What will be the result of applying ReLU to the following values?:
-2.7, 3.9, -1.0, 4.2, 6.4, -7.3

Options :

6406531512277. ✖ 1, 3.9, 1, 4.2, 6.4, 1

6406531512278. ✖ 0, 1, 0, 1, 1, 0

6406531512279. ✖ -2.7, 0, -1.0, 0, 0, -7.3

6406531512280. ✖ -1, +1, -1, +1, +1, -1

6406531512281. ✔ 0, 3.9, 0, 4.2, 6.4, 0 ✓✓

$ReLU = max(0, value)$

$max(0, -2.7) = 0$

$max(0, 3.9) = 3.9$

Suppose you run gradient descent for linear regression for 500 iterations with a learning rate 0.02. You observe that the training loss (sum of squared loss) is increasing after every iteration. What may be the reason? What changes would you make to the set-up for the gradient descent to converge to a solution?

Options :

6406531512285. ✖ Number of features in the training data may be too low, try increasing them.

6406531512286. ✖ Number of features in the training data may be too high, try reducing them.

6406531512287. ✖ Learning rate may be too low; Try increasing it.

6406531512288. ✔ Learning rate may be too high; Try reducing it. ✓

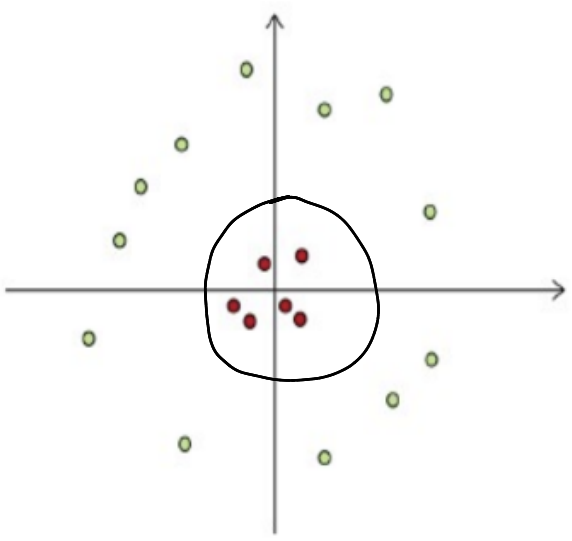
Gradient Descent

$\omega^{t+1} = \omega^t - \eta_t \nabla f(\omega^t)$

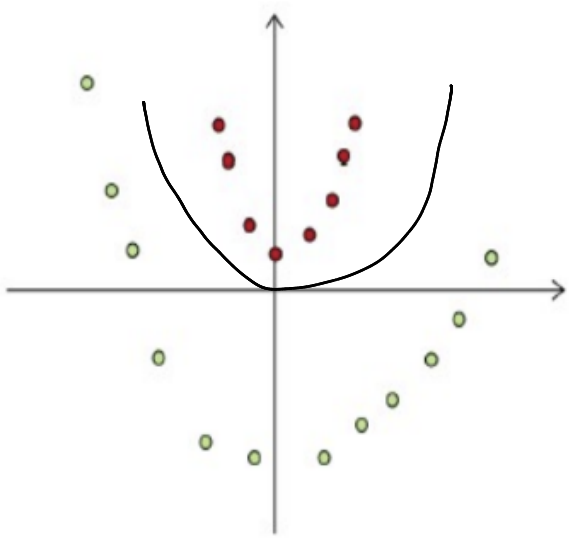
Learning Rate

Each of the datasets given below corresponds to a binary classification problem. The labels are `red` and `green` for the two classes.

Dataset-1



Dataset-2



On which of these two datasets can we train a hard-margin, kernel-SVM with quadratic kernel?

Options :

6406531512241. ✖ Only dataset-1

6406531512242. ✖ Only dataset-2

6406531512243. ✔ On both dataset-1 and dataset-2

6406531512244. ✖ Neither dataset-1 nor dataset-2

= Decision boundry

let $k : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be a valid kernel. Is $x_1^3 x_2^3 k(x_1, x_2)$ a valid kernel? Here, $x_1, x_2 \in \mathbb{R}$?

Options :

6406531512249. ✔ Yes

6406531512250. ✖ No

$$K(x_1, x_2) = \phi(x)^T \phi(x)$$

$$x_1^3 x_2^3 k(x_1, x_2) = x_1^3 x_2^3 \phi(x)^T \phi(x)$$

$$\Rightarrow x_1^3 \phi(x)^T \; x_2^3 \phi(x)$$

$$\Rightarrow \phi'(x)^T \phi'(x) \Rightarrow \text{Valid}$$

$$\phi'(x) = x^3 \phi(x)$$

A set of data points is generated by the following process:

$y_i = w_0 + w_1 x_i + w_2 x_i^2 + w_3 x_i^3 + w_4 x_i^4 + w_5 x_i^5 + \epsilon$ where ϵ is a Gaussian noise.

You use two models to fit the data:

Model 1: $\hat{y} = a_0 + a_1 x + a_2 x^2$ (underfit)

Model 2: $\hat{y} = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \ldots + a_{10} x^{10}$ (Overfit)

Using a fixed number of training examples, Model 2 will have _____ bias than Model 1, and Model 1 is more likely to _____

Options :

6406531512289. ✖ Higher, underfit

6406531512290. ✖ Lower, overfit

6406531512291. ✖ Higher, overfit

6406531512292. ✔ Lower, underfit

Overfit Models have,
- High Variance
- Low Bias

Underfit models have,
- High Bias
- Low Variance

In model-2, if x changes slightly, the $\hat{y}$ will change a lot.
$\Rightarrow$ Model-2 has high variance.
$\Rightarrow$ Model-2 is Overfit.

