Endterm-Practice-2

This document has 20 questions.

Question-1

Statement

Which of these statements are true *in general*?

Options

(a)

Deep trees certainly perform well on the training data.

(b)

Deep trees perform well on both training and test data.

(c)

Deep trees perform well on the training data but may not perform well on the test data.

(d)

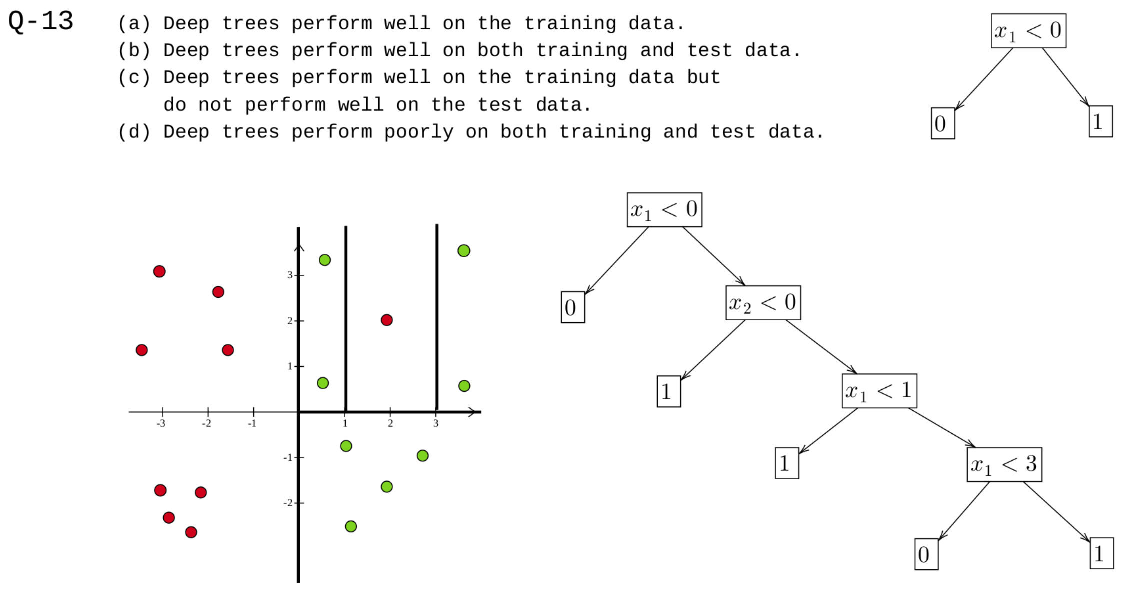
Deep trees perform poorly on both training and test data.

Answer

(a), (c)

Solution

Ignore the options given in the figure below. Only look at the two trees. One of them is a stump and the other is a deep tree. Notice how the stump is already doing a good job of classifying the training data and will generalize well on the test data. The deep tree however overfits and may not perform well on the test data. The solid lines shown are the decision boundaries corresponding to the deep tree.



```
Q-13   (a) Deep trees perform well on the training data.
       (b) Deep trees perform well on both training and test data.
       (c) Deep trees perform well on the training data but
           do not perform well on the test data.
       (d) Deep trees perform poorly on both training and test data.
```

Question-2

Statement

Which of the following are valid covariance matrices for centered datasets in $\mathbb{R}^3$?

$$\mathbf{C}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 4 & 3 & 0 \\ 1 & 9 & 2 \end{bmatrix}, \ \mathbf{C}_2 = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \end{bmatrix}, \ \mathbf{C}_3 = \begin{bmatrix} 5 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \mathbf{C}_4 = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

Options

(a)

Only $\mathbf{C}_1$

(b)

Only $\mathbf{C}_2$

(c)

$\mathbf{C}_2$ and $\mathbf{C}_3$

(d)

$\mathbf{C}_2, \mathbf{C}_3$ and $\mathbf{C}_4$

(e)

All four are valid covariance matrices

(f)

None of them

Answer

(c)

Solution

$\mathbf{C}_1$ cannot be a covariance matrix as it is not symmetric. $\mathbf{C}_2$ and $\mathbf{C}_3$ are valid covariance matrices. They are symmetric, positive semi-definite. As for $\mathbf{C}_4$, it is not p.s.d:

$$|\mathbf{C}_4 - \lambda I| = (1-\lambda)(1-\lambda)(-\lambda) - (1-\lambda)$$

$$= (1-\lambda)(\lambda^2 - \lambda - 1)$$

We have $\lambda = 1, \frac{1+\sqrt{5}}{2}, \frac{1-\sqrt{5}}{2}$. Since one of the eigenvalues is negative, $\mathbf{C}_4$ is not positive semi-definite. Hence it cannot be a valid covariance matrix. In another sense, the variance along a direction cannot be negative.

Question-3

Statement

The following is the vector output by some hidden layer in a neural network after the activation function has been applied: $[0.1, 0.8, 0.4, 0.5, 0.7, 0.9]^T$ Which of the following could be the activation function used in this layer?

