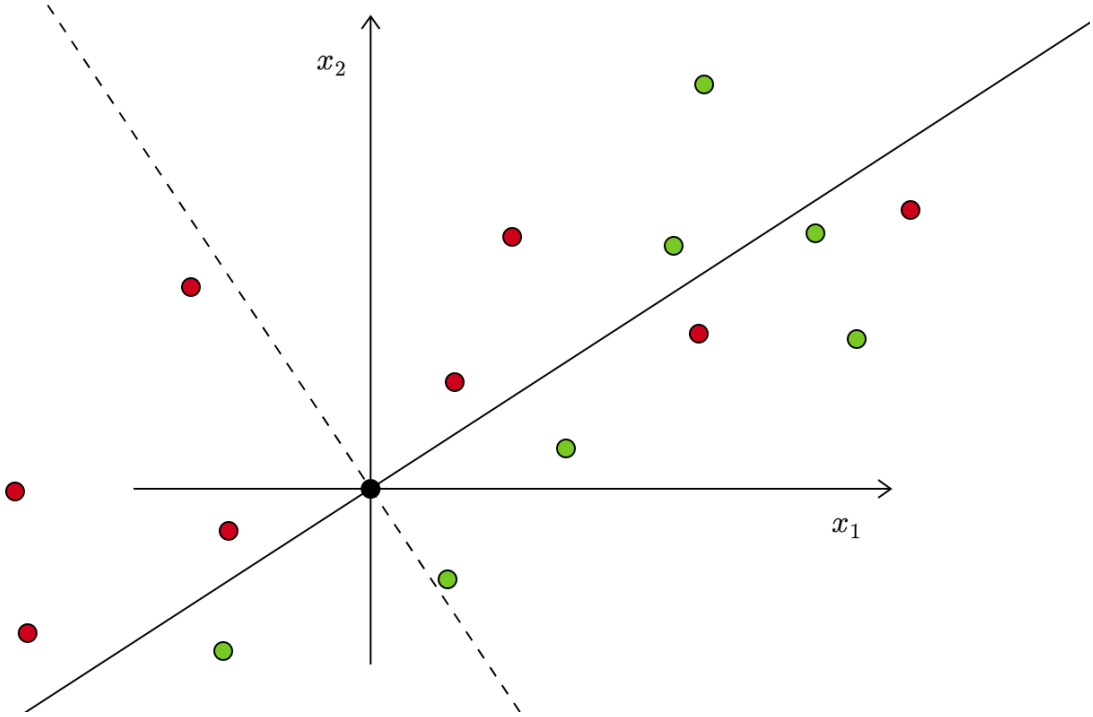
Solution

Options (a) and (b) are incorrect as the impurity increases from $p = 0$ to $p = 0.5$ and then decreases. Option-(c) is incorrect for obvious reasons.

Question-9

Statement

Consider a binary classification problem in which all data-points are in $\mathbb{R}^2$. The red points belong to class $+1$ and the green points belong to class -1. A linear classifier has been trained on this data. The decision boundary is given by the solid line.



This classifier misclassifies four points. Which of the following could be a possible value for the weight vector?

Options

(a)

$$\begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

(b)

$$\begin{bmatrix} -1 \\ 2 \end{bmatrix}$$

(c)

$$\begin{bmatrix} -1 \\ -2 \end{bmatrix}$$

(d)

$$\begin{bmatrix} 1 \\ -2 \end{bmatrix}$$

Answer

(b)

Solution

The weight vector is orthogonal to the decision boundary. So it will lie on the dotted line. This gives us two quadrants in which the vector can lie in: second or fourth. In other words, we only need to figure out its direction. If it is pointing in the second quadrant, then there will be four misclassifications. If it is pointing in the fourth quadrant then all but four points will be misclassified.

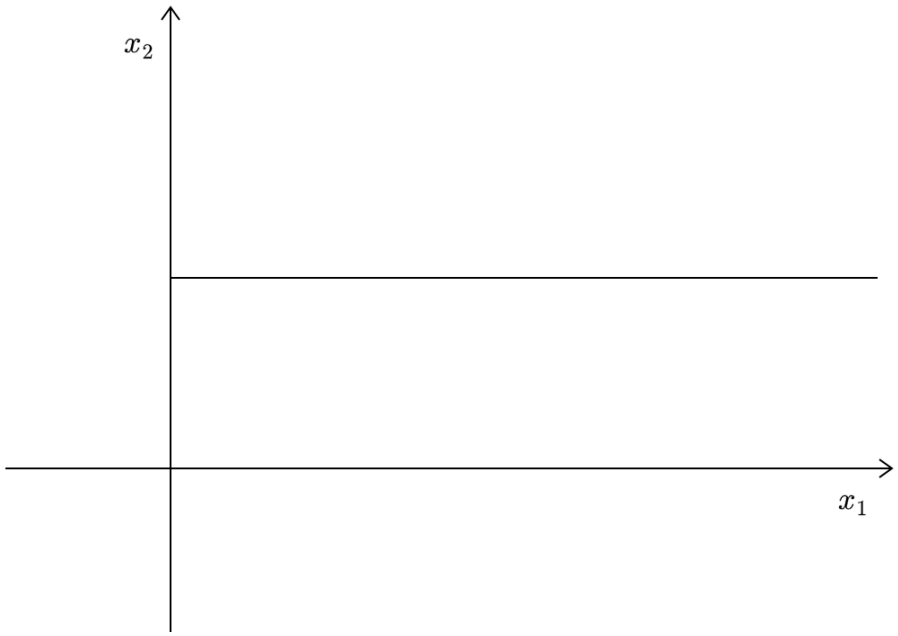
Question-10

Statement

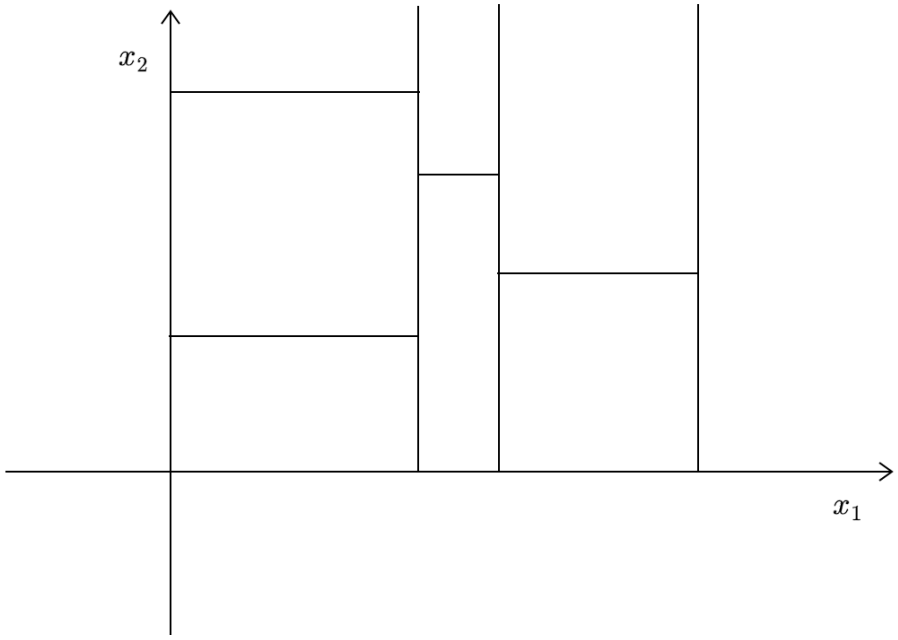
Which of the following are valid decision regions for a decision tree classifier for datapoints in $\mathbb{R}^2$? The question in every internal node is of the form $f_k \leq \theta$. Both the features are positive real numbers.

Options

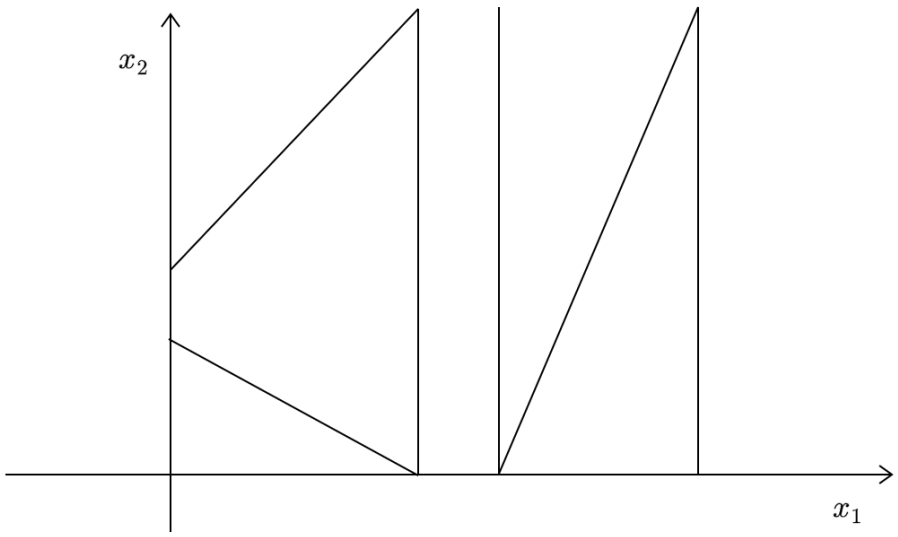
(a)



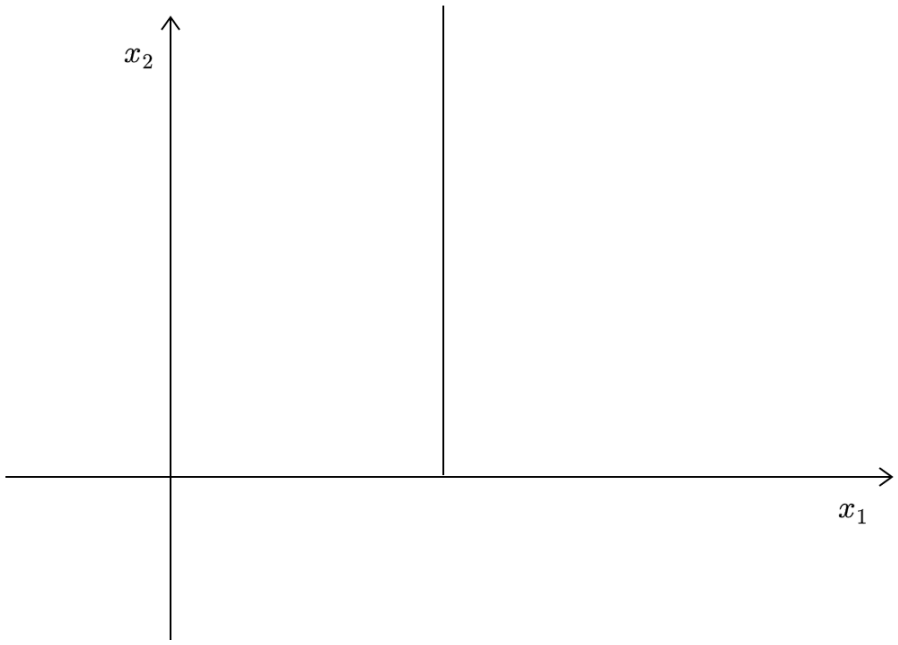
(b)



(c)



(d)



Answer

(a), (b), (d)

Solution

A question of the form $f_k \leq \theta$ can only result in one of these two lines:

- a horizontal line
- a vertical line

It cannot produce a slanted line as shown in option-(c). Options (a) and (d) correspond to what are called decision stumps: a single node splitting into two child nodes.