Consider the following linearly separable training dataset for a binary classification problem in $\mathbb{R}^2$:

$$\mathbf{x}_1 = \begin{bmatrix} 0 \\ -1 \end{bmatrix}, y_1 = 1 \qquad \mathbf{x}_2 = \begin{bmatrix} -2 \\ 3 \end{bmatrix}, y_2 = 1$$

$$\mathbf{x}_3 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, y_3 = -1 \qquad \mathbf{x}_4 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, y_4 = -1$$

A hard-margin, linear-SVM is trained on this dataset. Among the four options given below, one of them is the optimal weight vector $\mathbf{w}^*$. Identify this vector. Recall that the optimal weight vector is the solution to the primal problem.

Options :

6406531512232. ✔ $\begin{bmatrix} -2 \\ -1 \end{bmatrix}$ ω_1

6406531512233. ✖ $\begin{bmatrix} -100 \\ -50 \end{bmatrix}$ ω_2

6406531512234. ✖ $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ ω_3

6406531512235. ✖ $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$ ω_4

$(\omega^T x_i) y_i \geq 1$

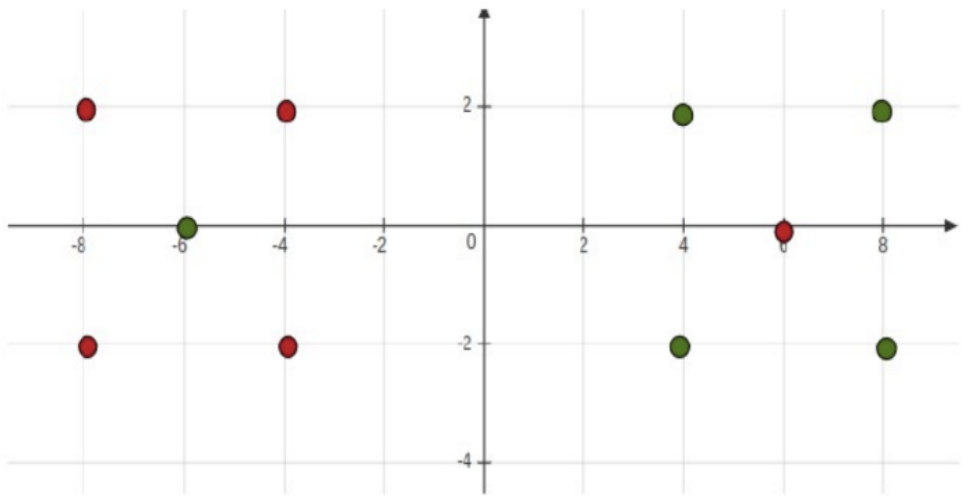
constraint for optimal ω^*.

$(\omega^T x_1) y_1$ $(\omega^T x_2) y_2$ $(\omega^T x_3) y_3$ $(\omega^T x_4) y_4$

ω_1 $\begin{bmatrix} -2 & -1 \end{bmatrix} \begin{bmatrix} 0 \\ -1 \end{bmatrix} (1)$ $\begin{bmatrix} -2 & -1 \end{bmatrix} \begin{bmatrix} -2 \\ 3 \end{bmatrix} (1)$ $\begin{bmatrix} -2 & -1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} (-1)$ $\begin{bmatrix} -2 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} (-1)$

Similarly check for ω_2 ω_3 ω_4

Consider the following data set:



Which of the following will have lower leave-one-out cross-validation error?

Options :

6406531512293. ✖ 1-Nearest Neighbor

6406531512294. ✔ 3-Nearest Neighbor

Consider that we introduce negative marking in this exam. After getting the results, you observe that eight of your friends $\{f_1, \ldots, f_8\}$ have scored the following marks respectively:

{5, 6, -2, -3, 1, 7, -4, -1}

You want to cluster your friends into two groups based on their marks by using the Lloyd's algorithm.

You initialize the algorithm by keeping the first four friends, i.e., $\{f_1, f_2, f_3, f_4\}$ in cluster 1 (C_1) and the last four friends, i.e., $\{f_5, f_6, f_7, f_8\}$ in cluster 2 (C_2).

How would the clusters look like after executing one step of Lloyd's algorithm?

Options :

6406531512295. ✖ $C_1: (f_1, f_2, f_3, f_4), C_2: (f_5, f_6, f_7, f_8)$

6406531512296. ✖ $C_1: (f_1, f_2, f_5, f_6), C_2: (f_3, f_4, f_7, f_8)$

6406531512297. ✖ $C_1: (f_2, f_4, f_6), C_2: (f_1, f_3, f_5, f_7, f_8)$

6406531512298. ✔ $C_1: (f_1, f_2, f_6), C_2: (f_3, f_4, f_5 f_7, f_8)$

$$\mu_1 = \frac{5+6+(-2)+(-3)}{4}$$

$$\Rightarrow \quad 1.5$$

$$\mu_2 = \frac{1+7+(-4)+(-1)}{4}$$

$$\Rightarrow \quad 0.75$$

$-2, -3, 1, -4, -1$

$\downarrow \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow$

$f_3 \quad f_4 \quad f_5 \quad f_7 \quad f_8$ are closer to μ_2.

Let w^*, ξ^* be the optimal primal solutions, and α^*, β^* be the optimal dual solutions of the soft-margin SVM problem. Select the options which are always true.

Options :

6406531512267. ✔ a)

If $\beta_i^* = 0$, the i^{th} point is either incorrectly classified by w^* or correctly classified with a margin less than 1.