Options

(a)

Only ReLU

(b)

Only Sigmoid

(c)

Either ReLU or Sigmoid

(d)

Neither ReLU nor Sigmoid

Answer

(c)

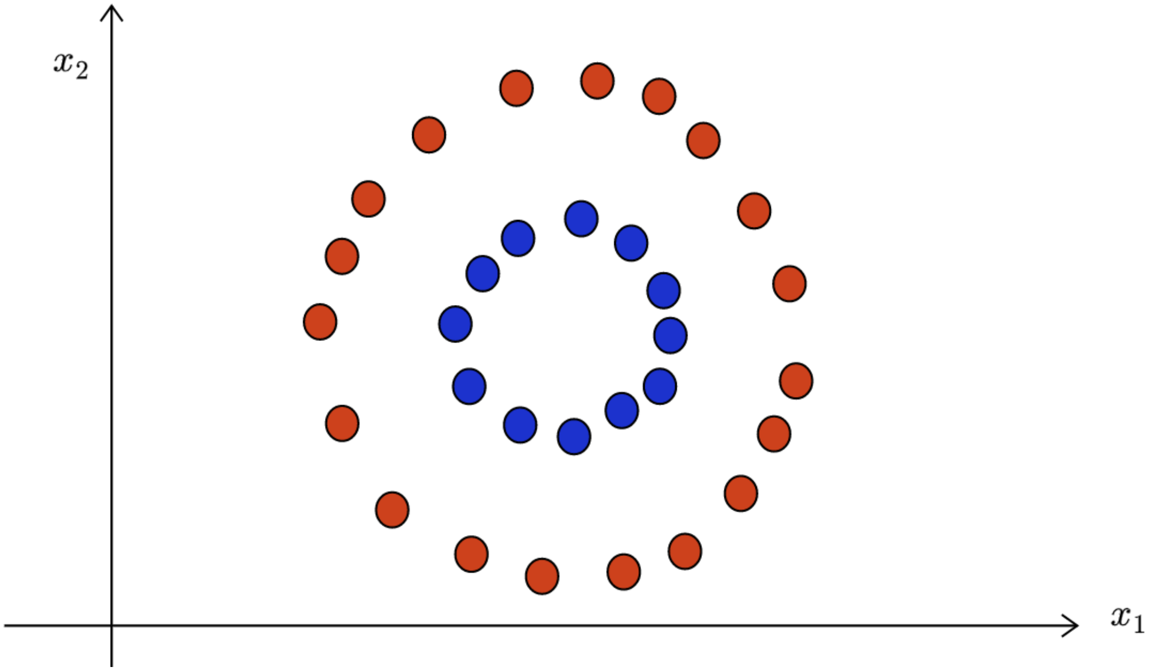
Solution

Since all activation values are in the range $(0, 1)$, the activation function could be either sigmoid or ReLU.

Question-4

Statement

Consider the following dataset in $\mathbb{R}^2$ for a binary classification problem. The red and blue points belong to two different classes:



Each data-point is of the form $\left([x_1 \quad x_2]^T, y_i\right)$. The data-points are explicitly transformed that result in a new set of features. A linear classifier with weight vector $\mathbf{w}$ has been learnt on this transformed data which perfectly separates these two classes. Which of the following could be the transformed feature vector $\mathbf{x}$ that was used?

Note: This is a case of explicit feature transformation. The way to achieve this using the concepts learned in our course is using a kernel.

Options

(a)

$$\begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix}$$

(b)

$$\begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_1^2 \\ x_2^2 \end{bmatrix}$$

(c)

$$\begin{bmatrix} 1 \\ x_1^2 \\ x_2^2 \end{bmatrix}$$

(d)

$$\begin{bmatrix} x_1 \\ x_2 \\ x_1^2 \\ x_2^2 \end{bmatrix}$$

Answer

(b)

Solution

The decision boundary is a circle. The general equation of a circle is:

$$(x_1 - a)^2 + (x_2 - b)^2 = r^2$$

If this is expanded, we get:

$$x_1^2 + x_2^2 + (-2a)x_1 + (-2b)x_2 + (a^2 + b^2 - r^2) = 0$$

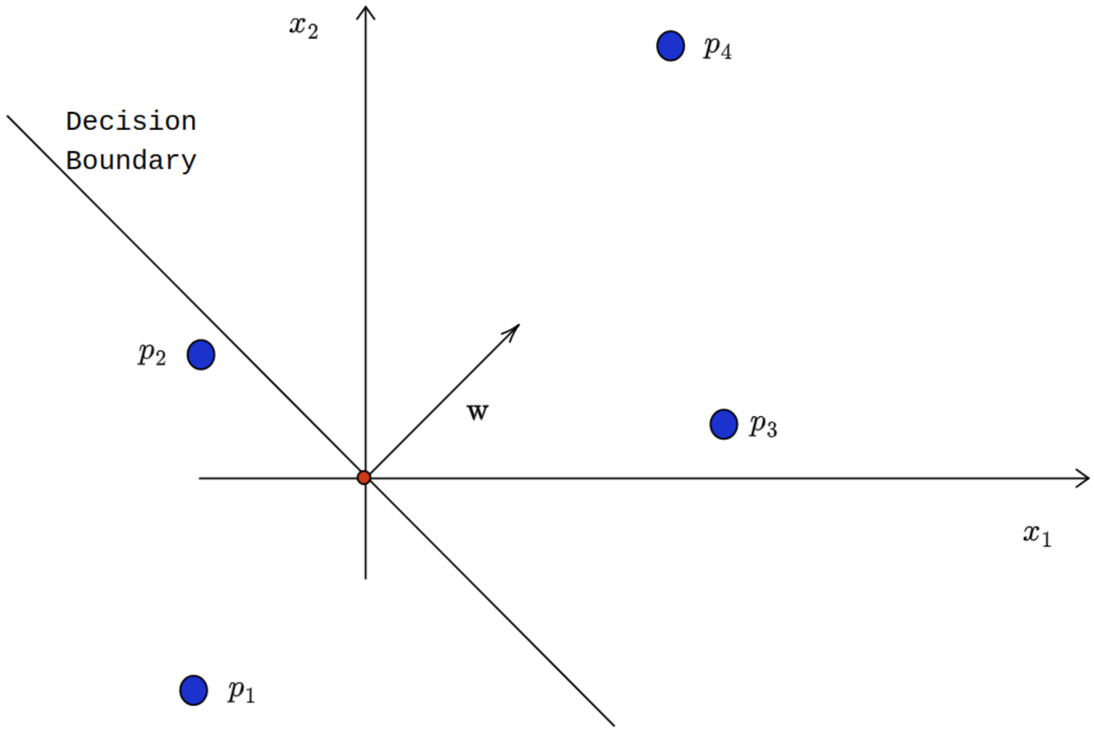
We see that the features required are:

$$[x_1^2 \quad x_2^2 \quad x_1 \quad x_2 \quad 1]$$

Question-5

Statement

A logistic regression model has been trained for a binary classification problem with labels 0 and 1. The weight vector and the corresponding decision boundary are displayed in the figure given below:



Now, the model is tested on four points. The probability corresponding to the i^{th} data-point $\mathbf{x}_i$ returned by the logistic regression model is given as follows:

$$P(y = 1 \mid \mathbf{x}_i) = p_i$$

We don't know the true labels for any of the four points. We are only talking about the predicted probabilities here. Which of the following relationships is correct?

Options

(a)

$$p_1 < p_2 < p_3 < p_4$$

(b)

$$p_1 > p_2 > p_3 > p_4$$

(c)

$$p_3 < p_4 < p_2 < p_1$$

(d)

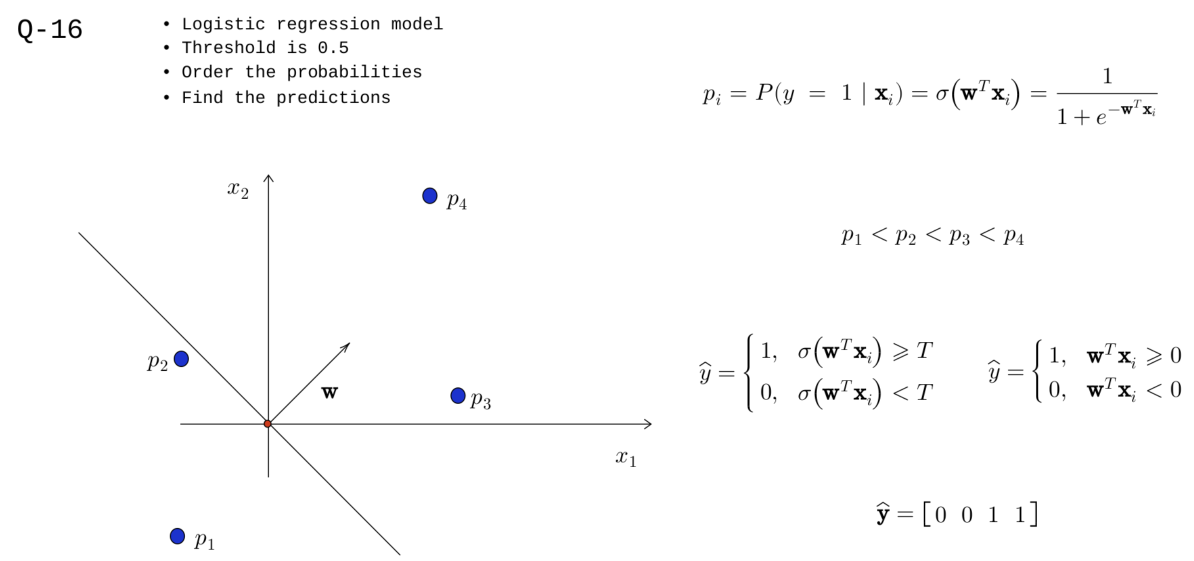
$$p_1 > p_2 \text{ and } p_4 > p_3$$

Answer

(a)

Solution

Here, note that $T = 0.5$.



