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

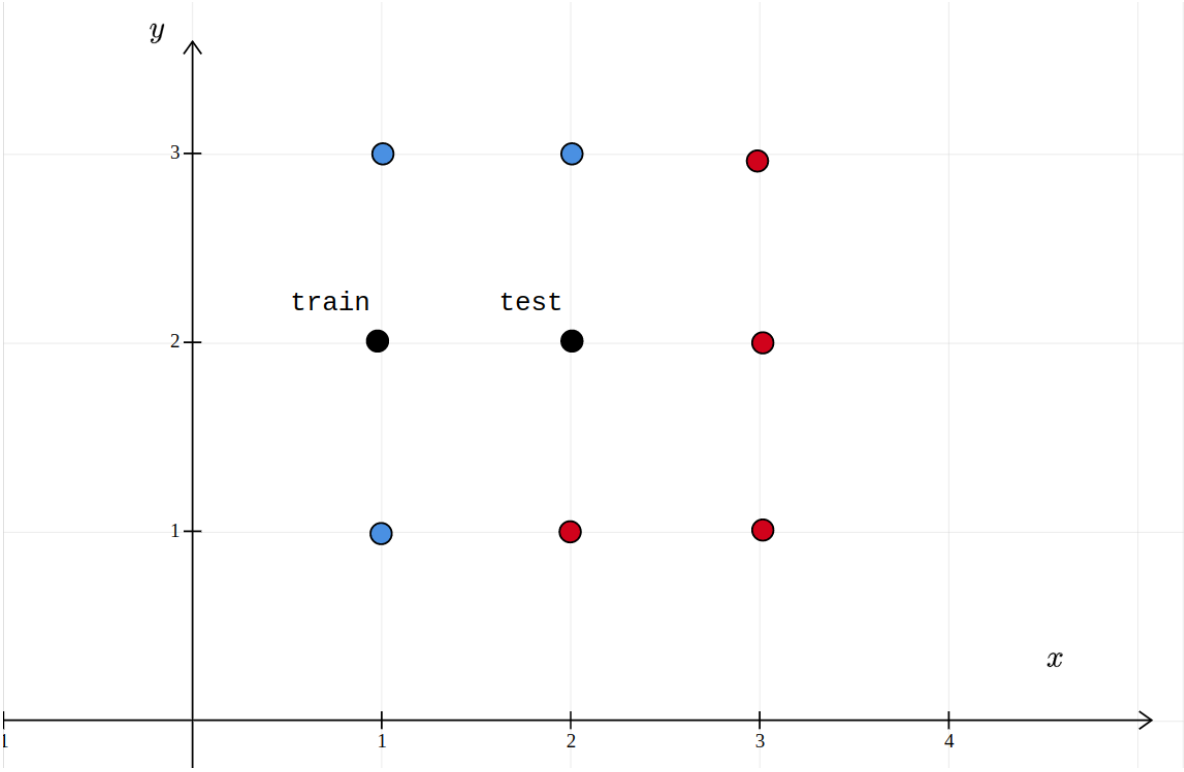
The entire training dataset has to be stored in memory. For predicting the label of a test-point, we have to perform the following steps:

- 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:

- $d(\mathbf{x}_{\text{test}}, \mathbf{x}_1)^2 = 5$
- $d(\mathbf{x}_{\text{test}}, \mathbf{x}_2)^2 = 149$
- $d(\mathbf{x}_{\text{test}}, \mathbf{x}_3)^2 = 14$
- $d(\mathbf{x}_{\text{test}}, \mathbf{x}_4)^2 = 2$
- $d(\mathbf{x}_{\text{test}}, \mathbf{x}_5)^2 = 134$

