Endterm Practice-1

This document has 20 questions.

Question-1

Statement

Consider the following modification to the prediction of the label for a data-point $\mathbf{x}$ in a logistic regression model.

$$\hat{y} = \begin{cases} 1, & P(y = 1 \mid \mathbf{x}) \geq T \\ 0, & \text{otherwise} \end{cases}$$

T is called the threshold and is some real number in the interval $(0, 1)$. $\hat{y}$ stands for the predicted label. Given this setup, the equation of the decision boundary is given below:

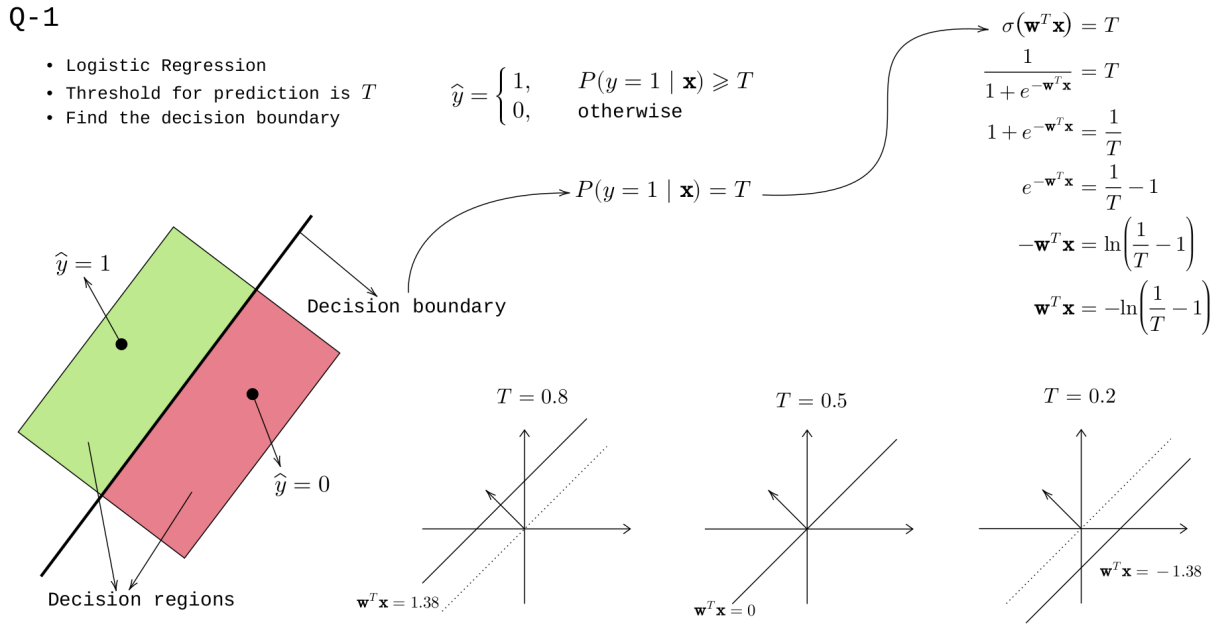
$$\mathbf{w}^T\mathbf{x} - u = 0$$

If $T = \dfrac{e}{1+e}$, what is the value of the unknown quantity u? Enter the closest integer as your answer.

Answer

1

Solution



Question-2

Statement

Consider a modified loss function for linear regression that is of the following form for a training dataset that has n points:

$$L(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n} r_i(\mathbf{w}^T\mathbf{x} - y_i)^2$$

Here, r_i is some constant in $[0, 1]$ associated with each data-point in the training dataset. What is the expression of the gradient of $L(\mathbf{w})$ with respect to $\mathbf{w}$?

Options

(a)

$$\sum_{i=1}^{n} r_i(\mathbf{w}^T\mathbf{x}_i - y_i)\mathbf{x_i}$$

(b)

$$\sum_{i=1}^{n} r_i(\mathbf{w}^T\mathbf{x}_i - y_i)$$

(c)

$$\sum_{i=1}^{n} (\mathbf{w}^T\mathbf{x}_i - y_i)\mathbf{x_i}$$

(d)

$$\sum_{i=1}^{n} r_i(\mathbf{w}^T\mathbf{x}_i - y_i)^2\mathbf{x_i}$$

Answer

(a)

Solution

The only new term in the gradient is r_i for each data-point:

$$\nabla L(\mathbf{w}) = \sum_{i=1}^{n} r_i(\mathbf{w}^T\mathbf{x}_i - y_i)\mathbf{x_i}$$