Q-16

- Logistic regression model
- Threshold is 0.5
- Order the probabilities
- Find the predictions

$$p_i = P(y \ = \ 1 \mid \mathbf{x}_i) = \sigma\left(\mathbf{w}^T \mathbf{x}_i\right) = \frac{1}{1 + e^{-\mathbf{w}^T \mathbf{x}_i}}$$

$$p_1 < p_2 < p_3 < p_4$$

$$\widehat{y} = \begin{cases} 1, & \sigma\left(\mathbf{w}^T \mathbf{x}_i\right) \geqslant T \\ 0, & \sigma\left(\mathbf{w}^T \mathbf{x}_i\right) < T \end{cases} \qquad \widehat{y} = \begin{cases} 1, & \mathbf{w}^T \mathbf{x}_i \geqslant 0 \\ 0, & \mathbf{w}^T \mathbf{x}_i < 0 \end{cases}$$

$$\widehat{\mathbf{y}} = \begin{bmatrix} 0 & 0 & 1 & 1 \end{bmatrix}$$

Question-6

Statement

A logistic regression model is being trained on a dataset of size $2n$. The first n data-points belong to class-1 (label is 1) and the rest in class-0 (label is 0). Note that we are talking about the true label here.

$$\begin{aligned}\text{Class-1} &= \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}\\ \text{Class-0} &= \{\mathbf{x}_{n+1}, \cdots, \mathbf{x}_{2n}\}\end{aligned}$$

The probability output by the model at any step in the training process is given by:

$$P(y = 1 \mid \mathbf{x}_i) = p_i$$

Which of the following expressions is the negative log-likelihood of the model on this dataset? This is also called the binary cross entropy loss.

Options

(a)

$$\sum_{i=1}^{n} -\log p_i + \sum_{i=n+1}^{2n} -\log(1 - p_i)$$

(b)

$$\sum_{i=1}^{2n} -\log p_i$$

(c)

$$\sum_{i=1}^{2n} -\log(1 - p_i)$$

(d)

$$\sum_{i=1}^{2n} -p_i \log p_i$$

Answer

(a)

Solution

Recall that the negative log-likelihood is as follows:

$$\sum_{i=1}^{2n} -y_i \log[\sigma(\mathbf{w}^T \mathbf{x}_i)] - (1 - y_i) \log[1 - \sigma(\mathbf{w}^T \mathbf{x}_i)]$$

Using $p_i = \sigma(\mathbf{w}^T \mathbf{x}_i)$, we get:

$$\sum_{i=1}^{2n} -y_i \log p_i - (1 - y_i) \log(1 - p_i)$$

Since the first n data-points have $y_i = 1$ and the rest have $y_i = 0$, we get:

$$\sum_{i=1}^{n} - \log p_i + \sum_{i=n+1}^{2n} - \log(1 - p_i)$$

Question-7

Statement

Consider a logistic regression model that is trained on videos to detect objectionable content. Videos with objectionable content belong to the positive class (label 1). Harmless videos belong to the negative class (label 0).

A good detector should be able to correctly identify almost all videos that are objectionable. If it incorrectly classifies even a single video that has inappropriate content in it, that could have serious consequences, as millions of people might end up watching it. In this process the detector may classify some harmless videos as belonging to the positive class. But that is a price we are willing to pay.

How should we choose the threshold (for inference) of this logistic regression model?

Options

(a)

The threshold should be a low value.

(b)

The threshold should be a high value.

(c)

The performance of the classifier is independent of the threshold.

Answer

(a)

Solution

The threshold should be a low value, say something like 0.2. For example, even if we are only 25% confident of predicting a video as objectionable, we would like to go ahead and flag it.

Question-8

Statement

A logistic regression model has been trained on a dataset in a binary classification setup. It is now tested on two separate datasets, each having 14 data-points, 7 from each class. The loss (negative log-likelihood) of the **same model** on the two test-datasets is L_1 and L_2. It is also given that the classification accuracy of the model on both these test-datasets is 100%. Now consider the images of the two test datasets along with the decision boundary of the model:

Image for L_1:

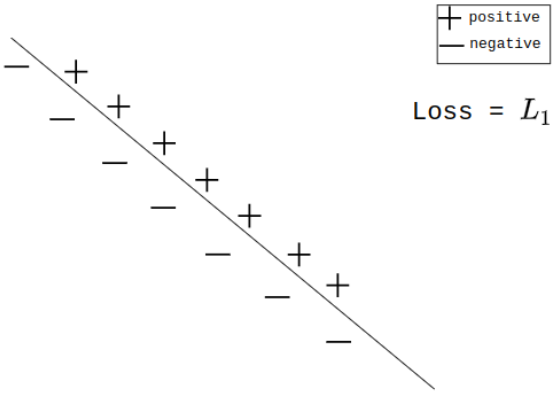
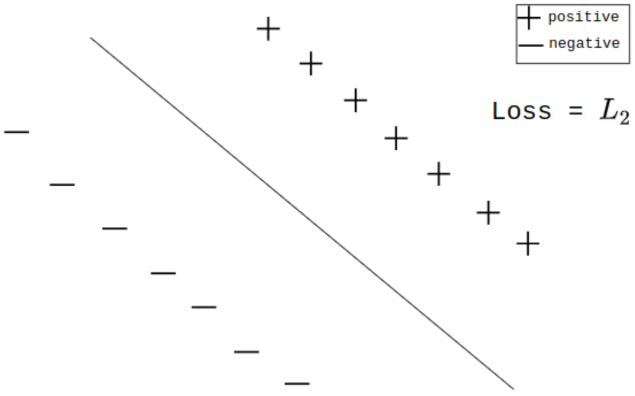


Image for L_2:



Which of the following statements are true? Assume that the loss is computed mathematically to arbitrary precision. For example, 10^{-20} is not rounded off to 0. The label is 1 for the positive class and 0 for the negative class.