Graded

This document has 10 questions.

Question-1

Statement

Assume that for a certain linear regression problem involving 4 features, the following weight vectors produce an equal amount of mean square error:

w_1 = [2, 2, 3, 1]

w_2 = [1, 1, 3, 1]

w_3 = [3, 2, 4, 1]

w_4 = [1, 2, 1, 1]

Which of the weight vector is likely to be chosen by ridge regression?

Options

(a)

w_1

(b)

w_2

(c)

w_3

(d)

w_4

Answer

(d)

Solution

Total error = MSE + $\lambda||w||^2$

If MSE for all the given weights is same, the weight vector whose squared length is the least will be chosen by Ridge Regression.

Question-2

Statement

Assuming that in the constrained version of ridge regression optimization problem, following are the weight vectors to be considered, along with the mean squared error (MSE) produced by each:

w_1 = [2, 2, 3, 1], MSE = 3

w_2 = [1, 1, 3, 1], MSE = 5

w_3 = [3, 2, 4, 1], MSE = 8

w_4 = [1, 2, 1, 1], MSE = 9

If the value of θ is 13, which of the following weight vectors will be selected as the final weight vector by ridge regression?

Note: θ is as per lectures. That is, $||w||^2 \leq \theta$

Options

(a)

w_1

(b)

w_2

(c)

w_3

(d)

w_4

Answer

(b)

Solution

We need to minimize MSE such that $||w||^2 \leq \theta$

$||w_1||^2 = 18, ||w_2||^2 = 12, ||w_3||^2 = 30, ||w_4||^2 = 7$

$||w||^2 \leq 13$ for w_2 and w_4.

However, the MSE for w_2 is lesser than w_4.

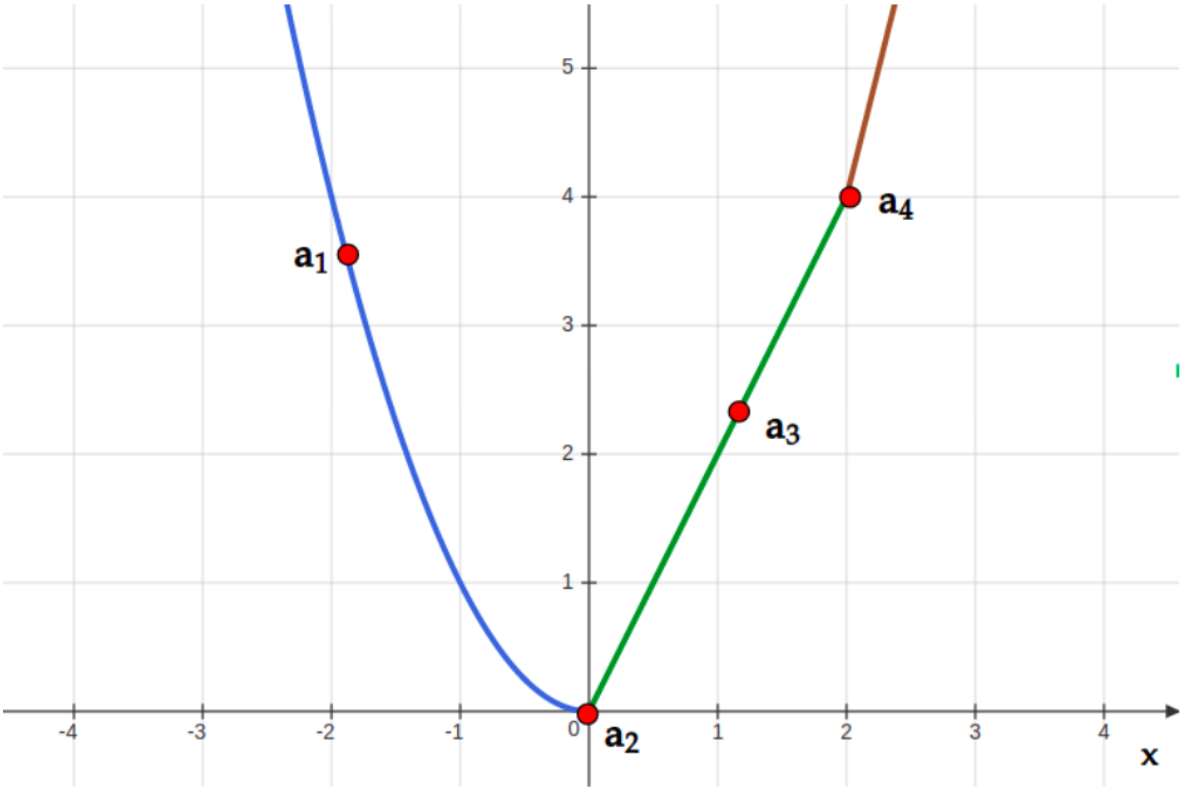
Hence, w_2 will be chosen.

Question-3

Statement

Consider the following piece-wise function as shown in the image:

$$y(x) = \begin{cases} x^2 & x \leq 0 \\ 2x & 0 \leq x \leq 2 \\ 4x - 4 & 2 \leq x \end{cases}$$



How many sub-gradients are possible at points a_1, a_2, a_3 and a_4?

Options

(a)

$a_1 : Many, a_2 : One, a_3 : Many, a_4 : One$

(b)

$a_1 : One, a_2 : Many, a_3 : Many, a_4 : One$

(c)

$a_1 : One, a_2 : Many, a_3 : One, a_4 : Many$

(d)

$a_1 : Many, a_2 : One, a_3 : One, a_4 : Many$

Answer

(c)

Solution

$a_1$lies on the part of the function which is differentiable. For a differentiable function (subpart), only one sub-gradient is possible which is the gradient itself.

a_2 lies at the intersection of two x^2 and $2x$. The function is not differentiable at this point (as left slope is different from right slope). Hence there are multiple sub-gradients possible at a_2.

a_3 lies on the part of the function which is differentiable. For a differentiable function (subpart), only one sub-gradient is possible which is the gradient itself.

a_4 lies at the intersection of two $2x$ and $4x - 4$. The function is not differentiable at this point (as left slope is different from right slope). Hence there are multiple sub-gradients possible at a_2

Question-4

Statement

For a data set with 1000 data points and 50 features, 10-fold cross-validation will perform validation of how many models?

Options

(a)

10

(b)

50

(c)

1000

(d)

500

Answer

(a)

Solution

In 10-fold cross validation, the data will be divided into 10 parts. In each of ten iterations, a model will be built using nine of these parts and the remaining part will be used for validation. Hence, in total, ten models will be validated.

Question-5

Statement

For a data set with 1000 data points and 50 features, assume that you keep 80% of the data for training and remaining 20% of the data for validation during k-fold cross-validation. How many models will be validated during cross-validation?

Options

(a)

80

(b)

20

(c)

5

(d)

4

Answer

(c)

Solution

If 20% of the data is used for validation, that means, 1/5th part is used for validation, which means, 5-fold cross validation is being performed. In each iteration, one model will be validated. Hence, total 5 models will be validated.

Question-6

Statement

For a data set with 1000 data points and 50 features, how many models will be trained during Leave-One-Out cross-validation?

Options

(a)

1000

(b)

50

(c)

5000

(d)

20

Answer

(a)

Solution

In leave one out cross-validation, only one data point is used for validation in each iteration, and the remaining n-1 data points are used for training. Hence a total of n = 1000 models will be trained.

Question-7

Statement

The mean squared error of $\hat{w}_{ML}$ will be small if

Options

(a)

The eigenvalues of XX^T are small.

(b)

The eigenvalues of $(XX^T)^{-1}$ are large.

(c)

The eigenvalues of XX^T are large.

(d)

The eigenvalues of $(XX^T)^{-1}$ are small.

Answer

(c), (d)

Solution

Mean Squared error of $\hat{w}_{ML} = \sigma^2 trace(XX^T)^{-1}$. Trace of a matrix = sum of eigenvalues.

If the eigenvalues of XX^T are large, the eigenvalues of $(XX^T)^{-1}$ will be small. Hence, trace will be small and in turn MSE will be small.

Question-8

Statement

The eigenvalues of a 3×3 matrix A are 2, 5 and 1. What will be the eigenvalues of the matrix A^{-1}

Options

(a)

4, 25, 1

(b)

2, 5, 1

(c)

0.5, 0.2, 1

(d)

0.6, 0.9, 0.1

Answer

(c)

Solution

If the eigenvalues of A are a, b and c, then the eigenvalues of A^{-1} will be 1/a, 1/b and 1/c.

Graded assignment

Question 1

Statement

Consider the two different generative model-based algorithms.