If we have to solve for $\mathbf{w}$, here is how we would go about it:

Q-8

- Weighted linear regression
- Weight for each data-point in the loss function
- $\mathbf{R} = \texttt{diag}(r_1, \cdots . r_n)$

$$L(\mathbf{w}) = \frac{1}{2} \cdot \sum_{i=1}^{n} r_i \left(\mathbf{w}^T \mathbf{x}_i - y_i \right)^2$$

$$\mathbf{X} = \begin{bmatrix} | & & | \\ \mathbf{x}_1 & \cdots & \mathbf{x}_n \\ | & & | \end{bmatrix}, \ \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

$$\begin{aligned} \nabla L(\mathbf{w}) &= \sum_{i=1}^{n} r_i \left(\mathbf{w}^T \mathbf{x}_i - y_i \right) \mathbf{x}_i \\ &= \sum_{i=1}^{n} (r_i \mathbf{x}_i) \mathbf{x}_i^T \mathbf{w} - \sum_{i=1}^{n} r_i y_i \mathbf{x}_i \end{aligned}$$

$$\begin{aligned} \mathbf{XR} &= \begin{bmatrix} | & & | \\ \mathbf{x}_1 & \cdots & \mathbf{x}_n \\ | & & | \end{bmatrix} \begin{bmatrix} r_1 & & \\ & \ddots & \\ & & r_n \end{bmatrix} \\ &= \begin{bmatrix} | & & | \\ r_1 \mathbf{x}_1 & \cdots & r_n \mathbf{x}_n \\ | & & | \end{bmatrix} \end{aligned}$$

$$\mathbf{XRX}^T \mathbf{w} = \mathbf{XRy}$$

$$\mathbf{w} = \left(\mathbf{XRX}^T \right)^{-1} \mathbf{XRy}$$

Question-3

Statement

A hard-margin, linear-SVM is trained for a 2D problem. The optimal weight vector is $\mathbf{w} = [2 \quad -1]^T$. Consider a unit square whose corners are at:

$$(0,0), (1,0), (0,1), (1,1)$$

A point is picked at random from the square. What is the probability that this point is predicted as belonging to class $+1$ by the model?

Answer

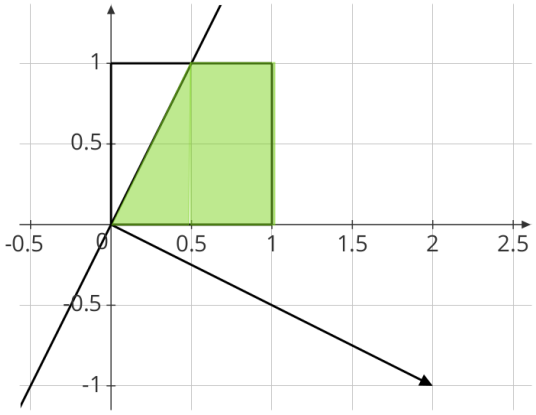
0.75

Range: $(0.74, 0.76)$

Solution

Q-9

- Hard-margin, linear-SVM
- $\mathbf{w}^* = [2 \quad -1]^T$
- Unit square (0,0), (1, 0), (0, 1), (1, 1)
- Probability that a point picked at random from the unit square is is predicted as 1



$$P(y = 1 \mid \mathbf{x}) = 1 - \frac{1}{2} \cdot 1 \cdot \frac{1}{2}$$

$$= 1 - 0.25$$

$$= 0.75$$

Question-4

Statement

Consider the MNIST digit classification problem. It has has 10 classes. The training dataset has n data-points, with an equal number of points from each of the 10 classes. Consider a dummy classifier that does prediction as follows: for each input data-point, it picks one of the 10 classes at random and outputs that as its prediction.

Accuracy of a model on a dataset is defined as the proportion of points that it classifies correctly. What is the expected accuracy of this model? Your answer should be between 0 and 1.

Answer

0.1

Range: $(0.09, 0.11)$

Solution

Let there be n points in the dataset. For each point, let C_i be a random variable that denotes the outcome of the classification. The C_is are i.i.d.:

$$C_i = \begin{cases} 1, & \text{correctly classified} \\ 0, & \text{incorrectly classified} \end{cases}$$

For a classifier that picks points uniformly at random, we see that:

$$P(C_i = 1) = \frac{1}{k}$$

Let us define the average number of correct classifications by the random variable A:

$$A = \frac{1}{n} \cdot \sum_{i=1}^{n} C_i$$

If we run multiple rounds of prediction with this random classifier on a fixed dataset of size n, each time we will get a different value for A (the realization). Here, the randomness comes from the nature of the classifier. To get an estimate of the accuracy of the model, it makes sense to talk about the expected value of A. By the linearity of expectation:

$$E[A] = \frac{1}{n} \cdot \sum_{i=1}^{n} E[C_i] = \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{1}{k} \cdot 1 = \frac{1}{k}$$

Here, we have used the fact that $C_i \sim Br(1/k)$. On an average, how does the random classifier do on this dataset? It classifies about 10% of the points correctly for $k = 10$. In the special case of $k = 2$ (binary classification), the accuracy for such a random classifier is 50%. This squares with our intuition of tossing a coin and determining the label of a data-point.