6406531512268. ✖ If the i^{th} data point pays a nonzero bribe, then $\alpha_i^* = 0$. b)

6406531512269. ✖ If i^{th} data point lies on one of the supporting hyperplanes, then $\alpha_i^* = C$. c)

6406531512270. ✔ If i^{th} data point lies on the correct supporting hyperplane, it does **not** pay any bribes. d)

Keep these conditions in mind,

$\alpha_i^* + \beta_i^* = C$ —— (1)

$\alpha_i^* (1 - w^{*T} x_i y_i - \xi_i^*) = 0$ —— (2)

$\beta_i^* (\xi_i^*) = 0$ —— (3)

Also, ξ_i^* is called bribe

a) if $\beta_i^* = 0$,

From (1)
$\alpha_i^* + 0 = C$
$\Rightarrow \alpha_i^* = C$

From (3)
if $\beta_i^* = 0$
$\Rightarrow \xi_i^* \geq 0$

From (1)
if $\alpha_i^* = C$
$\Rightarrow (1 - w^{*T} x_i y_i - \xi_i^*) = 0$
$\Rightarrow 1 - w^{*T} x_i y_i = \xi_i^*$
$\Rightarrow 1 - w^{*T} x_i y_i \geq 0$
$\Rightarrow w^{*T} x_i y_i \leq 1$

When $w^{*T} x_i y_i \leq 1$, either points are classified incorrectly or classified correctly but not with enough margin.

d) No bribes means $\xi_i^* = 0$

From ③
$\Rightarrow \beta_i^* \geq 0$

From ①
$\alpha_i^* \geq 0$

From ②
if $\alpha_i^* \geq 0$
$\Rightarrow 1 - w^{*T} x_i y_i - \xi_i^* = 0$
$\Rightarrow 1 - w^{*T} x_i y_i - 0 = 0$
$\Rightarrow w^{*T} x_i y_i = 1$

When $w^{*T} x_i y_i = 1$, the points lie on supporting hyperplanes and are classified correctly.

Which of the following estimators are more likely to be preferred for bagging? Select all that apply.

Options :

6406531512271. ✖ A decision stump (underfit)

6406531512272. ✖ A decision stump with randomly selected features for splitting the nodes (underfit)

6406531512273. ✔ *k* -NN classifier with a smaller value of *k*. (many clusters, overfit)

6406531512274. ✖ *k*-NN classifier with a larger value of *k* . (underfit)

6406531512275. ✔ A fully grown decision tree with randomly selected features for splitting the nodes (overfit)

Bagging (Bootstrap Aggregation) helps in reduction of variance of weak learners (models).

We know that overfit models have high variance.
⇒ To reduce variance, Bagging will be used.
⇒ Overfit models are preferred for Bagging.

Let $\{15,\ 21.3,\ 8.5,\ 2,\ 40,\ 33,\ 28.4\}$ be 7 points sampled independently and **uniformly** from $[0, a]$ for some unknown $a > 0$. Find the maximum likelihood estimator $\hat{a}_{ML}$ of a given these samples.

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

40

Here,

$x_i \in \text{Uniform}[0, a]$

$\downarrow \quad \downarrow$

$x \quad y$

$$L(x) = \prod_{i=1}^{n} \frac{1}{y-x} = \frac{1}{(y-x)^n}$$

$$\Rightarrow \log(L(x)) = \log\left(\frac{1}{(y-x)^n}\right)$$

$$\Rightarrow n\left[\log(1) - \log(y-x)\right]$$

$$\Rightarrow n\left[0 - \log(y-x)\right]$$

$$\Rightarrow -n\log(a-0) \qquad (x=0, y=a)$$

$$\Rightarrow -n\log(a)$$

Taking derivative w.r.t. a

$$\Rightarrow \frac{-n}{a} = 0$$

$$\Rightarrow \max(x_1, x_2, x_3 \ldots\ldots x_n)$$

$\Rightarrow$ Here max value is $\underline{\underline{40}}$.

Consider a linearly separable binary classification data set with 500 data points and 50 features. Assume that there exists a w such that $||w|| = 1, y_i(w^T x_i) \geq 0.25 \; \forall i$. Also assume that $||x||_2 \leq 1 \; \forall i$

What is the maximum number of mistakes that the Perceptron algorithm can make in this data set?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

16

mistakes $\leq \frac{R^2}{\gamma^2}$

R = distance of farthest point in dataset

γ = distance of hyper-planes

Given that,

$||x_i|| \leq 1 \quad \forall i$

$\Rightarrow$ Maximum distance of any point is at most 1.

$\Rightarrow R = 1$

$(w^T x_i) y_i \geq 0.25$

$\Rightarrow$ Distance of hyperplanes $= 0.25$

$\Rightarrow \gamma = 0.25$

$\Rightarrow$ mistakes $\leq \frac{R^2}{\gamma^2}$

$\Rightarrow$ mistakes $\leq \frac{(1)^2}{(0.25)^2}$

$\Rightarrow$ mistakes $\leq \underline{\underline{16}}$

$\Rightarrow$ At most 16 mistakes are made.

Consider a neural network with 8 inputs and 2 outputs. If there are 4 hidden layers each with 4 neurons, how many parameters need to be learnt if there is a bias associated with each neuron in the hidden and output layers?