1. Model 1: chances of occurring a feature are affected by the occurrence of other features and the model does not impose any additional condition on conditional independence of features.
2. Model 2: chances of occurring a feature are not affected by the occurrence of other features and therefore, the model assumes that features are conditionally independent of the label.

Which model has more independent parameters to estimate?

Options

(a)

Model 1

(b)

Model 2

Answer

(a)

Solution:

In the first model, features are not independent, therefore, we need to find the probabilities (or density) for each and every possible example given the labels whereas in model 2, the features are independent, therefore we need to find the pmf (or pdf) of the features only.

That is the model 1 has more parameters to estimate.

Question 2

Statement

Which of the following statement is/are always correct in context to the naive Bayes classification algorithm for binary classification with all binary features? Here, $\hat{p}_j^y$ denotes the estimate for the probability that the j^{th} feature value of a data point is 1 given that the point has the label y.

Options

(a)

If $\hat{p}_j^y = 0.2$ for $y = 0$, then $\hat{p}_j^y = 0.8$ for $y = 1$

(b)

$\sum_{j=1}^{d} \hat{p}_j^y = 1$ for any y

(c)

If $\hat{p}_j^y = 0$ for $y = 0$, then $\hat{p}_j^y = 0$ for $y = 1$

(d)

If $\hat{p}_j^1 = 0$, no labeled 1 example in the training dataset takes j^{th} feature values as 1.

(e)

None of the above

Answer

(d)

Solution

In general, $\hat{p}_j^y =$ estimate for $P(f_j = 1|y)$.

It means that $\hat{p}_j^y$ denotes the parameters of different distributions $f_j|y$ for different y and for different j.

Therefore, If $\hat{p}_j^y = 0.2$ for $y = 0$, then it is not mandatory that $\hat{p}_j^y = 0.8$ for $y = 1$ as they come from different distributions.

For different j, distributions $f_j|y$ are different distributions. Therefore, it is not necessary that $\sum_{j=1}^{d} \hat{p}_j^y = 1$. What we can say is that $\sum_{i \in R(f_j)} P(f_j = i|y) = 1$

If $\hat{p}_j^y = 0$ for $y = 0$, It implies that there is no labeled zero examples such that j^{th} feature value is 1. It doesn't mean that all j^{th} feature value is 1 for all labeled one examples.

If $\hat{p}_j^1 = 0$, no labeled 1 example in the training dataset takes j^{th} feature values as 1.

Question 3

Statement

A naive Bayes model is trained on a dataset containing d features $f_1, f_2, \ldots, f_d$. Labels are 0 and 1. If a test point was predicted to have the label 1, which of the following expression should be sufficient for this prediction?

Options

(a)

$P(y=1) > P(y=0)$

(b)

$\prod_{i=1}^{d} P(f_i|y=1) > \prod_{i=1}^{d} P(f_i|y=0)$

(c)

$\left(\prod_{j=1}^{d} (\hat{p}_j^1)^{f_j}(1-\hat{p}_j^1)^{1-f_j}\right) P(y=1) > \left(\prod_{j=1}^{d} (\hat{p}_j^0)^{f_j}(1-\hat{p}_j^0)^{1-f_j}\right) P(y=0)$

(d)

None of the above

Answer

(c)

Solution

A test example is predicted label 1, it implies that

$$
\begin{aligned}
& P(y=1|x) > P(y=0|x) \\
\Rightarrow & \frac{P(x|y=1).P(y=1)}{P(x)} > \frac{P(x|y=0).P(y=0)}{P(x)} \\
\Rightarrow & P(x|y=1).P(y=1) > P(x|y=0).P(y=0) \\
\Rightarrow & \left(\prod_{j=1}^{d} (\hat{p}_j^1)^{f_j}(1-\hat{p}_j^1)^{1-f_j}\right) P(y=1) > \left(\prod_{j=1}^{d} (\hat{p}_j^0)^{f_j}(1-\hat{p}_j^0)^{1-f_j}\right) P(y=0)
\end{aligned}
$$

Question 4

Statement

Consider a binary classification dataset contains only one feature and the data points given the label follow the given distribution

$$
\begin{aligned}
x|(y=0) &\sim \mathrm{N}(0,2) \\
x|(y=1) &\sim \mathrm{N}(2,\sigma^2)
\end{aligned}
$$

If the decision boundary learned using the gaussian naive Bayes algorithm is linear, what is the value of σ^2?

Answer

2

Solution

Since the decision boundary is linear, both theiances will be the same. That is $\sigma^2 = 2$

Question 5

Statement

Consider a binary classification dataset with two binary features f_1 and f_2. The f_2 feature values are 0 for all label '0' examples but the label '1' examples take both values 1 and 0 for the feature f_2. If we apply the naive Bayes algorithm on the same dataset, what will be the prediction for point $[1, 1]^T$?

Options

(a)

Label 0

(b)

Label 1

(c)

Insufficient information to predict.

Answer

(c)

Solution

Given that the f_2 feature values are 0 for all label '0' examples it implies that $\hat{p}_2^0 = 0$.

Therefore, $p(y = 0|x) = 0$

But the label '1' examples take both values 1 and 0 for the feature f_2. It implies that $\hat{p}_2^1 > 0$.

Still the value of $p(y = 1|x)$ can be 0 if the value of $\hat{p}_1^1$ is zero. So, we need the value of $\hat{p}_1^1$ to make any conclusion.

Common data for questions 6 and 7

Statement

Consider the following binary classification dataset with two features f_1 and f_2. The data points given the labels follow the Gaussian distribution. The dataset is given as

f_1	f_2	**label** y
0.5	1.3	1
0.7	1.1	1
1.3	2.0	0
2.3	2.4	0

Question 6

Statement

What will be the value of $\hat{p}$, the estimate for $P(y = 1)$?

Answer

0.5

Solution

$$\begin{aligned}\hat{p} &= \frac{\sum_{i=1}^{n} y_i}{n} \\ &= 2/4 = 0.5\end{aligned}$$

Question 7

Statement

What will be the value of $\hat{\mu}_0$?

Options

(a)

$(1.8, 2.2)$

(b)

$(0.6, 1.2)$

(c)

$(2.0, 2.0)$

(d)

$(0.8, 1.2)$

Answer

(a)

Solution

$$\begin{aligned}\hat{\mu}_0 &= \frac{\sum_{i=1}^{n} \mathbb{1}(y_i = 0) x_i}{\sum_{i=1}^{n} \mathbb{1}(y_i = 0)} \\ &= \frac{(1.3, 2.0) + (2.3, 2.4)}{2} \\ &= (1.8, 2.2)\end{aligned}$$

Question 8

Statement

Consider a binary classification dataset containing two features f_1 and f_2. The feature f_1 is categorical which can take three values and the feature f_2 is numerical that follows the Gaussian distribution. How many independent parameters must be estimated if we apply the naive Bayes algorithm to the same dataset?

Answer

9

Solution

We need one parameter for $P(y = 1)$ as y takes only two values.

For a given label (say $y = 1$)

feature f_1 can take three values, therefore we need two estimates for $P(f_1 = 0|y = 1)$ and $P(f_1 = 1|y = 1)$.

Similarly, two estimates if $y = 0$

For feature f_2, we need μ_0, μ_1, Σ_0 and Σ_1.

Therefore, total number of parameters = $1 + 2 + 2 + 4 = 9$

Common data for questions 9 and 10

Statement

A binary classification dataset has 1000 data points belonging to $\{0,1\}^2$. A naive Bayes algorithm was run on the same dataset that results in the following estimate:

$\hat{p}$, **estimate for** $P(y=1)$	**0.3**
$\hat{p}_1^0$, estimate for $P(f_1=1\|y=0)$	0.2
$\hat{p}_2^0$, estimate for $P(f_2=1\|y=0)$	0.3
$\hat{p}_1^1$, estimate for $P(f_1=1\|y=1)$	0.1
$\hat{p}_1^2$, estimate for $P(f_2=1\|y=1)$	0.02

Question 9

Statement

What is the estimated value of $P(f_2=0|y=1)$? Write your answer correct to two decimal places.

Answer

0.98 Range: [0.97, 0.99]

Solution

$$\begin{aligned}\text{estimate for } P(f_2=0|y=1) &= 1-(\text{estimate for } P(f_2=1|y=1)) \\ &= 1-0.02=0.98\end{aligned}$$

Question 10

Statement

What will be the predicted label for the data point $[0,1]$?

Answer

0 No range is required

Solution

$$\begin{aligned}P(y=0|x) &\propto P(x|y=0).P(y=0) \\ &\propto P(f_1=0|y=0).P(f_2=1|y=0).P(y=0) \\ &\propto (1-0.2)(0.3)(1-0.3) \\ &= 0.168\end{aligned}$$

$$\begin{aligned}P(y=1|x) &\propto P(x|y=1).P(y=1) \\ &\propto P(f_1=0|y=1).P(f_2=1|y=1).P(y=1) \\ &\propto (1-0.1)(0.02)(0.3) \\ &= 0.054\end{aligned}$$

Since $P(y=0|x) > P(y=1|x)$, therefore the point will be predicted label 0.

Graded

Question-1

Statement

Assume that Perceptron algorithm is applied to a data set in which the maximum of the lengths of the data points is 4 and the value of margin (γ) of the optimal separator is 1. If the algorithm has made 10 mistakes at some point of the execution of the algorithm, which of the following can be valid squared length(s) of the weight vector obtained in the 11th iteration?

Options

(a)

90

(b)

150

(c)

170

(d)

190

Answer

(b), (c)

Solution

Given, $R = 4, \gamma = 1$

$l = 10$

Need to find w^{l+1}.

$||w^{l+1}||^2 \leq (l+1)R^2$
and
$||w^{l+1}||^2 \geq (l+1)^2\gamma^2$

Hence,

$||w^{l+1}||^2 \leq (10+1)4^2$

$\mathbf{||w^{l+1}||^2 \leq 176}$

and

$||w^{l+1}||^2 \geq (10+1)^2 1^2$

$\mathbf{||w^{l+1}||^2 \geq 121}$

Hence both (b) and (c) will be correct.