Question-5

Statement

Consider a neural network for an image classification problem. When the network is trained on the images as they are, it does a good job on the test data. Call the dataset (train + test) for this setup D_1 and the network N_1. Assume that we now turn all images upside down, in both the training and test dataset. Now, the network with the same architecture is trained from scratch on this modified dataset. Call this dataset (train + test) for this setup D_2 and network N_2. Select the most appropriate option?

Options

(a)

The network N_2 will not be able to learn anything from D_2. Its test-accuracy on D_2 will be very low.

(b)

The network N_2 will be able to learn useful patterns from D_2. In fact, the performance of network N_2 on D_2 will be similar to N_1 on D_1.

(c)

The network N_2 will be able to learn somewhat useful patterns from D_2. But the performance of N_1 on D_1 will be much better than N_2 on D_2.

Answer

(b)

Solution

Assume that each neuron in the input layer corresponds to a pixel in the image. There is a connection from each neuron in the input layer to every neuron in the first hidden layer. Such a layer is called a fully-connected layer. We only deal with fully connected layers in our course. For this setup, the way in which we order the input neurons is immaterial. Even if we shuffle the input neurons, the network will still be able to learn a classifier. Irrespective of the ordering, there is always going to be a connection between any node in the input layer to any node in the first hidden layer.

Question-6

Statement

Find the hinge loss for a soft-margin, linear-SVM on the dataset given below. The weight vector is $\begin{bmatrix}0 & 1\end{bmatrix}^T$.

x_1	x_2	y
2	1	1
-2	1	1
-1	2	1
0	2	-1
1	-1	-1
2	-2	-1
-2	0	-1

Answer

4

Solution

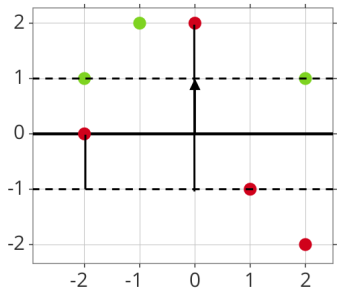
- Soft-margin, Linear-SVM
- $\mathbf{w} = \begin{bmatrix}0 & 1\end{bmatrix}^T$
- Hinge loss

$$L(\mathbf{X}, \mathbf{y}, \mathbf{w}) = \sum_{i=1}^{n} \max\left[1 - \left(\mathbf{w}^T \mathbf{x}_i\right) y_i, 0\right]$$

$\mathbf{X}, \mathbf{y} =$

x_1	x_2	y
2	1	1
-2	1	1
-1	2	1
0	2	-1
1	-1	-1
2	-2	-1
-2	0	-1

x_1	x_2	y	$1 - \left(\mathbf{w}^T \mathbf{x}\right) y$	L
2	1	1	0	0
-2	1	1	0	0
-1	2	1	-1	0
0	2	-1	3	3
1	-1	-1	0	0
2	-2	-1	-1	0
-2	0	-1	1	1



Question-7

Statement

Consider a logistic regression model for a binary classification problem with two features x_1 and x_2 and labels 1 and 0. Let x_1 be the horizontal axis and x_2 to be the vertical axis. You are given two feature vectors:

$$\mathbf{x_1} = \begin{bmatrix} 1 \\ \sqrt{3} \end{bmatrix}, \mathbf{x_2} = \begin{bmatrix} -1 \\ \sqrt{3} \end{bmatrix}$$

The weight vector makes an angle of θ with the positive x_1 axis (horizontal). Each θ corresponds to a different classifier. For what range of values of θ are both $\mathbf{x_1}$ and $\mathbf{x_2}$ predicted to belong to class-1?