Response Type : Numeric

Evaluation Required For SA : Yes

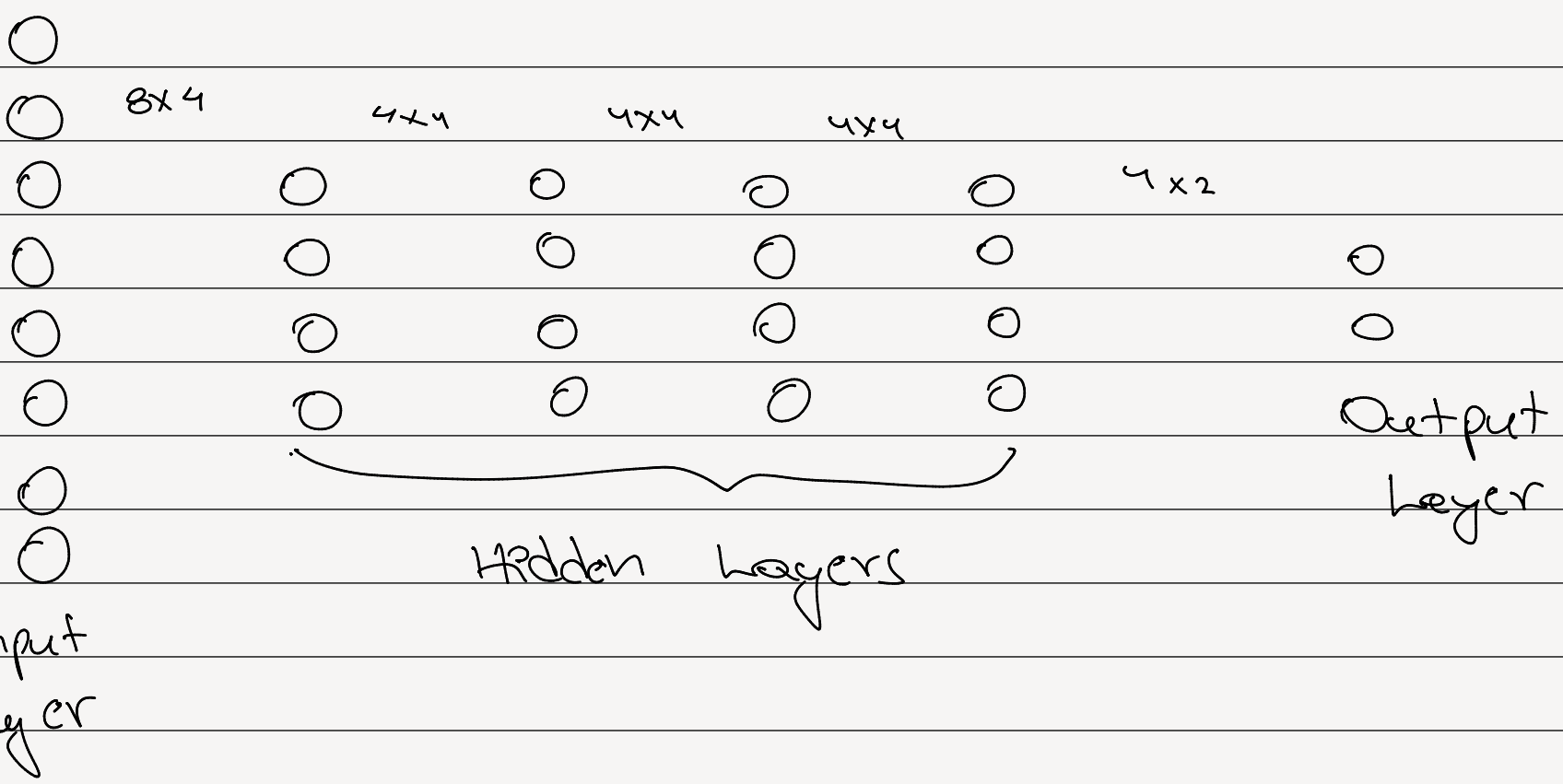
Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

106



$\Rightarrow$ Parameters $= 8\times4 + 4\times4 + 4\times4 \times 4\times4 + 4\times2$

$\Rightarrow$ 88 (without bias)

$\Rightarrow$ Parameters $= 88 + 4 + 4 + 4 + 4 + 2$

$\Rightarrow$ 106 (with bias)

We fit the following two models on a given one-dimensional dataset:

Model 1: $\hat{y}_i = w_0 + w_1 x + w_2 x^2$

Model 2: $\hat{y}_i = w_0 + w_1 x + w_2 x^2 + w_3 x^3$

The training dataset for both models is the same. The test dataset used to evaluate both models is the same.

Based on the above data, answer the given subquestions.

Sub questions

Question Number : 39 Question Id : 640653454721 Question Type : MCQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 2.5

Question Label : Multiple Choice Question

Which of the two models is more likely to fit the training data better?

Options :

6406531512251. ✖ Model 1

6406531512252. ✔ Model 2

6406531512253. ✖ Both will fit equally well

6406531512254. ✖ Can not say

Which model is more likely to give less test error?

Options :

6406531512255. ✖ Model 1

6406531512256. ✖ Model 2

6406531512257. ✔ It will depend upon the underlying distribution that generates the dataset and therefore, can not say.

6406531512258. ✖ Both will give the equal error

Upon performing standard PCA on a centered dataset in $\mathbb{R}^3$, we get the principal components to be:

$$\mathbf{w}_1 = [1 \quad 0 \quad 0]^T, \quad \mathbf{w}_2 = [0 \quad 1 \quad 0]^T, \quad \mathbf{w}_3 = [0 \quad 0 \quad 1]^T$$

$\mathbf{C}$ is the covariance matrix of the centered dataset. The off-diagonal entries are hidden from your view:

$$\mathbf{C} = \begin{bmatrix} 12 & a & b \\ a & 6 & c \\ b & c & 3 \end{bmatrix}$$

$[x_1 \quad x_2 \quad x_3]^T$ denotes a data-point. Here, x_1, x_2, x_3 are the three features.

Note: The word standard indicates that no kernel has been used.

What is the variance along the first principal component?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

12

ω_1 is eigen vector of first principal component.

$C\omega = \lambda\omega$

$$\Rightarrow \begin{bmatrix} 12 & a & b \\ a & 6 & c \\ b & c & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \lambda_1 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \quad \text{(for } \omega_1\text{)}$$

$$\Rightarrow \begin{bmatrix} 12 \\ a \\ b \end{bmatrix} = \lambda \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

$\Rightarrow \lambda_1 = 12,$
$a = 0,$
$b = 0$

The off-diagonal elements of $\mathbf{C}$ correspond to the covariance between a pair of features, (x_i, x_j) with $i \neq j$. Which of the following statements about the features of this dataset is true?

Options :

6406531512228. ✖ Each pair of features has a strong positive correlation. For instance is, if x_1 increases, then x_2 also increases.

6406531512229. ✖ Each pair of features has a strong negative correlation. For instance is, if x_1 increases, then x_2 decreases.

6406531512230. ✔ Each pair of features is uncorrelated. For instance is, if x_1 increases, we can say nothing about the trend of x_2.