Question-2

Statement

Consider the following data set:

f_1	f_2	y
-1	-1	-1
0	1	+1
1	0	+1
1	1	+1

If Perceptron algorithm is applied on this data set with the weight vector initialized to [0, 0], how many times the weight vector will be updated during the training process?

Options

(a)

0

(b)

1

(c)

2

(d)

3

Answer

(b)

Solution

$w_0 = [0\ 0]$

$\mathbf{I_1}$:

The predictions for each of the four data points x_1, x_2, x_3, x_4 as per $w^T x$ will be +1, +1, +1, +1.

The mistake occurs for x_1

Hence, $w_1 = w_0 + x_1 y_1$

Resulting into $w_1 = [1\ 1]$

$\mathbf{I_2}$:

The predictions for each of the four data points x_1, x_2, x_3, x_4 as per $w^T x$ will be -1, +1, +1, +1, which are correct.

Hence, the weight update happened once.

Question-3

Statement

Consider the following data set with three data points:

$$\left(\begin{bmatrix}2\\2\end{bmatrix}, +1\right), \left(\begin{bmatrix}2\\-2\end{bmatrix}, +1\right), \left(\begin{bmatrix}-2\\1\end{bmatrix}, -1\right)$$

If the Perceptron algorithm is applied to this data with the initial weight vector w^0 to be a zero vector, what will be the outcome?

Options

(a)

The algorithm will converge with $w = \begin{bmatrix}-2\\1\end{bmatrix}$

(b)

The algorithm will converge with $w = \begin{bmatrix}2\\1\end{bmatrix}$

(c)

The algorithm will converge with $w = \begin{bmatrix}2\\-1\end{bmatrix}$

(d)

The algorithm will never converge.

Answer

(c)

Solution

$w_0 = [0\ 0]$

$\mathbf{I_1}$:

The predictions for each of the three data points x_1, x_2, x_3 as per $w^T x$ will be +1, +1, +1

The mistake occurs for x_3

Hence, $w_1 = w_0 + x_3 y_3$

Resulting into $w_1 = [2\ -1]$

$\mathbf{I_2}$:

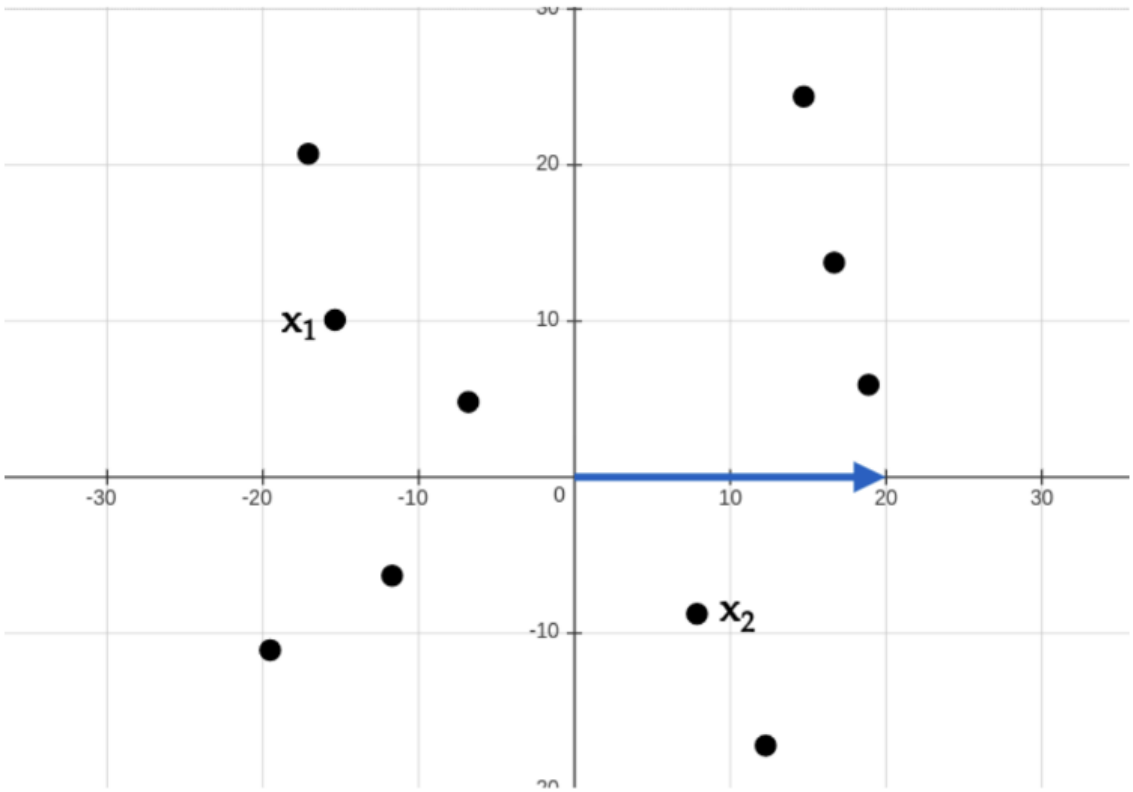
The predictions for each of the three data points x_1, x_2, x_3 as per $w^T x$ will be +1, +1, -1, which are correct.

Hence, the algorithm will converge with $w = [2\ -1]$

Question-4

Statement

Consider ten data points as shown in the following image:



The blue line represents the weight vector. As per this weight vector, the Perceptron algorithm will predict which classes for the data points x_1 and x_2?

Options

(a)

$x_1 : +1, x_2 : -1$

(b)

$x_1 : -1, x_2 : +1$

(c)

$x_1 : +1, x_2 : +1$

(d)

$x_1 : -1, x_2 : -1$

Answer

(b)

Solution

The decision boundary will be perpendicular to w. For the data points on the right side of it, $w^T x$ will be greater than equal to zero, on the LHS, it will be less than zero.

Accordingly, RHS data points (and hence x_2) will be predicted as +1.

And, LHS data points (and hence x_1) will be predicted as -1.

Question-5

Statement

In the previous question, if the weight vector is multiplied by -1, which classes will be predicted by the Perceptron for the data points x_1 and x_2?

Options

(a)

$x_1 : +1, x_2 : -1$

(b)

$x_1 : -1, x_2 : +1$

(c)

$x_1 : +1, x_2 : +1$

(d)

$x_1 : -1, x_2 : -1$

Answer

(a)

Solution

If the weight vector is multiplied by -1, then for the data points on the RHS, $w^T x$ will be less than 0 and on the LHS, it will be greater than equal to zero.

Accordingly, LHS data points (and hence x_1) will be predicted as +1.

And, RHS data points (and hence x_2) will be predicted as -1.

Question-6

Statement

Given a data set with R = 4, γ = 2, what is the maximum number of mistakes that Perceptron algorithm can make on the data?

Options

(a)

2

(b)

4

(c)

8

(d)

16

Answer

(b)

Solution

The maximum number of mistakes is given by R^2/γ^2

Which is $\frac{4^2}{2^2} = 4$

Question-7

Statement

If the scores (i.e, $w^T x$ values) for some data points are -4, 3, 1, 2, -6 respectively, what will be the probabilities returned for these points by Logistic Regression?

Options

(a)

0.25, 0.1875, 0.0625, 0.125, 0.375

(b)

-1, 1, 1, 1, -1

(c)

0.017, 0.95, 0.73, 0.88, 0.002

(d)

0, 1, 1, 1, 0

Answer

(c)

Solution

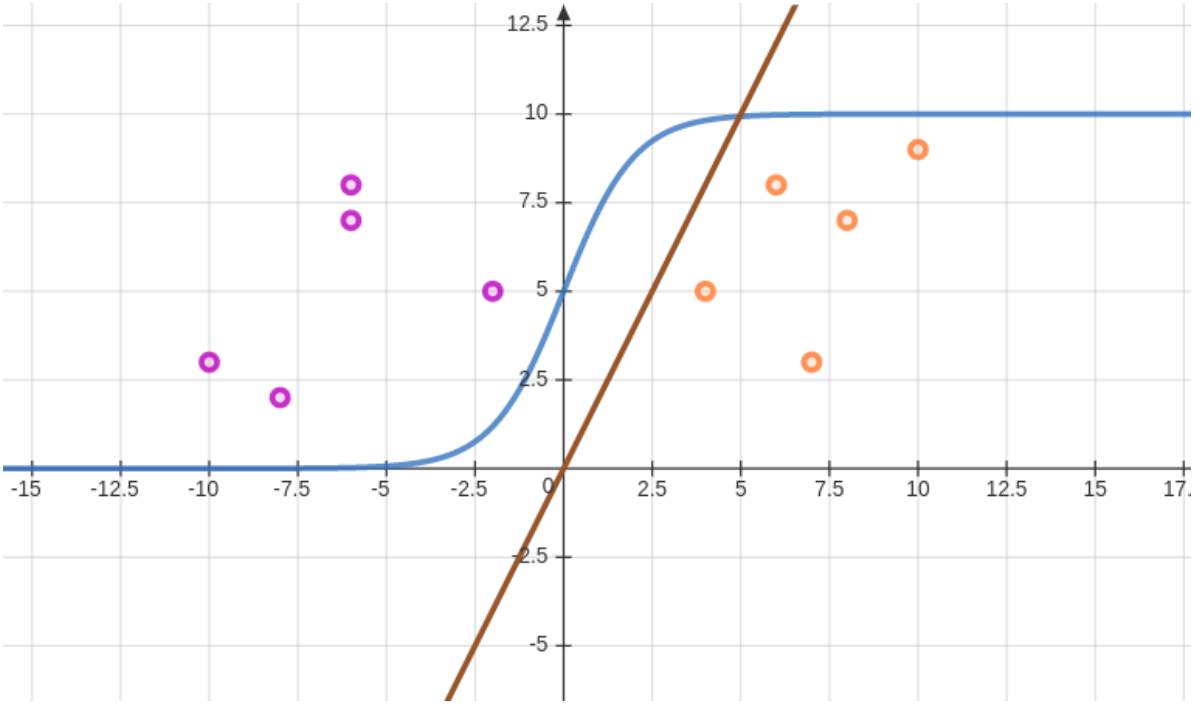
Put the values -4, 3, 1, 2, -6 in the formula $g(z) = \frac{1}{1+e^{-z}}$ one by one.