Options

(a)

$$30^\circ < \theta < 150^\circ$$

(b)

$$60^\circ < \theta < 120^\circ$$

(c)

$$0^\circ < \theta < 180^\circ$$

(d)

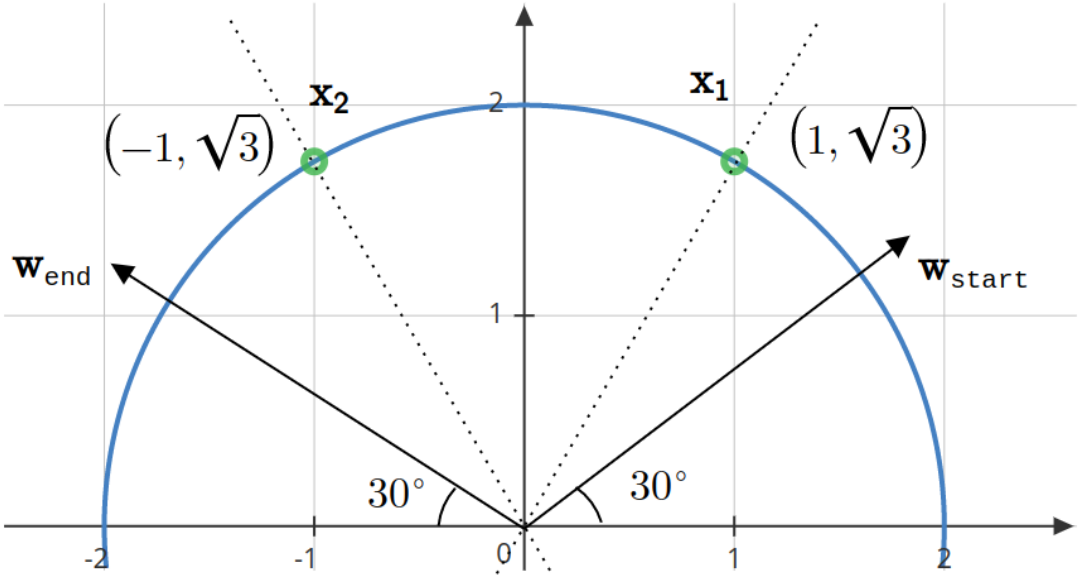
$$0 < \theta < 360^\circ$$

Answer

(a)

Solution

Dotted lines are decision boundaries



Question-8

Statement

Consider the following data-points for a regression problem with a single feature. The points are of the form (x_i, y_i), where x_i is the feature and y_i is the label.

$$(-3, 3), \quad (0, 4), \quad (1, 12), \quad (3, 15), \quad (4, 16)$$

If a model $y = c$ is fit for this problem with mean squared error as the loss function what is the best estimate for c? Note that we are fitting a constant here.

Answer

10

Solution

The estimate $\hat{c}$ is:

$$\hat{c} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i = 50/5 = 10$$

Question-9

Statement

In the context of a binary classification problem, are the following set of data-points in $\mathbb{R}^2$ linearly separable? Each row of X is a data-point. The labels of these points are given in the label vector y.

$$X = \begin{bmatrix} 1 & 2 \\ 3 & -4 \\ 5 & 0 \\ -1 & -2 \\ -3 & 4 \\ -2 & -6 \end{bmatrix}, \quad y = \begin{bmatrix} 1 \\ 1 \\ 1 \\ -1 \\ -1 \\ -1 \end{bmatrix}$$

Options

(a)

Yes

(b)

No

Answer

(a)

Question-10

Statement

Which of the following corresponds to the "naive assumption" in Naive Bayes classification? $\mathbf{x} = [x_1, \cdots, x_d]^T$ is a feature vector and y is its label.

Options

(a)

$$P(x_1, \cdots, x_d \mid y) = P(x_1, \cdots, x_d) \cdot P(y)$$

(b)

$$P(x_1, \cdots, x_d \mid y) = \prod_{i=1}^{d} P(x_i \mid y)$$

(c)

$$P(y \mid x_1, \cdots, x_d) = P(y)$$

(d)

$$P(x_1, \cdots, x_d, y) = P(x_1, \cdots, x_d) \cdot P(y)$$

Answer

(b)

Question-11

Statement

Consider the following statements in the context of a hard-margin linear-SVM:

(1) Every support vector lies on one of the two supporting hyperplanes.

(2) Every point on one of the two supporting hyperplanes is a support vector.