6406531512231. ✖ Unless we know the exact values of the off-diagonal elements a, b and c, we can't comment about the correlation among features.

$$C\omega = \lambda\omega$$

$$\Rightarrow \begin{bmatrix} 12 & a & b \\ a & 6 & c \\ b & c & 3 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = \lambda_2 \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \quad (\text{for } \omega_2)$$

$$\Rightarrow \begin{bmatrix} a \\ 6 \\ c \end{bmatrix} = \begin{bmatrix} 0 \\ \lambda_2 \\ 0 \end{bmatrix}$$

$$\Rightarrow a = 0$$
$$6 = \lambda_2$$
$$c = 0$$

Now we know, $a = b = c = 0$

Covariance matrix can also be written as,

$$C = \begin{bmatrix} \text{cov}(x_1, x_1) & \text{cov}(x_1, x_2) & \text{cov}(x_1, x_3) \\ \text{cov}(x_2, x_1) & \text{cov}(x_2, x_2) & \text{cov}(x_2, x_3) \\ \text{cov}(x_3, x_1) & \text{cov}(x_3, x_2) & \text{cov}(x_3, x_3) \end{bmatrix} = \begin{bmatrix} 12 & a & b \\ a & 6 & c \\ b & c & 3 \end{bmatrix}$$

As $a = b = c = 0$,

$\Rightarrow$ Covariance of off-diagonal elements $= 0$

$\Rightarrow$ If covariance $= 0$,

$\Rightarrow$ Features are independent.

A binary classification (labels are 0 and 1) dataset contains n examples belonging to $\{0, 1\}^4$ such that the first feature values for all n examples are 0. Assume that no smoothing is done.

What will be the value of $\hat{p}_1^1$? $\hat{p}_j^y$ is the estimate for the probability that the j^{th} feature value of an example is 1 given that the example belongs to the label y.

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

0

$$\hat{p}_1' = \frac{\sum_{i=1}^{n} 1(f_1' = 1 \mid y_1 = 1)}{\sum_{i=1}^{n} 1(y_1 = 1)}$$

Given that feature 1 is zero in all of the 'n' examples.

$\Rightarrow \hat{p}_1' = 0$

What will be the prediction for the point $x = [1, 1, 1, 0]$ using the naive Bayes classifier?

Options :

6406531512260. ✖ 0

6406531512261. ✖ 1

6406531512262. ✖ Indeterminate as $P(y = 0|x) = P(y = 1|x) = 1$

6406531512263. ✔ Indeterminate as $P(y = 0|x) = P(y = 1|x) = 0$

$$P(y=0/x) = \frac{P(x|y=0) \cdot P(y=0)}{P(x)}$$

$$\Rightarrow \frac{P([1,1,1,0]\,|\,y=0) \cdot P(y=0)}{P([1,1,1,0])}$$

Here the first feature (f_1) is 1.
We know that in our 'n' examples, first feature is always zero.

$\Rightarrow P(y=0/x) = 0$

Similarly, $P(y=1/x) = 0$

Consider a data set $x_1 = [1, 1]$, $x_2 = [1, -1]$, $x_3 = [-1, -1]$, $x_4 = [-1, 1]$ and the corresponding class labels being $y_1 = -1$, $y_2 = +1$, $y_3 = +1$, $y_4 = -1$.

Assume you try to find the w (no bias) 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?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

2

Perceptron Algo

Pick (x_i, y_i) pair from the dataset
If $\text{sign}(w^T x_i) = y_i$

Do nothing

else

$$\boxed{w^{t+1} = w^t + x_i y_i}$$

end

$w^0 = [0, 0]$

Prediction ($\hat{y}$)	Actual (y)	
$\text{sign}(w^T x_1) = 1$	-1	← mistake (1)

$\Rightarrow w' = w^0 + \begin{bmatrix} 1 \\ 1 \end{bmatrix}(-1)$

$\Rightarrow \omega^1 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$

Prediction ($\hat{y}$)	Actual (y)	
$\text{sign}(\omega^T x_1) = -1$	-1	
$\text{sign}(\omega^T x_2) = 1$	1	
$\text{sign}(\omega^T x_3) = 1$	1	
$\text{sign}(\omega^T x_4) = 1$	-1	← mistake (2)

$\Rightarrow \omega^2 = \begin{bmatrix} -1 \\ -1 \end{bmatrix} + \begin{bmatrix} -1 \\ 1 \end{bmatrix}(-1)$

$\Rightarrow \omega^2 = \begin{bmatrix} 0 \\ -2 \end{bmatrix}$

Prediction ($\hat{y}$)	Actual (y)
$\text{sign}(\omega^T x_1) = -1$	-1
$\text{sign}(\omega^T x_2) = 1$	1
$\text{sign}(\omega^T x_3) = 1$	1
$\text{sign}(\omega^T x_4) = -1$	-1

no mistakes are made, algo ends

$\Rightarrow$ Total mistakes $= \underline{\underline{2}}$

What is the squared length of the weight vector corresponding to the final linear separator your algorithm outputs?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