Ex: $g(-4) = \frac{1}{1+e^{4}} = \frac{1}{1+54.59} = \frac{1}{55.59} = 0.017$

Question-8

Statement

Which of the lines (blue or brown) in the following image may represent the decision boundary of Logistic Regression?



Options

(a)

Blue line

(b)

Brown line

(c)

Both

(d)

None of these

Answer

(b)

Solution

The decision boundary in logistic regression is always linear (Brown line).

It's just that the values obtained from a linear combination are reduced to values between 0 and 1 by the sigmoid function (blue line).

Question-9

Statement

Consider a data set $x_1 = [-1, -1]$, $x_2 = [-1, 0]$, $x_3 = [0, 1]$, $x_4 = [0, -1]$. Let the corresponding class labels be $y_1 = y_2 = y_4 = -1$ and $y_3 = +1$.

Assume you try to find the w using the Perceptron algorithm. You decide to cycle through points in the order $\{x_4, x_3, x_2, x_1\}$ repeatedly until you find a linear separator. How many mistakes does your algorithm make and what is the linear separator your algorithm outputs?

Options

(a)

3, [1, 0]

(b)

2, [1, 1]

(c)

3, [-1, 1]

(d)

4, [-1, 0]

Answer

(b)

Solution

When initial weight vector is not given, we will take zero vector as initial weight vector.

$w^0 = [0\ 0]$

The order in which we have to traverse the data points is given.

We start with x_4.

$w^T x_4$ predicts +1 class, which is a mistake.

Hence, $w = w + x_4 y_4$ gives $w = [0\ 1]$

$w^T x_3$ predicts +1 which is correct.

$w^T x_2$ predicts +1 which is a mistake.

Hence $w = w + x_2 y_2$ gives $w = [1\ 1]$

$w^T x_1$ predicts -1 which is correct.

Once again we check x_4, x_3, x_2, x_1 in that order, and they all are predicted correctly.

Hence final $w = [1\ 1]$ and w had been updated twice.

Graded

This document has 8 questions.

Question-1 [1 point]

Statement

Consider a linearly separable dataset for a binary classification problem in $\mathbb{R}^d$. Three linear classifiers have been trained on this dataset. All three pass through the origin and have the following property:

$$(\mathbf{w}_j^T \mathbf{x}_i) \cdot y_i \geq 1, \quad 1 \leq i \leq n$$

Here, $\mathbf{w}_j$ is the weight vector corresponding to the j^{th} classifier. Note that the above property is satisfied for each of the n data-points. If $\mathbf{w}_1$ is the weight vector corresponding to a hard-margin SVM, which of the following statements is always true? You can assume that the norms of all three weights are different from each other.

Options

(a)

$$||\mathbf{w}_1|| > ||\mathbf{w}_2|| > ||\mathbf{w}_3||$$

(b)

$$||\mathbf{w}_1|| < ||\mathbf{w}_2|| < ||\mathbf{w}_3||$$

(c)

$$||\mathbf{w}_1|| < ||\mathbf{w}_2|| \text{ and } ||\mathbf{w}_1|| < ||\mathbf{w}_3||$$

(d)

$$||\mathbf{w}_1|| > ||\mathbf{w}_2|| \text{ and } ||\mathbf{w}_1|| > ||\mathbf{w}_3||$$

Answer

(c)

Solution

$\mathbf{w}_1$ will have the smallest norm (maximum margin) among the three classifiers. The three weight vectors are feasible points for the primal. Among them, $\mathbf{w}_1$ is optimal.

Common Data for questions (2) to (4)

Statement

Common data for questions (2) to (4)

Consider the following training dataset for a binary classification problem in $\mathbb{R}^2$. Each data-point is represented by $\mathbf{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix}^T$ whose label is y.

Index	x_1	x_2	y
1	1	0	1
2	-1	0	-1
3	5	4	1
4	-5	-4	-1

We wish to train a hard-margin SVM for this problem. $\mathbf{w} = \begin{bmatrix} w_1 & w_2 \end{bmatrix}^T$ represents the weight vector. The index i is for the i^{th} data-point. α_i is the Lagrange multiplier for the i^{th} data-point.

Question-2 [1 point]

Statement

Select all primal constraints from the options given below.

Options

(a)

$$w_1 \geq 1$$

(b)

$$w_1 \leq 1$$

(c)

$$5w_1 + 4w_2 \geq 1$$

(d)

$$5w_1 + 4w_2 \leq 1$$

Answer

(a), (c)

Solution

Because of the symmetry in the problem, we effectively have only two constraints even though there are 4 data-points:

$$\begin{aligned} w_1 &\geq 1 \\ 5w_1 + 4w_2 &\geq 1 \end{aligned}$$

But in order to remain consistent with our formulation, let us list it down in the following manner:

$$\begin{aligned} w_1 &\geq 1 \quad (1) \\ w_1 &\geq 1 \quad (2) \\ 5w_1 + 4w_2 &\geq 1 \quad (3) \\ 5w_1 + 4w_2 &\geq 1 \quad (4) \end{aligned}$$

Question-3 [2 points]

Statement

Which of the following is the objective function of the dual problem? In all options, $\boldsymbol{\alpha} = \begin{bmatrix}\alpha_1 & \alpha_2 & \alpha_3 & \alpha_4\end{bmatrix}^T$ and $\mathbf{1} = \begin{bmatrix}1 & 1 & 1 & 1\end{bmatrix}^T$.

Options

(a)

$$\boldsymbol{\alpha}^T\mathbf{1} - \frac{1}{2} \cdot \boldsymbol{\alpha}^T\boldsymbol{\alpha}$$

(b)

$$\boldsymbol{\alpha}^T\mathbf{1} - \frac{1}{2} \cdot \boldsymbol{\alpha}^T \begin{bmatrix}1 & 1 & 5 & 5\\ 1 & 1 & 5 & 5\\ 5 & 5 & 41 & 41\\ 5 & 5 & 41 & 41\end{bmatrix} \boldsymbol{\alpha}$$

(c)

$$\boldsymbol{\alpha}^T\mathbf{1} - \frac{1}{2} \cdot \boldsymbol{\alpha}^T \begin{bmatrix}5 & 5 & 1 & 1\\ 5 & 5 & 1 & 1\\ 1 & 1 & 10 & 41\\ 1 & 1 & 41 & 10\end{bmatrix} \boldsymbol{\alpha}$$

(d)

$$\boldsymbol{\alpha}^T\mathbf{1} - \frac{1}{2} \cdot \boldsymbol{\alpha}^T \begin{bmatrix}1 & 1 & 30 & 30\\ 1 & 1 & 30 & 30\\ 5 & 5 & 10 & 10\\ 5 & 5 & 10 & 10\end{bmatrix} \boldsymbol{\alpha}$$

Answer

(b)

Solution

The objective function corresponding to the dual is:

$$\boldsymbol{\alpha}^T\mathbf{1} - \frac{1}{2} \cdot \boldsymbol{\alpha}^T(\mathbf{Y}^T\mathbf{X}^T\mathbf{X}\mathbf{Y})\boldsymbol{\alpha}$$

We have:

$$\mathbf{X} = \begin{bmatrix}1 & -1 & 5 & -5\\ 0 & 0 & 4 & -4\end{bmatrix}, \mathbf{Y} = \begin{bmatrix}1 & 0 & 0 & 0\\ 0 & -1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & -1\end{bmatrix}$$

Therefore:

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 1 & -1 & 5 & -5 \\ -1 & 1 & -5 & 5 \\ 5 & -5 & 41 & -41 \\ -5 & 5 & -41 & 41 \end{bmatrix}$$

And:

$$\mathbf{X}^T\mathbf{X}\mathbf{Y} = \begin{bmatrix} 1 & 1 & 5 & 5 \\ -1 & -1 & -5 & -5 \\ 5 & 5 & 41 & 41 \\ -5 & -5 & -41 & -41 \end{bmatrix}$$

Finally:

$$\mathbf{Y}^T\mathbf{X}^T\mathbf{X}\mathbf{Y} = \begin{bmatrix} 1 & 1 & 5 & 5 \\ 1 & 1 & 5 & 5 \\ 5 & 5 & 41 & 41 \\ 5 & 5 & 41 & 41 \end{bmatrix}$$

Question-4 [1 point]

Statement

What is the optimal weight vector, $\mathbf{w}^*$?

Hint: Plot the points and try to compute the answer using geometry; do not try to solve the dual algebraically!

Options

(a)

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

(b)

$$\begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

(c)

$$\begin{bmatrix} -1 \\ 0 \end{bmatrix}$$

(d)

$$\begin{bmatrix} 0 \\ -1 \end{bmatrix}$$

Answer

(a)

Solution

We see that the x_2 axis has to be the optimal separator as that is the one which has maximum margin. So, the equation of the decision boundary is $x_1 = 0$. This implies, $w_2 = 0$. Therefore, the weight vector becomes $\begin{bmatrix} w_1 & 0 \end{bmatrix}^T$. The choice of w_1 can be found out by noticing that $x_1 = 1$ and $x_1 = -1$ are the supporting hyperplanes. Therefore, $w_1 \cdot 1 = 1 \implies w_1 = 1$.

Question-5 [1 point]

Statement

Consider a kernel-SVM trained on a dataset of 100 points with polynomial kernel of degree 2. If $\boldsymbol{\alpha}^*$ is the optimal dual solution, what is the predicted label for a test-point $\mathbf{x}_{\text{test}}$?