Options

(a)

The loss of both the models is equal to 0. That is, $L_1 = L_2 = 0$.

(b)

$$L_1 > L_2$$

(c)

$$L_1 < L_2$$

(d)

$$L_1 = L_2$$

Answer

(b)

Solution

Consider the negative log-likelihood:

$$\sum_{i=1}^{n} -y_i \log p_i - (1 - y_i) \log(1 - p_i)$$

In the case of L_1, the points are very close to the decision boundary. p_i will be close to 0.5 for all these points, some slightly above 0.5 and the others slightly below 0.5. In the case of L_2, p_i will be close to 1 for the positively labeled points and close to 0 for the negatively labeled points. Therefore, $-\log p_i$ will be close to 0 for positively labeled data-points and $-\log(1 - p_i)$ will be close to zero for the negatively labeled data-points. In other words, L_2 will be very close to 0 while L_1 will have a higher value compared to L_2.

Question-9

Statement

Consider a training dataset of n data-points for a binary classification problem that satisfies the following condition for all i in $1, \cdots, n$:

$$x_{ij} = \begin{cases} a, & y_i = 1 \\ b, & y_i = -1 \end{cases}$$

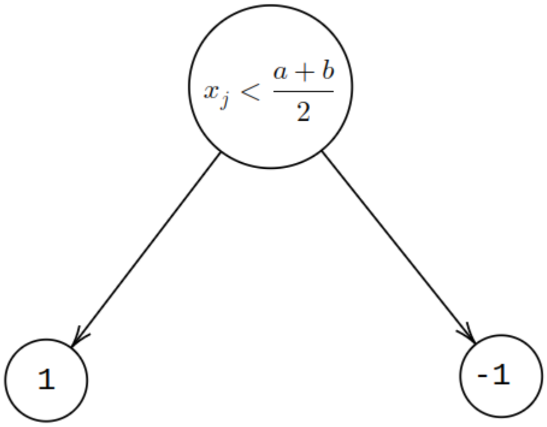
where, x_{ij} is the j^{th} feature for the i^{th} data-point and a and b are real numbers with $a \neq b$. If an AdaBoost model is fit on this dataset, how many rounds would be required to get a good classifier?

Answer

1

Solution

The j^{th} feature for all data-points from the positive class is a and it is b for all data-points from the negative class. Assume without loss of generality that $a < b$, then the following decision stump separates the data perfectly with zero misclassifications:



This will be the stump that will be learned in the first round. Since it is already perfect, one more round is not necessary.

Note Decision stumps are still weak learners. The reason a stump performs so well here is because the data is absurdly simple to classify. In reality, datasets will never be so simple. For such a simple dataset, one would never even consider boosting in the first place since a decision stump already works well.

Question-10

Statement

Which of the following models has the potential to achieve zero training error on every possible training dataset in $\mathbb{R}^2$?

Options

(a)

Decision Tree

(b)

Logistic Regression

(c)

Soft Margin Linear-SVM

(d)

Soft Margin Kernel-SVM with cubic kernel

Answer

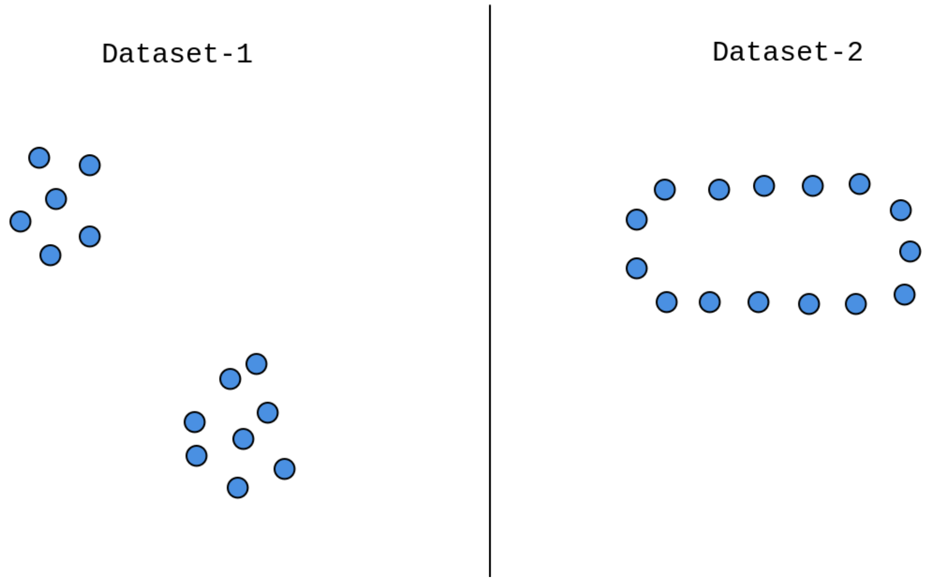
(a)

Logistic regression and soft-margin linear-SVM being linear models will produce zero training error only on linearly separable data. Soft-margin kernel-SVM with cubic kernel can be used on more complex datasets that are separable with a non-linear decision boundary. But every possible dataset cannot be separated with zero training error using a cubic-kernel SVM. A decision tree can be grown on any dataset until the leaves are perfectly pure. This will result in zero training error.