4

$\|\omega\|^2 = (0)^2 + (-2)^2$

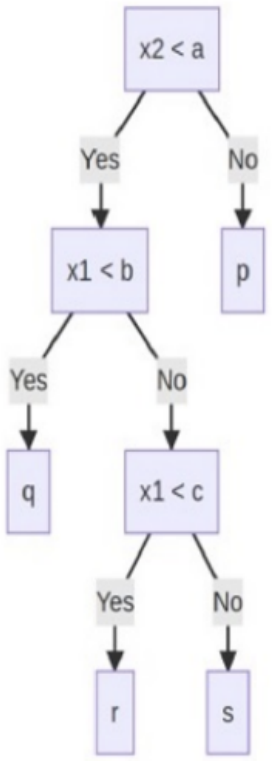
$\Rightarrow \underline{\underline{4}}$

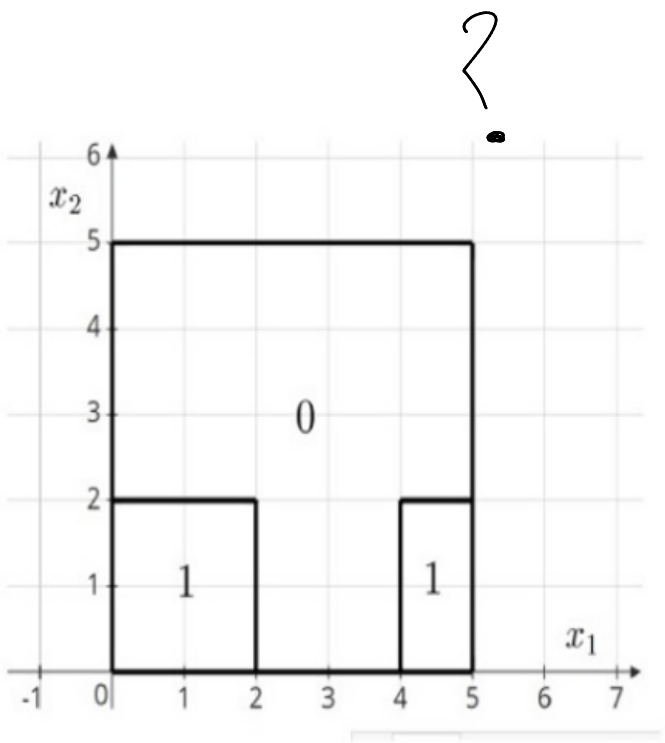
Consider a binary classification problem in $\mathbb{R}^2$. The features for this problem lie in a square:

$$0 \leq x_1 \leq 5$$
$$0 \leq x_2 \leq 5$$

A decision tree is trained on some training dataset for this problem. The tree and the decision regions corresponding to it are given below:

`x2 is the same as` x_2
`x1 is the same as` x_1





- a, b, c are values associated with the question nodes and are positive integers.
- p, q, r, s are values associated with the leaves and belong to the set of labels, in this case $\{0, 1\}$.

There are three bounded decision regions, denoted by solid lines, two of which have label 1 and one which has label 0.

What is the value of a?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

2

Question Label : Short Answer Question

What is the value of b?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

2

What is the value of c?

Response Type : Numeric

[redacted]

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

4

What is the value of the following expression?

$$(1 + p + r)(q + s)$$

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

2

Consider a single iteration of the AdaBoost algorithm that was run on three sample points, starting with uniform weights on the sample points. The labels are either +1 or −1. In the table below, some values have been omitted.

Data point	True label	Predicted label	Initial weight	Updated weight
x_1	?	1	$\frac{1}{3}$	$\frac{1}{2}$
x_2	−1	−1	$\frac{1}{3}$	?
x_3	−1	?	$\frac{1}{3}$	$\frac{1}{4}$

What will be the true label for point x_1 ? Enter 1 or -1 .

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

-1

For x_1, initial weight was 0.33 then it was updated to 0.5

In AdaBoost if datapoint is predicted incorrectly, then weight is increased.

$\Rightarrow$ True label of $x_1 = -1$

What will be the updated weight for point x_2 ? Enter your answer correct to two decimal places.

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Range

Text Areas : PlainText

Possible Answers :

0.23 to 0.27

All the weights should always sum up to 1.

$\Rightarrow \frac{1}{2} + \text{weight}(x_2) + \frac{1}{4} = 1$

$\Rightarrow \text{weight}(x_2) = \underline{0.25}$

How much training error will be incurred by the first estimator? The training examples consist of given three points. Enter your answer correct to two decimal places.

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Range

Text Areas : PlainText

Possible Answers :

0.31 to 0.35

First Estimator

	True Label ($h(x_i)$)	Predicted Label (y_i)
x_1	−1	1
x_2	−1	−1
x_3	−1	−1

$\text{Error} = \frac{1}{3} = \underline{0.33}$

There are 8 points in a training dataset in $\mathbb{R}^2$ for a binary classification problem that is linearly separable. Use the following notation: $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ for a data-point and $\mathbf{w}^* = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$ for the optimal weight vector of a hard-margin, linear-SVM.

	x_1	x_2	y
a	0	1	1
b	0	2	1
c	1	1	1
d	2	0	1
e	0	−1	−1
f	−2	0	−1
g	−3	0	−1
h	−5	1	−1

A hard-margin, linear-SVM is trained on this dataset. The optimal weight vector is $\begin{bmatrix} 0.5 \\ 1 \end{bmatrix}$. What is the maximum number of support vectors for this setup?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

4

A datapoint is considered to be on supporting vectors when $(\omega^T x_i) y_i = 1$

Check all the datapoints for, $(\omega^T x_i) y_i = 1$

The datapoints which satisfy this condition are,

a, d, e, f → 4 datapoints

The point $\begin{bmatrix}10\\0\end{bmatrix}$ with label 1 is added to the existing training dataset. We will now refer to these 9 points as the new dataset.

Options :

6406531512237. ✖ The new dataset is **not** linearly separable.

6406531512238. ✔ The new dataset is linearly separable.

6406531512239. ✔ If a hard-margin, linear-SVM is trained on the new dataset, the optimal weight vector will be $\begin{bmatrix}0.5\\1\end{bmatrix}$

6406531512240. ✖ If a hard-margin, linear-SVM is trained on this new dataset with 9 points, the optimal weight vector will **not** be $\begin{bmatrix}0.5\\1\end{bmatrix}$

If $(w^T x_i) y_i \geq 1$ for all datapoints, then dataset is linearly separable.

Here all datapoints satisfy $(w^T x_i) y_i \geq 1$,
$\Rightarrow$ Dataset is linearly separable

MLT End-Term Practice 1

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

?

Ans: 1

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 $\mathbf{w}$?

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

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

Gradient of Linear regression in matrix form is given by,

$$2(XX^T)w - 2(Xy) = 0$$

A hard-margin, linear-SVM is trained for a 2D problem. The optimal weight vector is $\mathbf{w} = \begin{bmatrix} 2 & -1 \end{bmatrix}^T$. 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?