Options

(a)

Only (1) is true

(b)

Only (2) is true

(c)

Both (1) and (2) are true

(d)

Both (1) and (2) are false

Answer

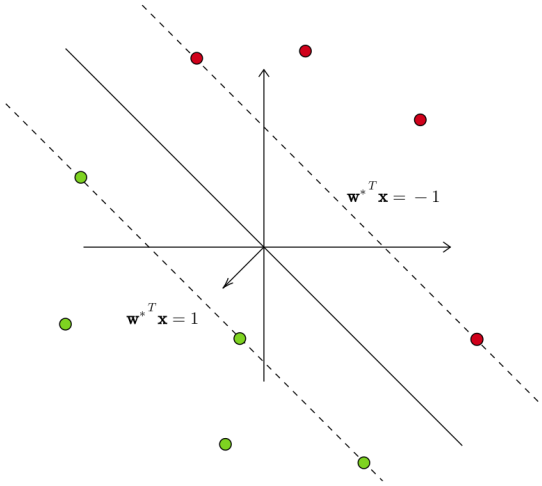
(a)

Solution

Q-10

Select all true statements regarding a hard-margin SVM:

(1) Every support vector lies on one of the two supporting hyperplanes.
(2) Every point on one of the two supporting hyperplanes is a support vector.



$$\left(\mathbf{w}^{*T}\mathbf{x}_i\right)y_i \geqslant 1 \implies 1 - \left(\mathbf{w}^{*T}\mathbf{x}_i\right)y_i \leqslant 0$$

$$\alpha_i^* \cdot \left[1 - \left(\mathbf{w}^{*T}\mathbf{x}_i\right)y_i\right] = 0$$

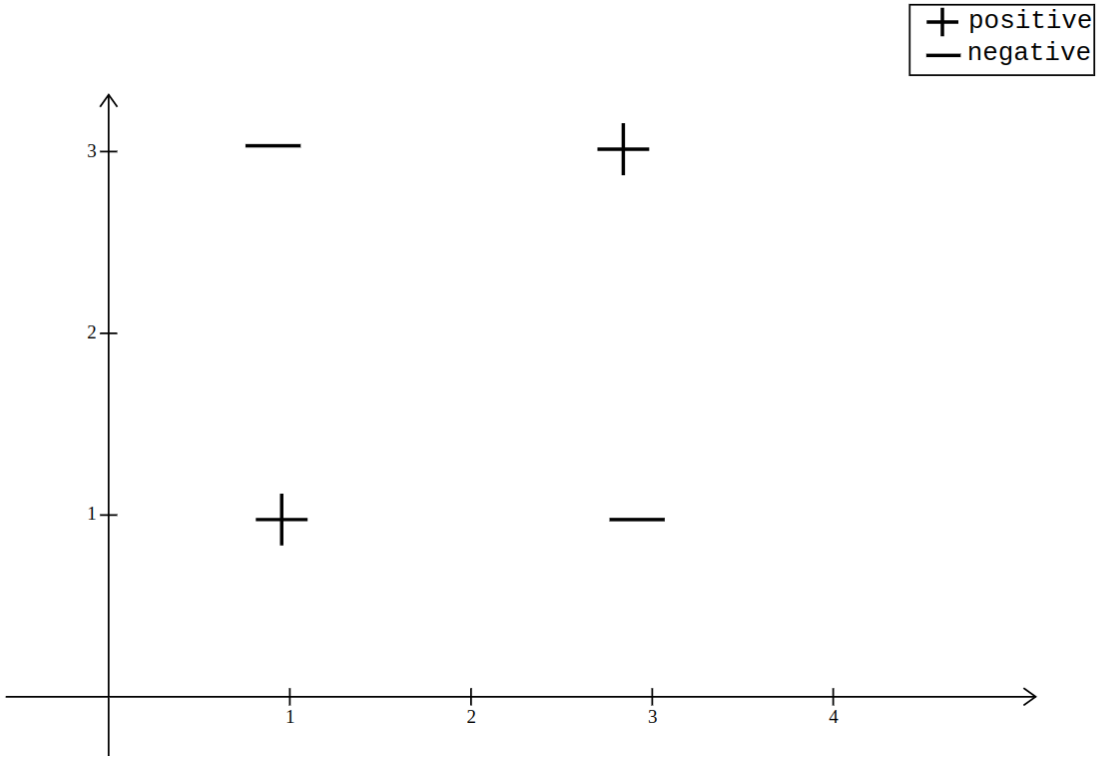
$$\text{Support vector} \implies \alpha_i^* > 0 \implies \left(\mathbf{w}^{*T}\mathbf{x}_i\right)y_i = 1$$

$$\left(\mathbf{w}^{*T}\mathbf{x}_i\right)y_i = 1 \quad \not\Longrightarrow \quad \alpha_i^* > 0$$

Question-12

Statement

Consider a binary classification task that has 2 features and an arbitrary linear classifier. This classifier is now tested on a dataset of four points given below:



What are the possible values of the accuracy (proportion of points correctly classified) of the classifier? All options are independent of each other. Assume that the decision boundary of the classifier does not pass through any one of the four points. Multiple options could be correct.