Question-11

Statement

Consider the two datasets given below:



If K-means algorithm with $k = 2$ is applied on each of these datasets, which of the following statements is true?

Options

(a)

The algorithm will terminate only in the case of dataset-1 after a certain number of iterations.

(b)

The algorithm will **never** terminate in the case of dataset-2 and will keep oscillating between different cluster configurations.

(c)

The algorithm will terminate for both datasets.

(d)

The algorithm will not terminate for both datasets.

Answer

(c)

Solution

Lloyd's algorithm will always terminate. The objective function that we are minimizing strictly decreases in every iteration and will stop decreasing after a finite number of iterations. This is independent of the configuration of the data-points and the choice of initial cluster centers.

Question-12

Statement

Consider the following data-points that make up the training dataset in a binary classification problem. They are of the form (x_1, x_2):

$$
\begin{gathered}
(-5,-3),(-4,1),(3,2),(4,5),(2,1) \\
(15,1),(21,-10),(8,4),(7,0),(9,-10)
\end{gathered}
$$

The first row has points which are labeled 0. The second row has points which are labeled 1. If you are allowed to ask a question of the form $f_k < \theta$, what is the information gain corresponding to the "best" question? Use $\log_2$ as always.

Answer

1

Solution

Notice that the question $x_1 < 6$ perfectly separates the data. This is one of the best questions we could ask. The entropy of the parent is 1. The entropy of each leaf is 0. Hence, the information gain is 1.

Question-13

Statement

Consider the following data-points for a regression problem:

$$\{(c_1 \cdot \mathbf{u}, y_1), \cdots, (c_n \cdot \mathbf{u}, y_n)\}$$

Here, c_i is some real number and $\mathbf{u} \in \mathbb{R}^d$ and $\mathbf{u} \neq 0$. Fit a linear regression model for this dataset and find the predicted value for the test-point $\mathbf{x}_{\text{test}} = 5 \cdot \mathbf{u}$. You can assume that $c_i \neq 0$ for some i. The following values are given to you:

$$\sum_i c_i \cdot y_i = 20, \quad \sum_i c_i^2 = 100$$

Answer

1

Solution

The loss is:

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

The gradient is:

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

$$= 2 \cdot \sum_{i=1}^{n} [(\mathbf{w}^T \mathbf{u})c_i^2 - c_i y_i]\mathbf{u}$$

$$= 0$$

$$\implies \mathbf{w}^T \mathbf{u} = \frac{\sum\limits_{i=1}^{n} c_i y_i}{\sum\limits_{i=1}^{n} c_i^2}$$

$$= 0.2$$

Now, the prediction for a point $5 \cdot \mathbf{u}$ is $\mathbf{w}^T(5\mathbf{u})$ which turns out to be 1.

Question-14

Statement

Let $f(\mathbf{w})$ be the the loss function for a linear regression problem. $\mathbf{X}$ is a $d \times n$ data-matrix, $\mathbf{w} \in \mathbb{R}^d$ and $\mathbf{y} \in \mathbb{R}^n$. Consider the following steps:

Step-1: f is a convex function.

Step-2: Every local minimum of f is a global minimum.

Step-3: There is a unique solution $\mathbf{w}'$ that minimizes f.

Step-4: If $\mathbf{X}\mathbf{X}^T$ is invertible, the unique solution to the minimization problem is given by $(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{y}$.

Which step is incorrect?

Options

(a)

Step-1

(b)

Step-2

(c)

Step-3

(d)

Step-4

(e)

All steps are correct. There are no incorrect steps.

Answer

(c)

Solution

The value of $\mathbf{w}$ that minimizes the squared loss is not unique. However, every $\mathbf{w}$ that minimizes the loss will give the same value of $f(\mathbf{w})$. In our course, the optimal solution is:

$$\mathbf{w}^* = \left[\mathbf{X}\mathbf{X}^T\right]^\dagger \mathbf{X}\mathbf{y}$$

This $\mathbf{w}^*$ is one possible solution. But it has the property that it has the least-norm among all possible solutions for the problem. When we refer to $\mathbf{w}^*$, it will always be this least-norm solution obtained using the pseudo-inverse.

Question-15

Statement

In the context of AdaBoost algorithm, for a classifier to be termed a weak learner, what should be the misclassification rate for it? The misclassification rate is the proportion of points misclassified by the classifier. Choose the most appropriate option.

Options

(a)

It should be slightly greater than 0.5.

(b)

It should be equal to 0.5.

(c)

It should be slightly less than 0.5.

(d)

It should be nearly zero.