No, the answer is incorrect.

Score: 0

Accepted Answers:

(Type: Range) 0.74,0.76

Here 2 points are on decision boundry ($w^T x = 0$) and 1 point is on right side of decision boundry.

$\Rightarrow$ 3 points are predicted as positive.
$\Rightarrow$ 1 point is predicted as negative.

$\Rightarrow P(\text{Positive point}) = \frac{3}{4} = \underline{\underline{0.75}}$

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 accuracy of this model as n becomes very large? Your answer should be between 0 and 1.

No, the answer is incorrect.

Score: 0

Accepted Answers:

(Type: Range) 0.09,0.11

A kernel k is defined as

$$k : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$$
$$k([x_1, x_2]^T, [y_1, y_2]^T) = 1 + 2x_1^2 y_1^2 + 2x_2^2 y_2^2$$

Which of the following transformation mappings corresponds to this kernel function?

Options :

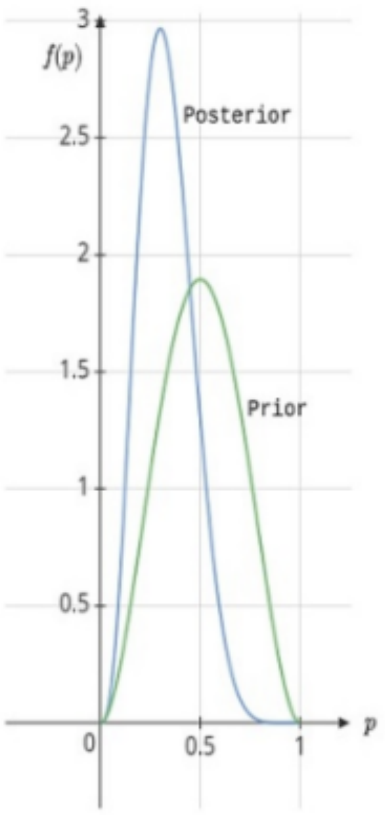
6406531884708. ✖ $\phi([x_1, x_2]^T) = [1,\ x_1^2 + x_2^2]^T$

6406531884709. ✖ $\phi([x_1, x_2]^T) = [1,\ \sqrt{2}x_1^2 + \sqrt{2}x_2^2]^T$

6406531884710. ✔ $\phi([x_1, x_2]^T) = [1,\ \sqrt{2}x_1^2,\ \sqrt{2}x_2^2]^T$

6406531884711. ✖ $\phi([x_1, x_2]^T) = [1,\ x_1^2,\ x_2^2]^T$

Consider a Bayesian estimation problem for a dataset of six observations, where each observation is either one or zero. The prior (green) and posterior (blue) distributions are shown below. Recall that both are Beta distributions in this case. p denotes the parameter and $f(p)$ denotes its pdf. Also recall that the observations are sampled from a Bernoulli distribution with parameter p



What can you say about the number of ones in the dataset? Choose the most appropriate option.

Options :

6406531884712. ✔ 1

6406531884713. ✖ 3

6406531884714. ✖ 5

Posterior mean here is less than 0.5 (From visual inspection)

Only option 1) gives posterior mean less than 0.5.

In regularized linear regression, what is a common issue that can occur when the regularization parameter λ is set too low?

Options :

6406531884715. ✔ Overfitting

6406531884716. ✖ Underfitting

Consider the following training dataset for a binary classification problem in $\mathbb{R}^2$:

x_1	x_2	y
1	2	1
1	−2	1
−5	0	1
5	0	−1
−2	1	−1
−2	−1	−1

If we try to learn a perceptron model for this dataset, will the algorithm ever converge to a weight vector? Select the most appropriate answer with the information available to you.

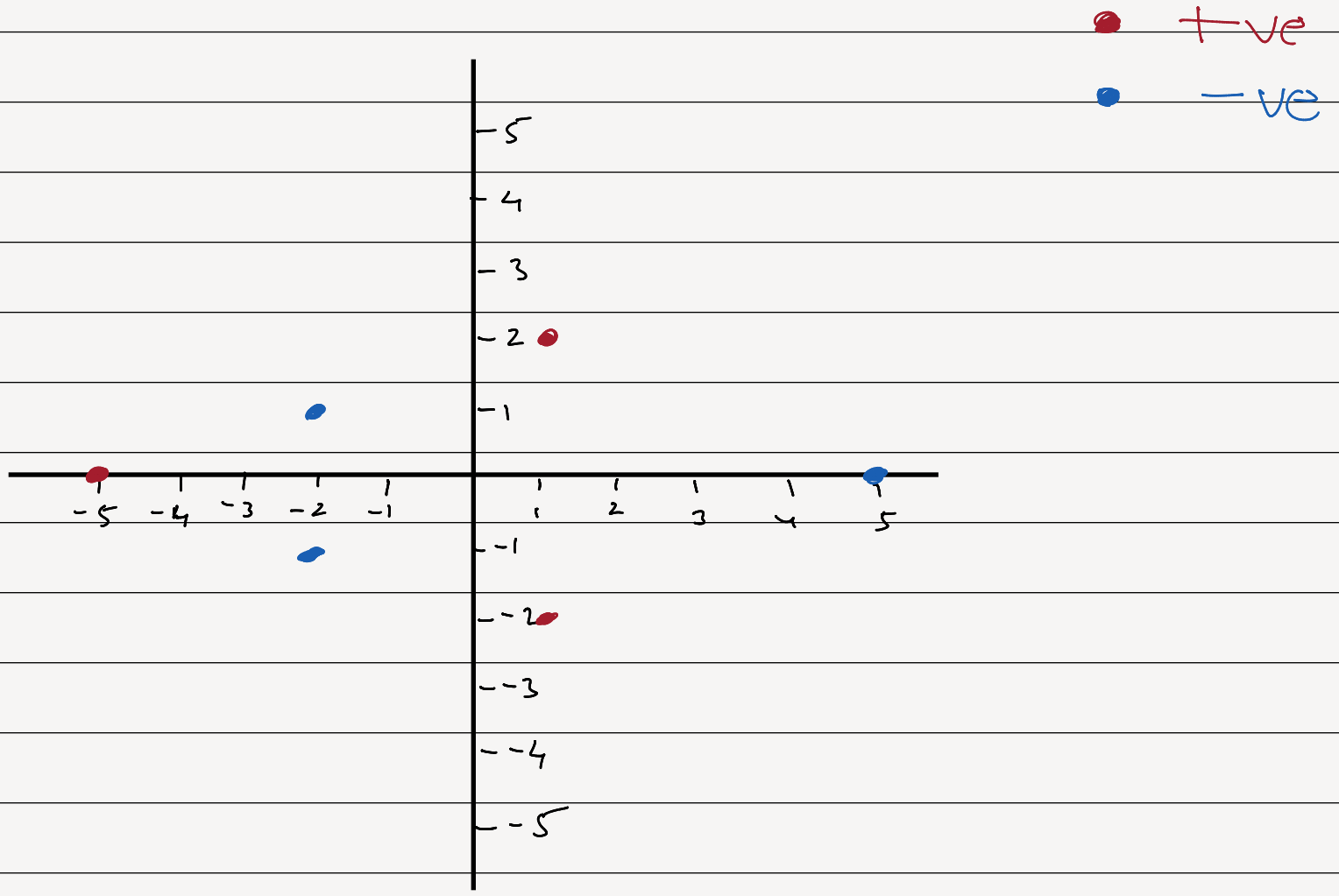
Options :