Options

(a)

$$\sum_{i=1}^{100} \alpha_i^* \cdot \mathbf{x}_{\text{test}}^T \mathbf{x}_i \cdot y_i$$

(b)

$$\text{sign}\left(\sum_{i=1}^{100} \alpha_i^* \cdot \mathbf{x}_{\text{test}}^T \mathbf{x}_i \cdot y_i\right)$$

(c)

$$\sum_{i=1}^{100} \alpha_i^* \cdot (1 + \mathbf{x}_{\text{test}}^T \mathbf{x}_i)^2 \cdot y_i$$

(d)

$$\text{sign}\left(\sum_{i=1}^{100} \alpha_i^* \cdot (1 + \mathbf{x}_{\text{test}}^T \mathbf{x}_i)^2 \cdot y_i\right)$$

Answer

(d)

Solution

The optimal weight vector is given by:

$$\mathbf{w}^* = \sum_{i=1}^{100} \alpha_i^* \cdot \phi(\mathbf{x}_i) \cdot y_i$$

Here, $\phi(\mathbf{x}_i)$ is the vector in the transformed space. First we compute the dot-product:

$$\mathbf{w}^{*T} \phi(\mathbf{x}_{\text{test}}) = \sum_{i=1}^{100} \alpha_i^* \cdot \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_{\text{test}}) \cdot y_i$$

$$= \sum_{i=1}^{100} \alpha_i^* \cdot k(\mathbf{x}_i, \mathbf{x}_{\text{test}}) \cdot y_i$$

$$= \sum_{i=1}^{100} \alpha_i^* \cdot (1 + \mathbf{x}_{\text{test}}^T \mathbf{x}_i) \cdot y_i$$

Finally, the prediction is:

$$\text{sign}\left(\mathbf{w}^{*T}\phi(\mathbf{x}_{\text{test}})\right) = \text{sign}\left(\sum_{i=1}^{100} \alpha_i^* \cdot (1 + \mathbf{x}_{\text{test}}^T \mathbf{x}_i)^2 \cdot y_i\right)$$

Common Data for questions (6) and (7)

Statement

Common data for questions (6) and (7)

Consider the transformation $\phi : \mathbb{R}^2 \rightarrow \mathbb{R}^6$ associated with the polynomial kernel with degree 2:

$$\phi(\mathbf{x}) = \begin{bmatrix} 1 \\ x_1^2 \\ x_2^2 \\ \sqrt{2}x_1x_2 \\ \sqrt{2}x_1 \\ \sqrt{2}x_2 \end{bmatrix}$$

A kernel-SVM is trained on a dataset with the above kernel. The optimal weight vector is as follows:

$$\mathbf{w}^* = \begin{bmatrix} -25 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

You can assume that the dataset is linearly separable in the transformed space.

Question-6 [1 point]

Statement

What is the shape of the decision boundary in $\mathbb{R}^2$?

Options

(a)

It is a parabola of the form $x_2 = 25x_1^2$

(b)

It is a parabola of the form $x_2 = -25x_1^2$

(c)

It is a straight line

(d)

It is a circle

Answer

(d)

Solution

The decision boundary in $\mathbb{R}^2$ is given by:

$$\mathbf{w^*}^T \phi(\mathbf{x}) = 0$$

$$x_1^2 + x_2^2 = 25$$

It is a circle centered at the origin with radius 5.

Question-7 [2 points]

Statement

Which of the following training data-points are certainly **not** support vectors?

Options

(a)

$[-1 \quad 5]^T$

(b)

$[3 \quad 5]^T$

(c)

$[-4 \quad -5]^T$

(d)

$[2\sqrt{5} \quad -2]^T$

(e)

$[\sqrt{30} \quad \sqrt{6}]^T$

Answer

(b), (c), (e)

Solution

All the support vectors will lie on the two supporting hyperplanes:

$$\mathbf{w^*}^T\phi(\mathbf{x}) = \pm 1$$

This gives two curves:

$$x_1^2 + x_2^2 = 24 \quad \text{OR} \quad x_1^2 + x_2^2 = 26$$

These curves are two circles, one smaller than the decision boundary and one larger than the decision boundary. From the points given here, there are two points that could be support vectors:

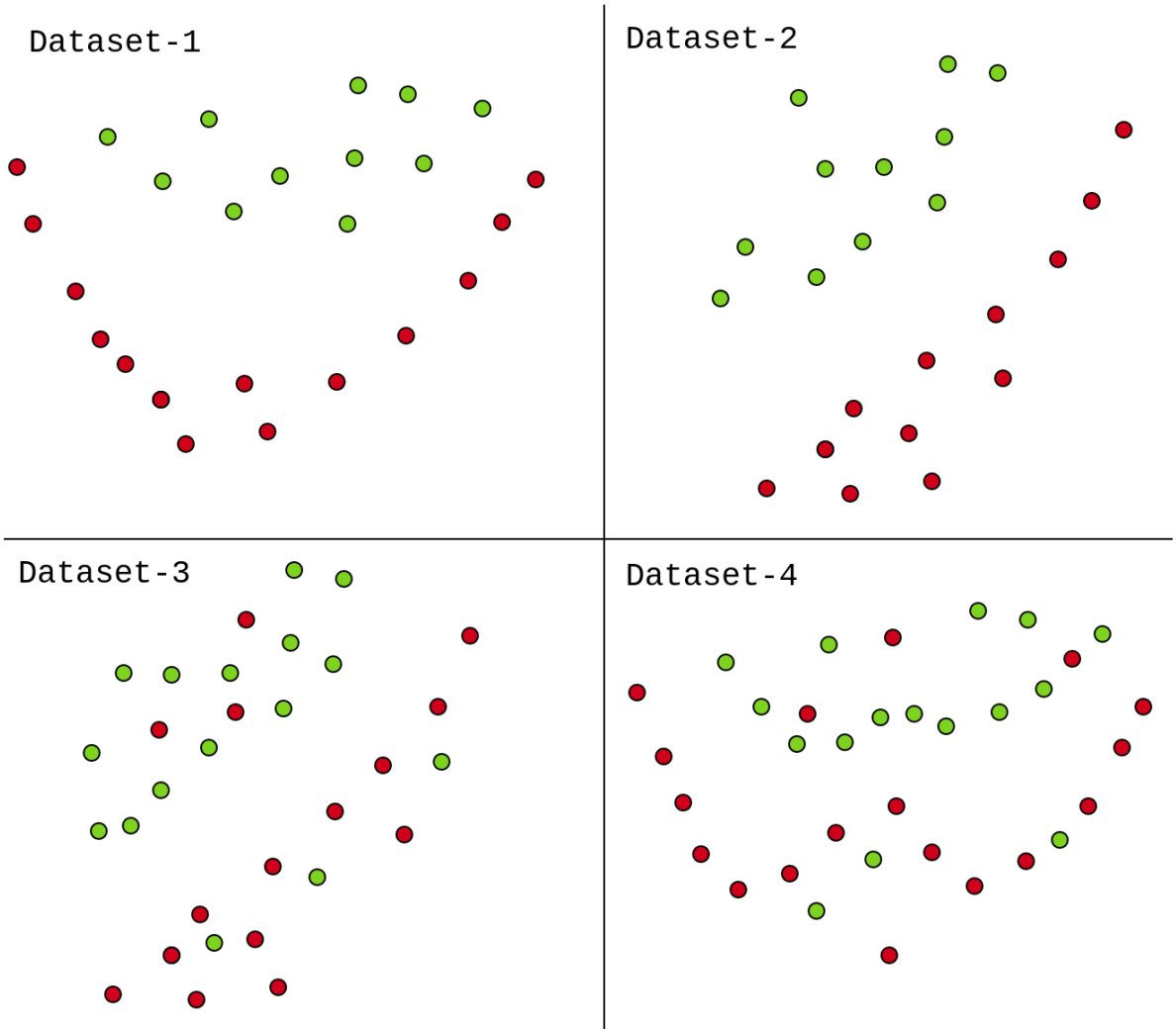
- $[-1 \quad 5]^T$: this point lies on $x_1^2 + x_2^2 = 26$
- $[2\sqrt{5} \quad -2]^T$: this point lies on $x_1^2 + x_2^2 = 24$

Note that these two are "potential" support vectors. Every support vector lies on the supporting hyperplanes. But every point on the supporting hyperplanes need not be a support vector.

Question-8 [1 point]

Statement

Match the following classification datasets with the most appropriate choice of adjectives:



Options

(a)

Dataset-1: kernel, hard-margin

Dataset-2: hard-margin

Dataset-3: soft-margin

Dataset-4: hard-margin

(b)

Dataset-1: kernel, hard-margin

Dataset-2: hard-margin

Dataset-3: soft-margin

Dataset-4: kernel, soft-margin

(c)

Dataset-1: soft-margin

Dataset-2: hard-margin

Dataset-3: kernel

Dataset-4: soft-margin

Answer

(b)

Solution

- Dataset-1: the decision boundary is non-linear. The structure is non-linear and the problem is linearly separable in some high dimensional space.
- Dataset-2: this is a clear case of hard-margin SVM
- Dataset-3: the boundary is linear. The presence of outliers suggests that this should be solved using a soft-margin SVM
- Dataset-4: The boundary is non-linear. In addition, the dataset has outliers. So, this should involve both a kernel and a soft-margin formulation.

Week 11

Question 1

Statement

In each round of AdaBoost, the weight for a particular training observation is increased going from round t to round $t+1$ if the observation was...

Options

(a)

classified incorrectly by the weak learner trained in round t

(b)

classified correctly by the weak learner trained in round t

(c)

classified incorrectly by a majority of the weak learners trained up to round t

(d)

classified correctly by a majority of the weak learners trained up to round t

Answer

(a)

solution

Since in AdaBoost, we increase the weight of the incorrectly classified points for the next bag, the option (a) is correct.

Question 2

Statement

Which of the following statements are true about bagging?

Options

(a)

In general, the final model has a higher bias than the individual learners.

(b)

In general, the final model has less bias than the individual learners.

(c)

In general, the final model has a higher variance than the individual learners.

(d)

In general, the final model has less variance than the individual learners.

Answer

(d)

Solution

Bagging on high variance models will reduce the variance without increasing the bias.
There is always a tradeoff between bias and variance. And reducing variance may cost increment in the bias. But bagging on high variance and low bias models reduces the variance without making the predictions biased.

Question 3

Statement

Is the following statement true or false?

If a point lies between the supporting hyperplanes in the soft-margin SVM problem, it always pays a positive bribe and plays a role in defining w^*.