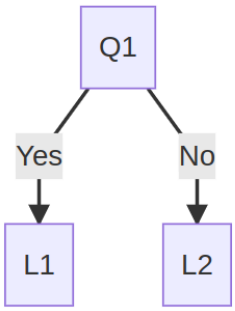
We see that among the three nearest neighbors, two have label 1 and one has label 0. Hence the predicted label is 1. For those interested in a code for the same:

```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:

$$IG = 1 - \left(\frac{80}{200} 0.954 + \frac{120}{200} 0.980 \right) \approx 0.030$$

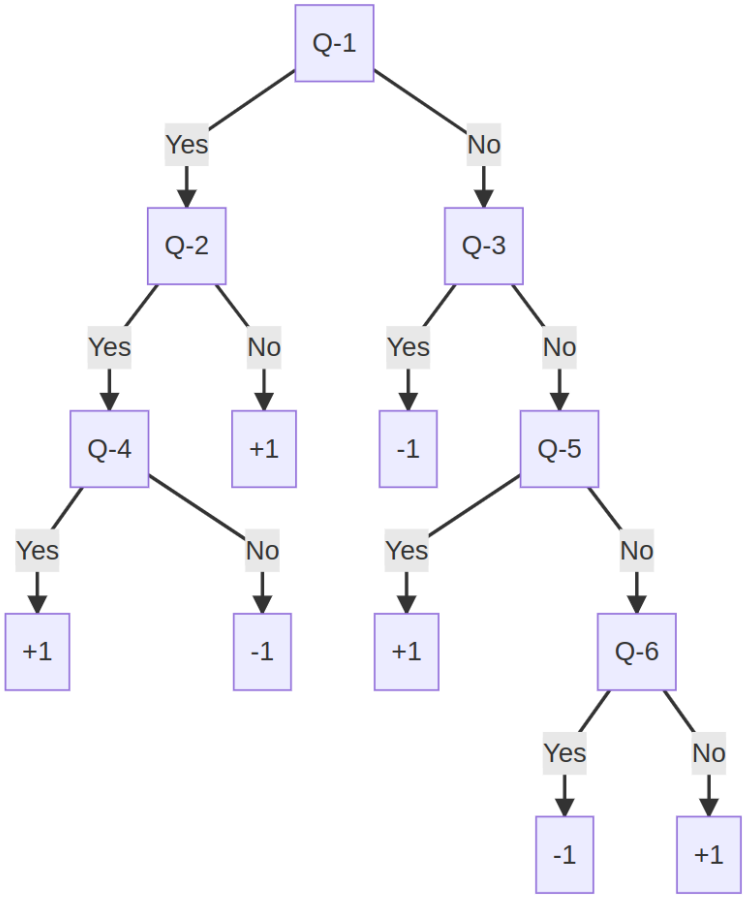
Code for reference:

```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)