Options

(a)

0

(b)

0.25

(c)

0.5

(d)

0.75

(e)

1

Answer

(b), (c), (d)

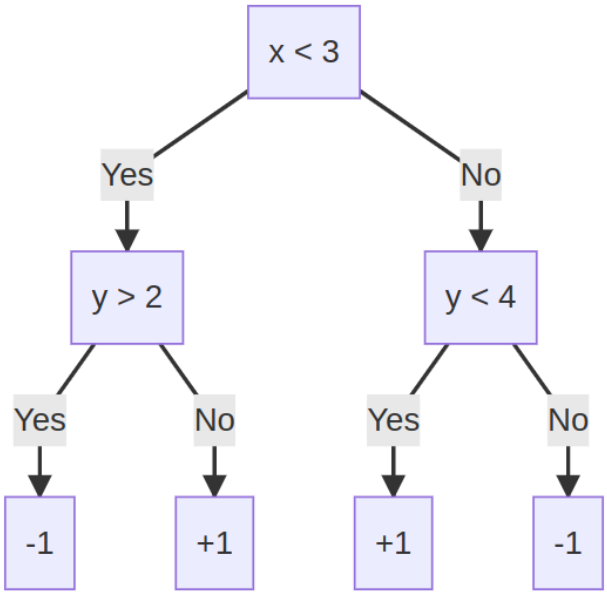
Solution

Try out different lines that separate these four points.

Question-13

Statement

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



This decision tree partitions the feature space into four regions, one corresponding to each of the four leaves. Call the leaves L_1, L_2, L_3, L_4 from left to right. Assume that $x, y \geq 0$ for all points. Now, consider the set of all points S in $\mathbb{R}^2$ that go into leaf L_2 when passed through the decision tree. That is:

$$S = \{(x, y) \mid x \geq 0, y \geq 0, (x, y) \text{ goes into } L_2, (x, y) \in \mathbb{R}^2\}$$

What is the area of the region corresponding to S?

Answer

6

Solution

$$S = \{(x, y) \quad | \quad 0 \leq x < 3 \text{ and } 0 \leq y \leq 2\}$$

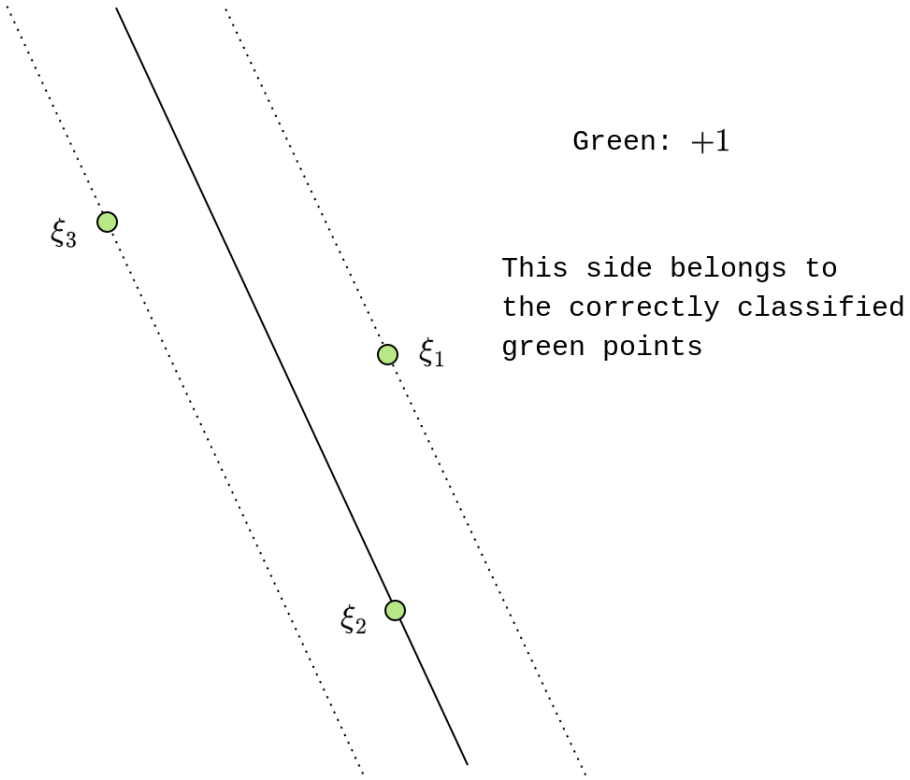
The area of the rectangle is 6.