Answer

(c)

Solution

The weak learner in AdaBoost should be slightly better than a random classifier. A random classifier that chooses the predicted label uniformly at random will have a misclassification rate of 0.5, that is, it will misclassify roughly half the examples in the dataset. A classifier that is slightly better than this will perform fewer mistakes and will hence have a misclassification rate that is slightly less than 0.5.

Question-16

Statement

Consider a soft-margin linear-SVM that has been trained for a binary classification problem. Now, consider a collection of 10 new data-points that are added to the positive class on the correct side of the margin, but very far away from the corresponding supporting hyperplane. If the SVM is "retrained" on this expanded dataset, what will happen to the decision boundary?

Options

(a)

The decision boundary will change drastically and reorient itself so that the distance of the newly added points from the boundary is minimized.

(b)

The decision boundary will change marginally.

(c)

The decision boundary will remain the same.

(d)

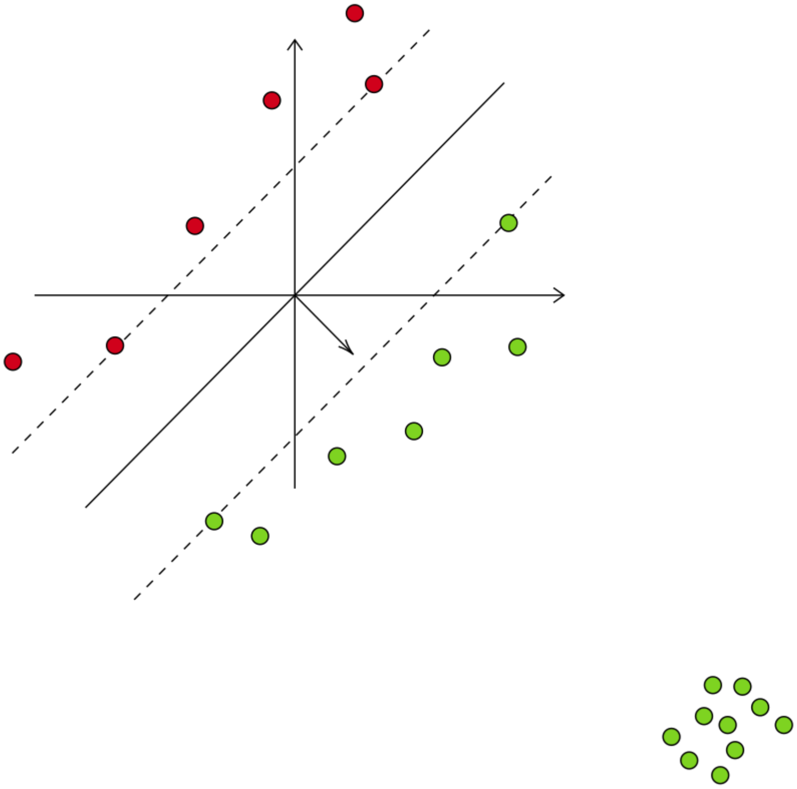
Cannot comment on the nature of the decision boundary with the given information.

Answer

(c)

Solution

Since the newly added data-points are far away from the supporting hyperplane and on the correct side, they don't disturb the feasible region in the primal problem. Since the feasible region remains the same, the solution to the primal problem remains the same. Whatever w^* is optimal for the original primal problem remains optimal for new problem.



Question-17

Statement

The accuracy of a classifier on a dataset is defined as the proportion of points in the dataset that are correctly classified by it. If $\mathbf{w}$ is the weight of a linear classifier that has an accuracy of 0.85 on a dataset, what is is the accuracy of a linear classifier whose weight is $\frac{\mathbf{w}}{2}$?

Answer

0.85

Solution

The prediction for a linear classifier depends on the sign of $\mathbf{w}^T\mathbf{x}$. Scaling $\mathbf{w}$ by a positive constant doesn't alter the sign of $\mathbf{w}^T\mathbf{x}$. Hence, the accuracy of the scaled classifier remains is the same as the accuracy of the original classifier.

Question-18

Statement

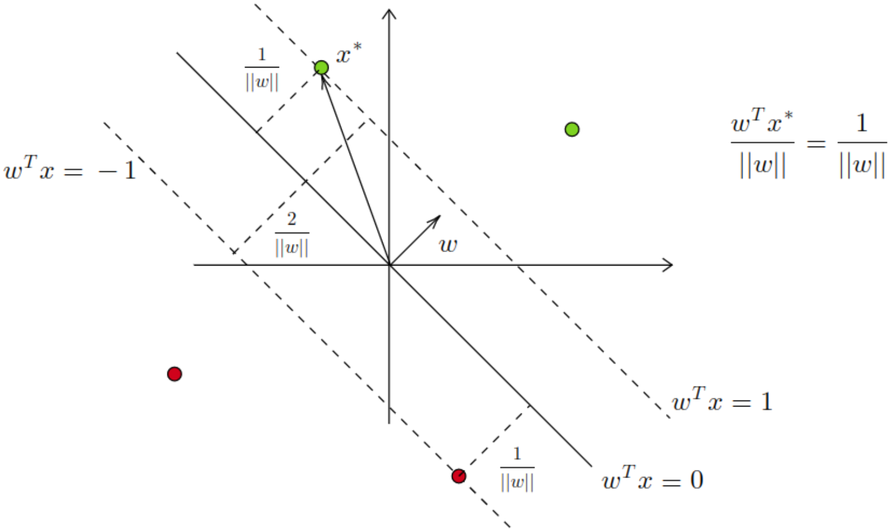
In a hard-margin SVM problem, let $\mathbf{w}^* = [1 \quad 2 \quad 3]^T$ be the optimal weight vector. What is the distance of the closest point in the dataset from the decision boundary? Enter your answer correct to three decimal places.