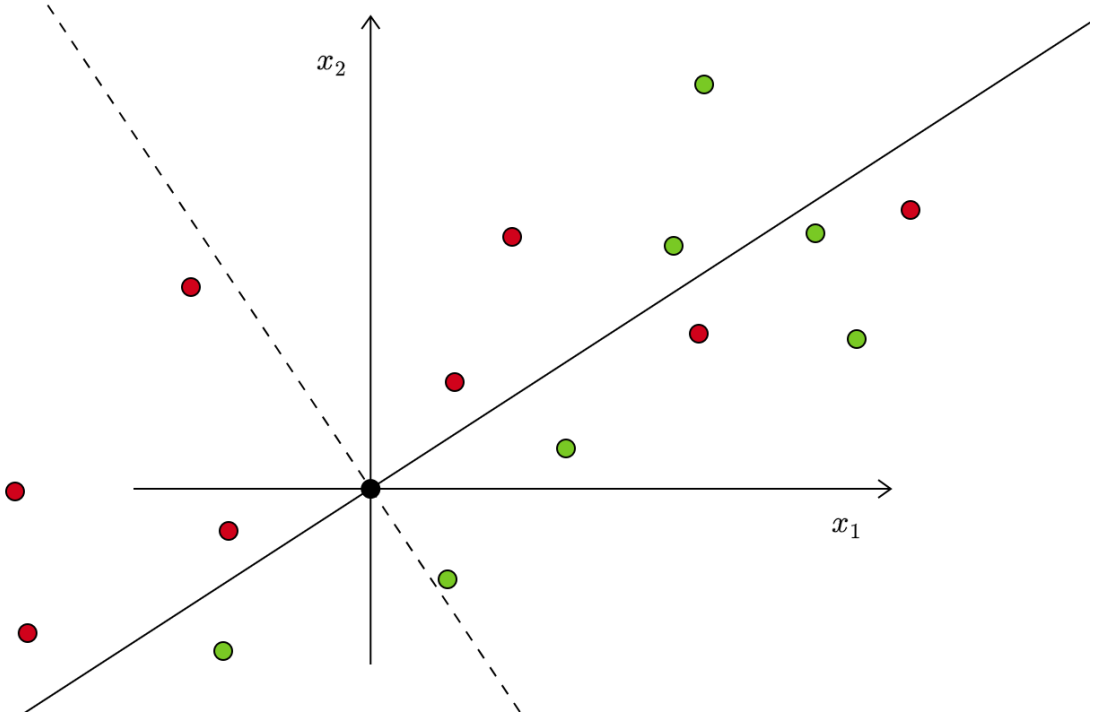
Solution

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.

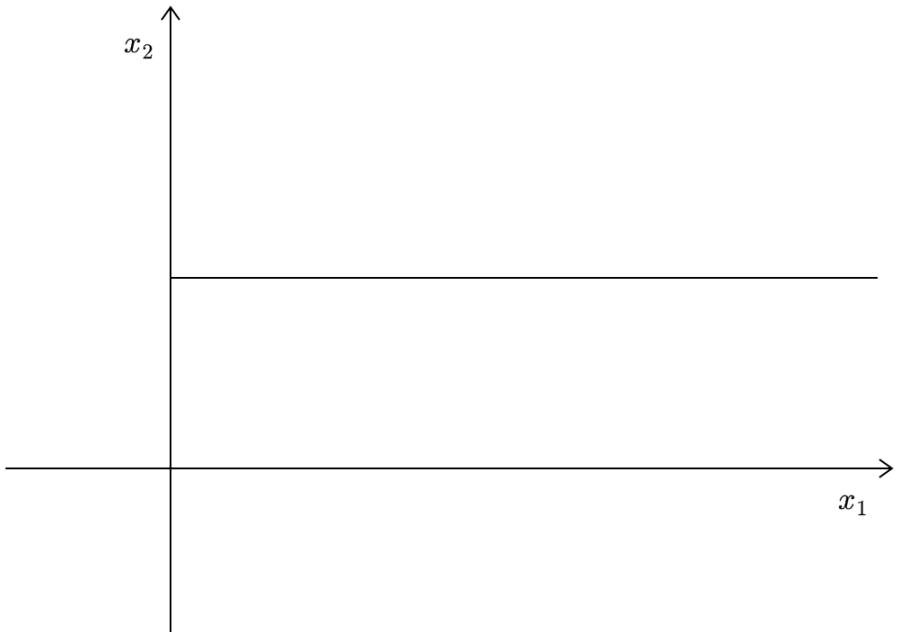
Question-10

Statement

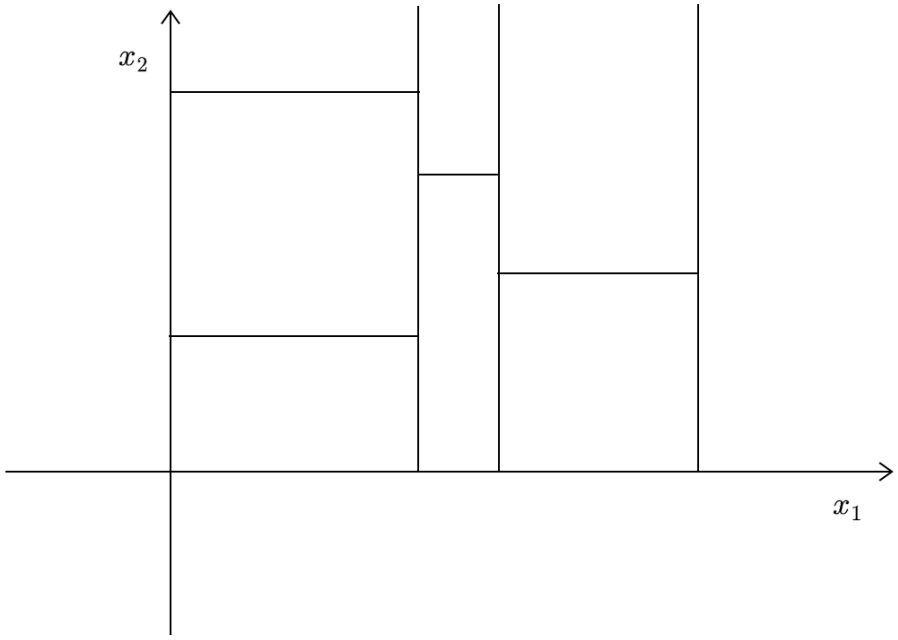
Which of the following are valid decision regions for a decision tree classifier for datapoints in $\mathbb{R}^2$? The question in every internal node is of the form $f_k \leq \theta$. Both the features are positive real numbers.