Comprehension type (14 - 16)

Statement

Common Data for questions (14) to (16)

Consider a soft-margin, linear-SVM that has been trained on a dataset. A subset of three data-points from the positive class (green) from this training dataset is shown below. The decision boundary (solid line) and the bounding planes (dotted lines) are also displayed here. The slack variables aka bribes for these three points are ξ_1, ξ_2, ξ_3 respectively.



Solution

We look at the (scaled) distance of a data-point from the correct supporting hyperplane to which it should belong.

Question-14

Statement

What is the value of ξ_1?

Answer

0

Question-15

Statement

What is the value of ξ_2?

Answer

1

Question-16

Statement

What is the value of ξ_3?

Answer

2

Question-17

Statement

In K-means clustering, if we decide to have 5 clusters given 100 data-points, what is the total number of cluster assignments possible?

Options

(a)

100^5

(b)

5^{100}

(c)

500

(d)

105

Answer

(b)

Question-18

Statement

In the context of Bayesian modeling, a MAP estimator $\hat{p}_{MAP}$, is a point estimate obtained by finding the mode of the posterior distribution. Find the MAP estimator for the dataset $\{1, 0, 1, 0, 1, 0\}$ modeled using a Bernoulli distribution with $\text{Beta}(3, 7)$ as the prior. Enter your answer correct to three decimal places.

Hint: If $\alpha, \beta > 1$, the mode for the $\text{Beta}(\alpha, \beta)$ can be found using some well known technique from calculus.

Answer

0.357

Range: [0.34, 0.37]

Solution

Q-19

- $D = \{1, 0, 1, 0, 1, 0\}$
- Prior is Beta$(3, 7)$
- Find the MAP estimate

$$\log(\texttt{posterior}) = 5 \log p + 9 \log(1 - p)$$

$$\hat{p}_{MAP} = \arg\max_{p} \quad \texttt{posterior}(p)$$

$$\texttt{Posterior} \propto \texttt{Prior} \times \texttt{Likelihood}$$

$$\propto \frac{p^2(1-p)^6}{B(3,7)} \times p^3(1-p)^3$$

$$\propto p^5(1-p)^9$$

$$\propto \texttt{Beta}(6, 10)$$

$$\frac{5}{p} - \frac{9}{1-p} = 0$$

$$5 - 5p - 9p = 0$$

$$\hat{p}_{MAP} = \frac{5}{14}$$

If $\alpha, \beta > 1$, then mode of Beta(α, β) is:

$$\frac{\alpha - 1}{\alpha + \beta - 2}$$

Question-19

Statement

A logistic regression model is as confident about its prediction that a point $\mathbf{x}_2$ belongs to class 1 as it is about its prediction that a point $\mathbf{x}_1$ belongs to class 0. What is the ratio of the distances (absolute values) of the points $\mathbf{x}_2$ and $\mathbf{x}_1$ from the decision boundary?

Answer

1

Solution

We are given:

$$P(y = 1 \mid \mathbf{x}_2) = P(y = 0 \mid \mathbf{x}_1)$$

From this, we get:

$$\frac{1}{1 + e^{-\mathbf{w}^T \mathbf{x}_2}} = 1 - \frac{1}{1 + e^{-\mathbf{w}^T \mathbf{x}_1}}$$

Simplifying this expression and then taking $\log$ on both sides, we get:

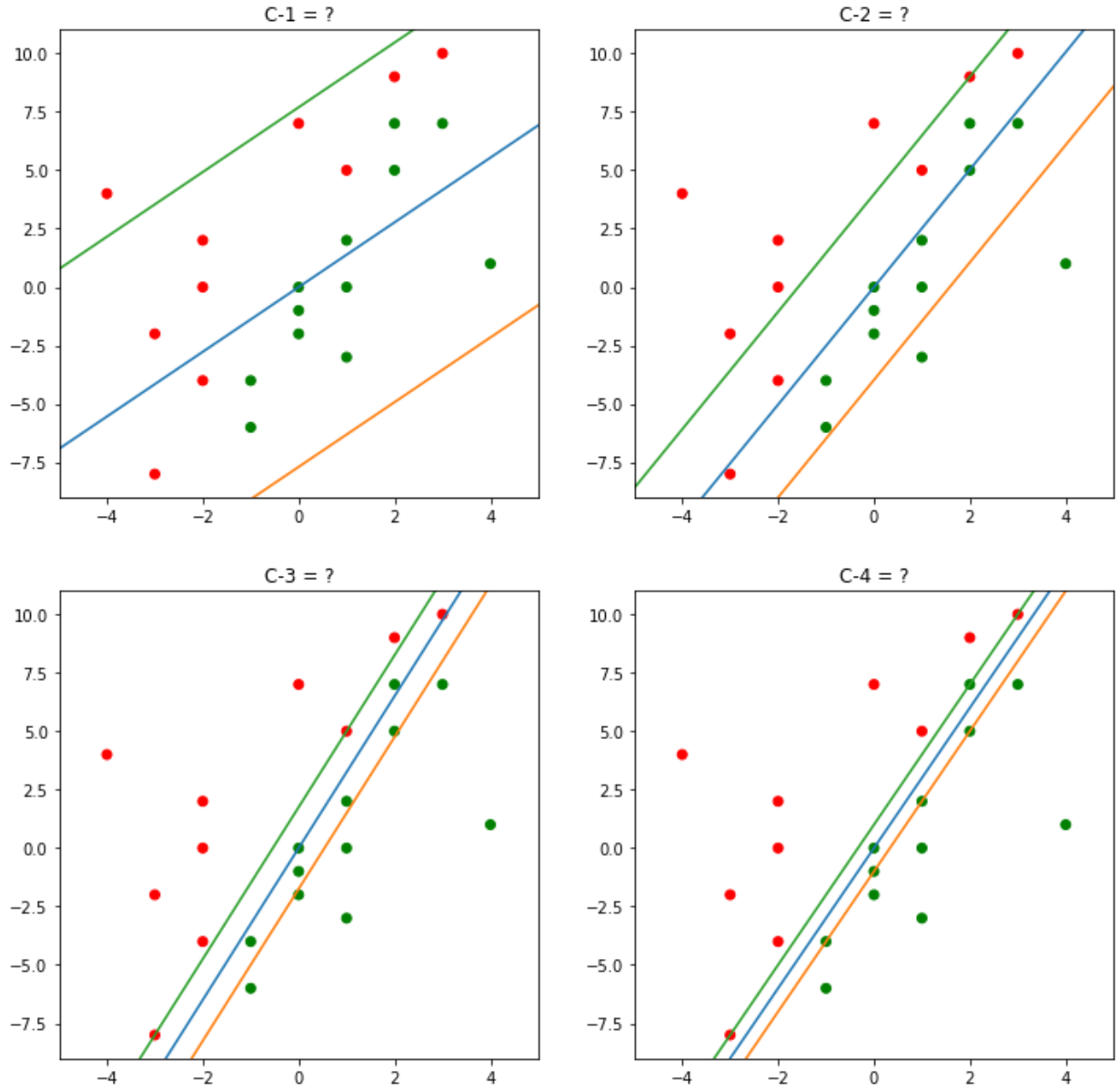
$$\mathbf{w}^T \mathbf{x}_2 = -\mathbf{w}^T \mathbf{x}_1$$

The ratio of these distances turns out to be 1. Geometrically, we see that both points are at the same distance from the decision boundary (assuming a threshold of 0.5), but on opposite sides.

Question-20

Statement

Consider four different soft-margin linear-SVM models trained on the same dataset with different values of C. The decision boundary and the supporting hyperplanes are plotted along with the dataset.



Select the most appropriate values for C_1, C_2, C_3 and C_4.

Options

(a)

$C_1 = 10, C_2 = 1, C_3 = 0.1, C_4 = 0.01$

(b)

$C_1 = 0.01, C_2 = 0.1, C_3 = 1, C_4 = 10$

(c)

$C_1 = C_2 = C_3 = C_4 = 1$

(d)

$C_1 = C_2 = C_3 = C_4 = 10$

Answer

(b)

Solution

Q-20

- Soft-Margin Linear-SVM
- Map C_i to $[0.01,\ 0.1,\ 1,\ 10]$ for $1 \leqslant i \leqslant 4$

$$\min_{\mathbf{w}} \quad \frac{||\mathbf{w}||^2}{2} + C\sum_{i=1}^{n} \max\left[0, 1 - \left(\mathbf{w}^T\mathbf{x}_i\right)y_i\right]$$

Loss = Margin + Hinge-loss

Ideal: Small $||\mathbf{w}||$ and small hinge loss

Large $||\mathbf{w}||$ $\implies$ Narrow margin $\implies$ Small hinge loss

Small $||\mathbf{w}||$ $\implies$ Wide margin $\implies$ Large hinge loss