Options

(a)

True

(b)

False

Answer

(a)

Solution

If a point lies between the supporting hyperplanes, it satisfies the following:

$$w^{*T} x_i y_i < 1 \tag{1}$$

Using the 1^{st} constraint,

$$
\begin{aligned}
1 - {w^*}^T x_i y_i - \xi_i^* &\leq 0 \\
\Rightarrow \xi_i^* &\geq 1 - {w^*}^T x_i y_i \\
\text{from } 1\text{, we can conclude that} & \\
\xi_i^* &> 0
\end{aligned}
$$

It implies that if a point lies between the supporting hyperplanes in the soft-margin SVM problem, it always pays a positive bribe

Using the CS 2,

$$
\begin{aligned}
\beta_i^* \xi_i^* &= 0 \\
\Rightarrow \beta_i^* &= 0 \quad (\text{as } \xi_i^* > 0)
\end{aligned}
$$

It implies that $\alpha_i^* = C$ and therefore If a point lies between the supporting hyperplanes in the soft-margin SVM problem, it plays a role in defining w^*.

Question 4

Statement

Is the following statement true or false?

If i^{th} point in soft-margin SVM pays a non-zero bribe $(\xi_i > 0)$, then the value of α_i is C.

Options

(a)

True

(b)

False

Answer

(a)

Solution

Using the CS 2,

$$
\begin{aligned}
\beta_i^* \xi_i^* &= 0 \\
\Rightarrow \beta_i^* &= 0 \quad (\text{as } \xi_i^* > 0)
\end{aligned}
$$

It implies that $\alpha_i^* = C$

Common data for question 5 and 6

Statement

Consider that an AdaBoost model is trained on the following binary classification dataset.

x_1	x_2	**Label** (y)
3.7	2	0
2.0	2	0
5	4	1
2.9	5	0
4.1	6	1

The first decision stump was created using the question $x_2 < 4$ or not. The error of a decision stump is defined as the proportion of misclassified points.

Question 5

Statement

Find the value of α_0. Notation is defined as per lecture.

Options

(a)

$\ln 2$

(b)

$\ln 4$

(c)

$\ln(4/5)$

(c)

$\ln(\sqrt{3/2})$

Answer

(a)

Solution

If we split the root node as per question $x_2 < 4$ or not, the left node will contain the points $(3.7, 2), (2.0, 0)$ and the labels of these points are $0, 0$ respectively. Therefore, the prediction in left node will be 0 (the majority class).

Similarly, in the right nodes, labels will be $1, 0$, and 1 and the prediction will be 1 (the majority class).

Only one point $(2.9, 5)$ is misclassified.

Therefore, error is $e = \frac{1}{5}$

$$\begin{aligned}\alpha_0 &= \ln \sqrt{\frac{1-e}{e}} \\ &= \ln \sqrt{\frac{1-1/5}{1/5}} \\ &= \ln 2\end{aligned}$$

Question 6

Statement

How will the weight corresponding to the last example change for creating the next stump?

Options

(a)

It will increase

(b)

It will decrease

Answer

(b)

Since the last example is correctly classified, its weight will decrease.

Question 7

Statement

A strong learner L is formed as per the AdaBoost algorithm by three weak learners L_1, L_2, and L_3. Their performance/weights (α) are $1, 0.4$, and 1.6, respectively. For a particular point, L_1 and L_2 predict that its label is positive, and L_3 predicts that it's negative. What is the final prediction the learner L makes on this point? Enter 1 or -1.

Answer

-1 No range is required

Solution

For the final prediction, we have

$$\alpha_1 h_1(x) + \alpha_2 h_2(x) + \alpha_3 h_3(x) = 1 + 0.4 + 1.6(-1) = -0.2 < 0$$

Therefore, prediction will be -1

Question 8

Statement

Which of the following options is correct? Select all that apply.

Options

(a)

In bagging, typically around $\frac{1}{3}rd$ data points remain unselected in bags if the number of data points is large.

(b)

Each weak leaner has equal importance in making the final prediction in Bagging.

(c)

Each weak leaner has equal importance in making the final prediction in AdaBoost.

(d)

Generally, weak learners in the random forest tend to overfit.

Answer

(a), (b), (d)

Solution

The probability that a point will not be selected in any one pick will be $(1 - 1/n)$.

The probability that a point will not be selected in n picks will be $(1 - 1/n)^n$.

as $n \to \infty, \ \ (1 - 1/n)^n \to 0.33$

that is In bagging, typically around $\frac{1}{3}rd$ data points remain unselected in bags if the number of data points is large.

In bagging, each learner has equal importance in making the final prediction as the majority of all the predictions are taken into account, and therefore each prediction counts.

But in AdaBoost, the weighted average is taken into account, and the estimator which has a higher value of α will have a higher importance in making the final prediction.

In the random forest, overfit models are preferred as they have high variance and low bias.

Common data for questions 9, 10, and 11

Statement

We have trained four models in the same dataset with different hyperparameters. In the following table, we have recorded the training and testing errors for each of the models.

Model	Training error	Test error
1	0.2	1.8

Model	Training error	Test error
2	1.0	1.1
3	0.5	0.7
4	1.9	2.3

Question 9

Statement

Which model tends to underfit?

Options

(a)

Model 1

(b)

Model 2

(c)

Model 3

(d)

Model 4

Answer

(d)

Solution

Model 4 has high training error as well as high test error. It means that model 4 as high variance and high bias and tends to underfit.

Question 10

Statement

Which model tends to overfit?

Options

(a)

Model 1

(b)

Model 2

(c)

Model 3

(d)

Model 4

Answer

(a)

Model 1 has less training error as well as high test error. It means that model 4 as high variance and low bias and tends to overfit.

Question 11

Statement

Which model would you choose?

Options

(a)

Model 1

(b)

Model 2

(c)

Model 3

(d)

Model 4

Answer

(c)

Model 3 has less training and test error and therefore, it is most preferred.

Graded

Question-1

Statement

Consider the following data and the hypothesis function $h(x) = sign(g(x))$:

g(x)	y
+30	+1
-20	-1
-1	-1
+1	+1

Which of the following will be true?

Options

(a)

The values of 0-1 loss and squared loss will be same, which will be equal to zero.

(b)

The values of 0-1 loss and squared loss will be same, which will be equal to some large positive quantity.

(c)

The value of 0-1 loss will be zero, while the value of squared loss will be some large positive quantity.

(d)

The value of squared loss will be zero, while the value of 0-1 loss will be some large positive quantity.

Answer

(c)

Solution

There is an ambiguity in this question. The given $h(x) = sign(g(x))$ is supposed to be explain the 0-1 loss and not the squared loss. In this case,

g(x)	y	sign(g(x))	0-1 loss	Squared loss ((g(x)-y)^2
+30	+1	+1	0	(29)^2
-20	-1	-1	0	(-19)^2
-1	-1	-1	0	0

g(x)	y	sign(g(x))	0-1 loss	Squared loss ((g(x)-y)^2
+1	+1	-1	0	0

The value of 0-1 loss is zero, while the value of squared loss is a large positive quantity. Hence, option (c) should be correct.

However, since h(x) is stated to be sign(g(x)) and if squared loss is computed on the column 2 and 3 of the above table, squared loss will also come out to be 0, resulting in option (a) to be correct.

Hence, both options (a) and (c) will fetch marks.

Common instructions for questions 2-5

Consider a neural network with the number of neurons in different layers as mentioned in the below list for a regression task:

[5, 5, 4, 3, 1]

Question-2

Statement

How many hidden layers are there in the network?

Answer

3 (No range required)

Solution

The first layer is input layer and the last one is hidden layer. The intermediate 3 layers are hidden layers.

Question-3

Statement

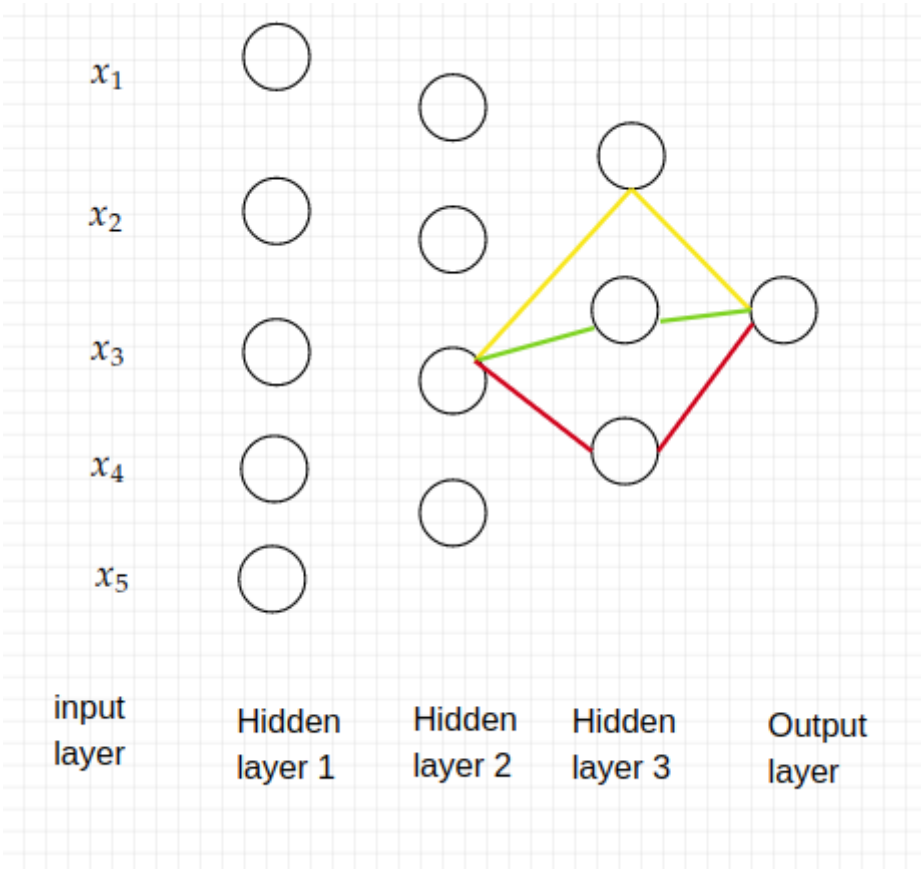
Through how many paths can the 3rd neuron in the 2nd hidden layer affect the final output?