Options

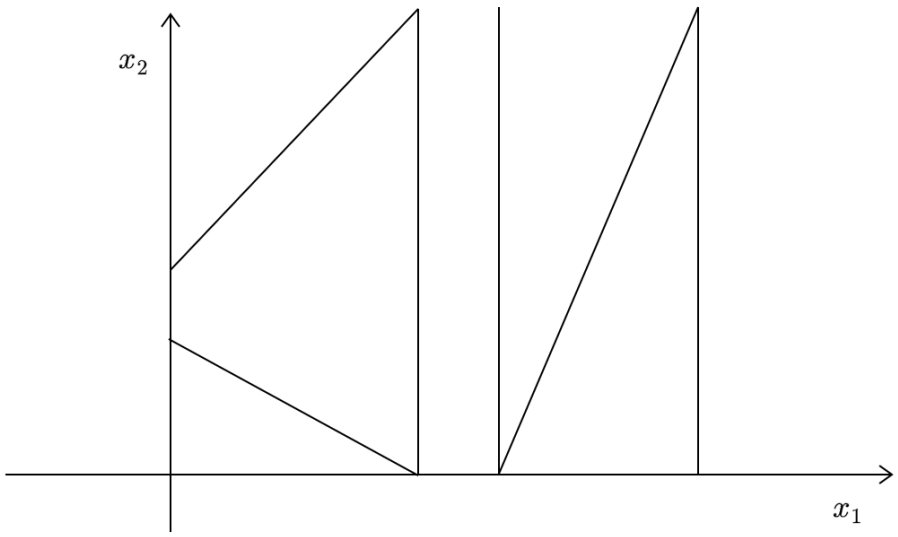
(a)



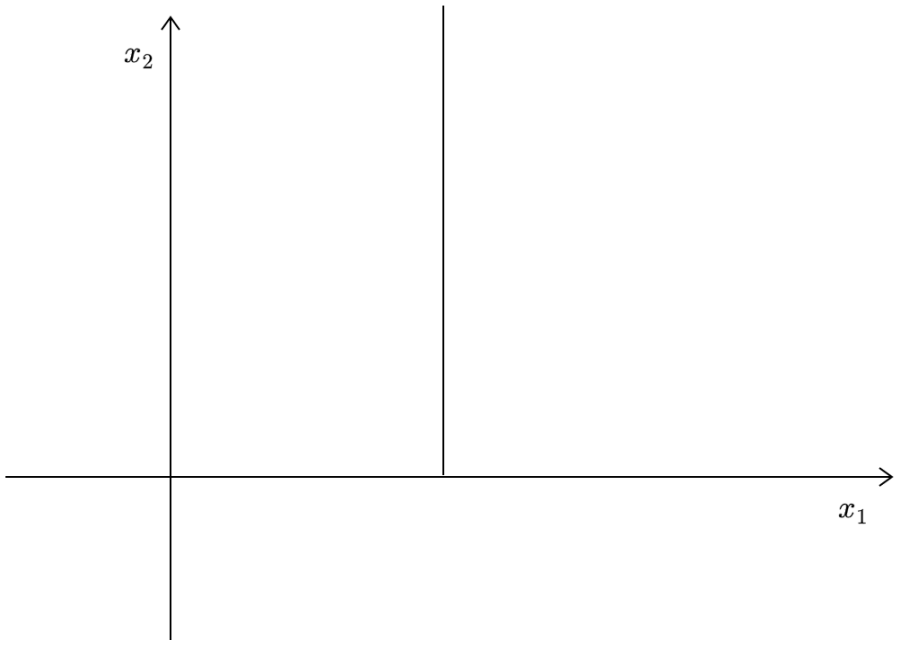
(b)



(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.