Answer

[0.25, 0.28]

Solution

Hard-margin, linear-SVM is a linear classifier that maximizes the margin. The margin is the distance of the nearest point to the boundary. If the closest point lies on $w^T x = \pm 1$, then the margin turns out to be $\dfrac{1}{||w||}$.



The nearest point to the boundary is at a distance of $1/||\mathbf{w}||^* = 0.267$

Question-19

Statement

Let the eigenvectors of a covariance matrix for a centered dataset in $\mathbb{R}^3$ be $\mathbf{w}_1$, $\mathbf{w}_2$ and $\mathbf{w}_3$. These three directions specify a new coordinate system to represent the data. If the entire dataset is represented in terms of these new coordinates, what can you say about the covariance matrix in this new coordinate system?

Options

(a)

The covariance matrix is invariant to change of coordinates

(b)

The covariance matrix becomes diagonal

(c)

The covariance matrix becomes identity

(d)

Insufficient information to answer this question

Answer

(b)

Solution

Let $\mathbf{Q}$ be the orthogonal matrix whose columns are the eigenvectors of $\mathbf{C}$. Then, we can decompose $\mathbf{C}$ as:

$$\mathbf{C} = \mathbf{Q}\mathbf{D}\mathbf{Q}^T$$

This comes from the spectral theorem. Here, $\mathbf{D}$ is the diagonal matrix of eigenvalues of $\mathbf{C}$. Now, let us represent all the data-points in the new basis of the eigenvectors of $\mathbf{C}$:

$$\mathbf{X}' = \begin{bmatrix} - & \mathbf{w}_1^T & - \\ & \vdots & \\ - & \mathbf{w}_d^T & - \end{bmatrix} \begin{bmatrix} | & & | \\ \mathbf{x}_1 & \cdots & \mathbf{x}_n \\ | & & | \end{bmatrix} = \mathbf{Q}^T\mathbf{X}$$

The covariance matrix of $\mathbf{X}'$ is now:

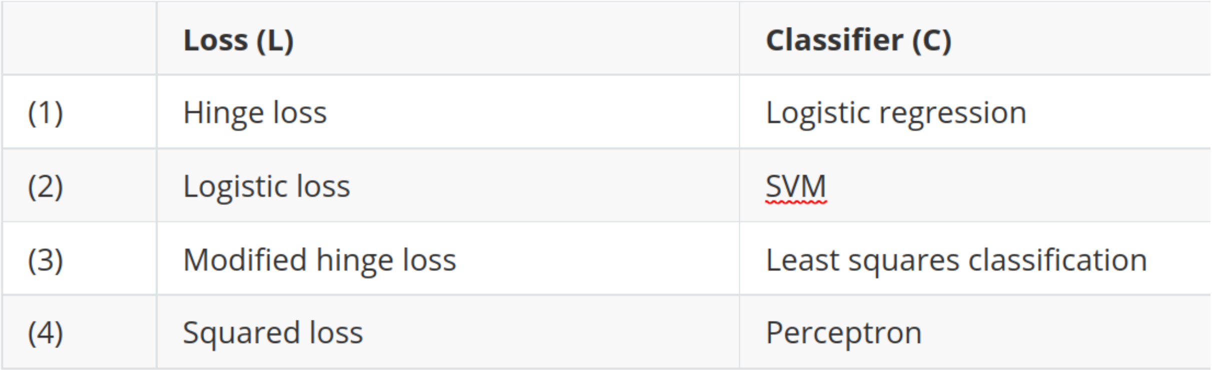
$$
\begin{aligned}
\frac{1}{n} \cdot \mathbf{X}'\mathbf{X}'^T &= \frac{1}{n} \cdot \mathbf{Q}^T\mathbf{X}\mathbf{X}^T\mathbf{Q} \\
&= \mathbf{Q}^T\mathbf{C}\mathbf{Q} \\
&= \mathbf{D}
\end{aligned}
$$

We see that the covariance matrix of the dataset in the new coordinate system is diagonal.

Question-20

Statement

Map the loss function to the classifier for which it is used:



	Loss (L)	Classifier (C)
(1)	Hinge loss	Logistic regression
(2)	Logistic loss	SVM
(3)	Modified hinge loss	Least squares classification
(4)	Squared loss	Perceptron

Options

(a)

A1-C1, A2-C2, A3-C3, A4-C4

(b)

A1-C2, A2-C1, A3-C4, A4-C3

(c)

A1-C4, A2-C2, A3-C2, A4-C3

(d)

A1-C1, A2-C3, A3-C4, A4-C2

Answer

(b)