Answer

3 (No range required)

Solution

The three paths are shown below in different colors.



Question-4

Statement

Assuming that there is a bias associated with each neuron, how many total parameters need to be computed?

Answer

73 (No range required)

Solution

#weights from input to hidden layer 1: 5*5 = 25

#weights from hidden layer 1 to hidden layer 2: 5*4 = 20

#weights from hidden layer 2 to hidden layer 3: 4*3 = 12

#weights from hidden layer 3 to output layer: 3*1 = 3

There will be a bias associated withe each neuron except the input layer neurons. (Input layer is used to simply pass on the inputs to the subsequent layers)

#biases = 5+4+3+1 = 13

Total number of parameters = 25+20+12+3+13 = 73

Question-5

Statement

What will be an appropriate activation function for the output layer?

Options

(a)

Sigmoid

(b)

Linear

(c)

ReLU

Answer

(b)

Solution

Sigmoid is used for binary classification at the output layer.

ReLU is mostly used at the hidden layers. (It does not make sense to be used at the output layer.)

Linear is an appropriate activation function for output layer in a regression problem.

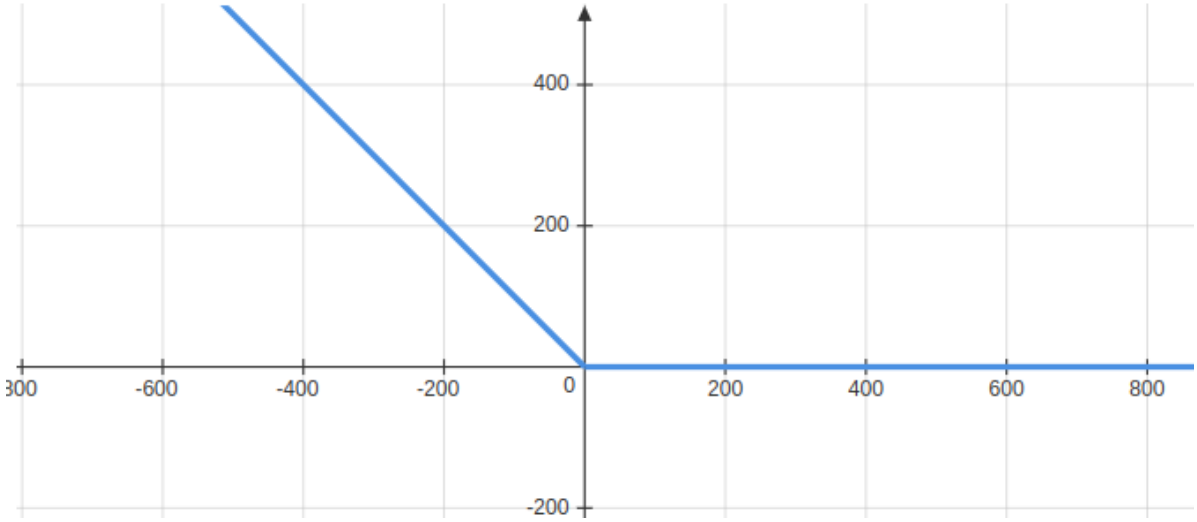
Question-6

Statement

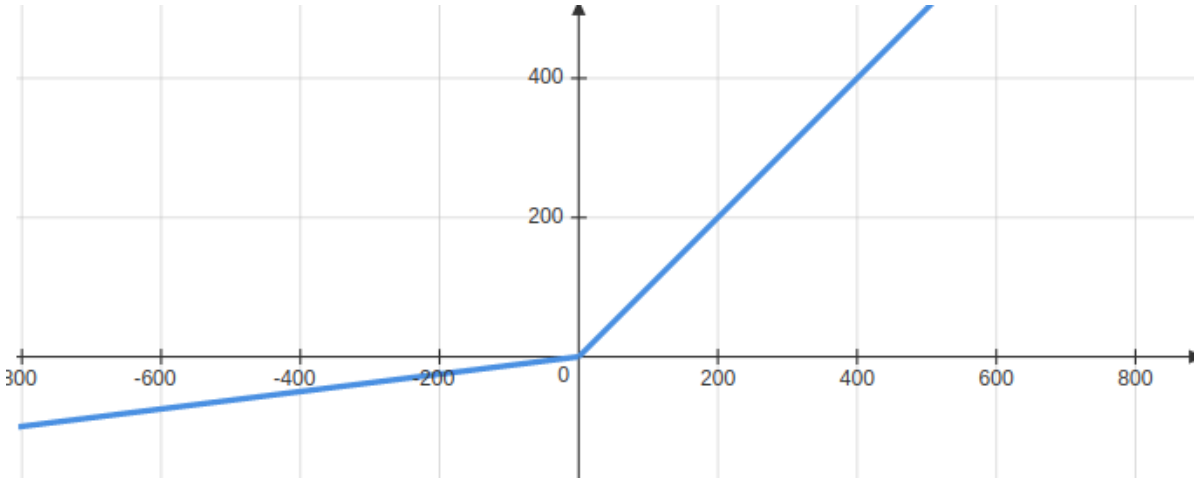
Which of the following represents ReLU activation function?

Options

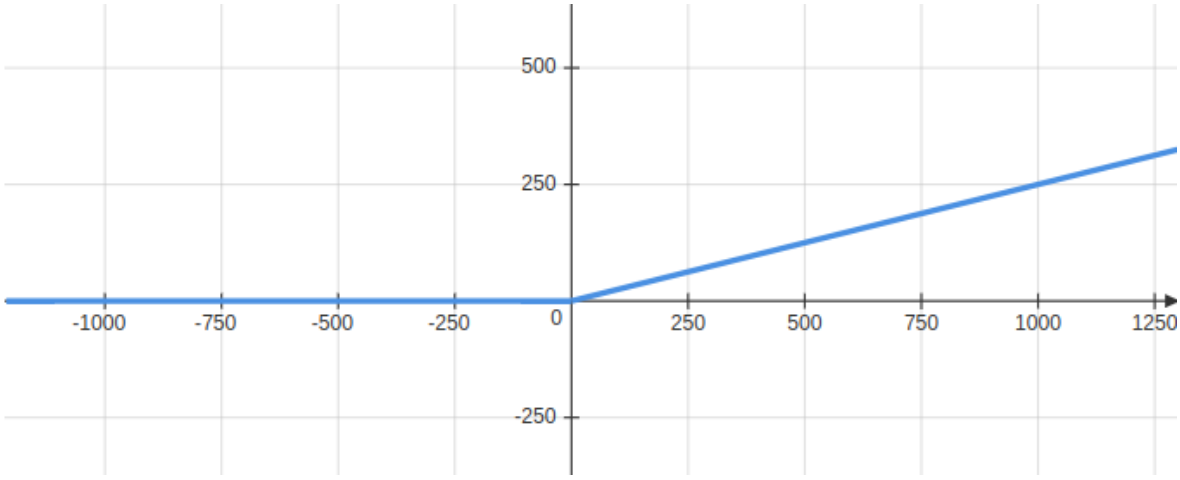
(a)



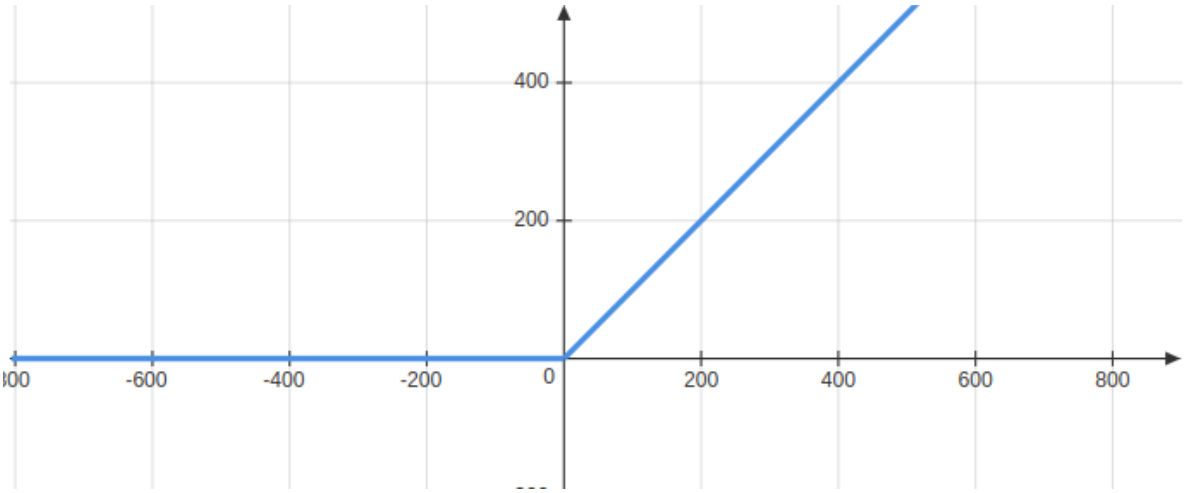
(b)



(c)



(d)



Answer

(d)

Solution

ReLU converts the negative inputs to zero, and keeps the positive inputs same.

Question-7

Statement

Suppose we build a neural network for a 5-class classification task. Suppose for a single training example, the true label is `[1 0 0 0 0]` while the predictions by the neural network are `[0.1 0.5 0.1 0.1 0.2]`. What would be the value of cross entropy loss for this example?

Answer

3. 322 (Range: 3.2 to 3.4)

Solution

Cross entropy = $-\sum y_i \; log_2 \; \hat{y}_i$

= $-1 * log_2(0.1) - 0 * log_2(0.5) - 0 * log_2(0.1) - 0 * log_2(0.1) - 0 * log_2(0.2)$

= $-log_2(0.1) = 3.322$

Question-8

Statement

State True or False:

If $CE(y_1, y_2)$ represents the value of cross entropy loss, then $CE(y_1, y_2) = CE(y_2, y_1)$ always.

Options

(a)

True

(b)

False

Answer

(b)

Solution

Cross entropy ($CE(y_1, y_2)$) = $-\sum y_1 log_2 y_2$ which is not commutative.