6406531884717. ✖ Yes, it will certainly converge to a weight vector.

6406531884718. ✔ No, it will never converge.

From below graph, we can see that this dataset isn't linearly separable.

⇒ Perceptron Algorithm will never converge.



Consider a logistic regression model that has been trained for a binary classification problem on a dataset in $\mathbb{R}^2$. The weight vector is $\begin{bmatrix} 3/5 \\ 4/5 \end{bmatrix}$. Given a test data-point as input to the model, it returns 1 as the predicted label if the probability output by the model is greater than 0.75 and 0 otherwise. What is the predicted label for the test data-point $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$? Note that the probability output by a logistic regression model is $P(y = 1 \mid \mathbf{x})$.

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

0

$$P(y=1|x) = \frac{1}{1+e^{-w^T x}}$$

Here,
$w = \begin{bmatrix} 3/5 \\ 4/5 \end{bmatrix}$ and $x = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$

$$\Rightarrow w^T x = \begin{bmatrix} 3/5 & 4/5 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \frac{3}{5} = 0.6$$

$$\Rightarrow P(y=1|x) = \frac{1}{1+e^{-0.6}}$$

$$\Rightarrow \frac{1}{1.5488}$$

$$\simeq 0.645$$

As $p < 0.75$,
$\Rightarrow$ Predicted label is $\underline{0}$.

While training a perceptron model, the weight vector at some iteration t is $\mathbf{w}^t$. The next data-point picked up by the perceptron algorithm in the course of its execution is $(\mathbf{x}, y)$, where y is the true label:

$$\mathbf{w}^t = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} -1 \\ 0 \\ 1 \\ -1 \end{bmatrix}, \quad y = 1$$

What is the value of $\mathbf{w}^{t+1}$?

$w^{t+1} = w^t + \eta y_i x_i$

$\Rightarrow \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix} + \begin{bmatrix} -1 \\ 0 \\ 1 \\ -1 \end{bmatrix} (1)$

$\Rightarrow \begin{bmatrix} 0 \\ 2 \\ 4 \\ 3 \end{bmatrix}$

Options :

6406531884719. ✔ $\begin{bmatrix} 0 \\ 2 \\ 4 \\ 3 \end{bmatrix}$

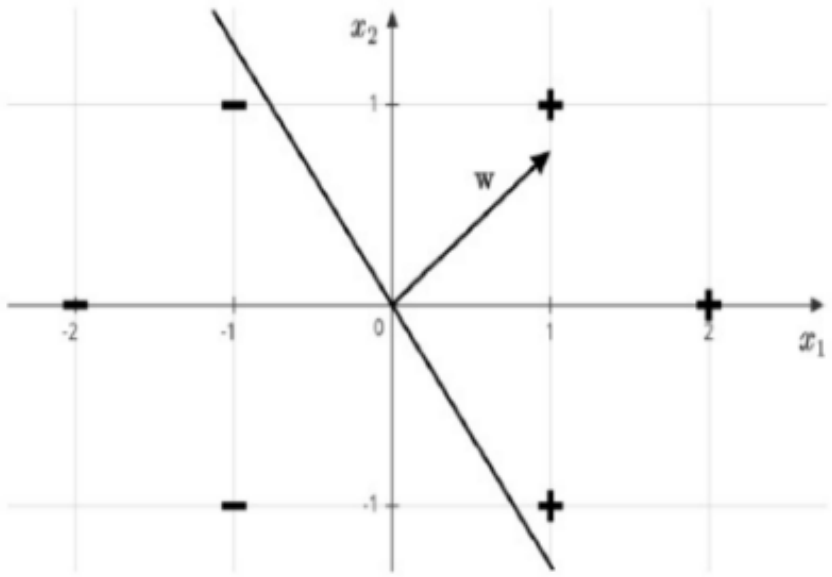
6406531884720. ✖ $\begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}$

6406531884721. ✖ $\begin{bmatrix} 2 \\ 2 \\ 2 \\ 5 \end{bmatrix}$

6406531884722. ✖ No update will happen as the point $(\mathbf{x}, y)$ is not a mistake with respect to $\mathbf{w}^t$

Question Label : Multiple Choice Question

Consider a linearly separable binary classification problem. The training dataset is shown below. Is **w** the optimal weight vector corresponding to a hard-margin, linear-SVM? The symbol + corresponds to label 1 and − corresponds to label −1.

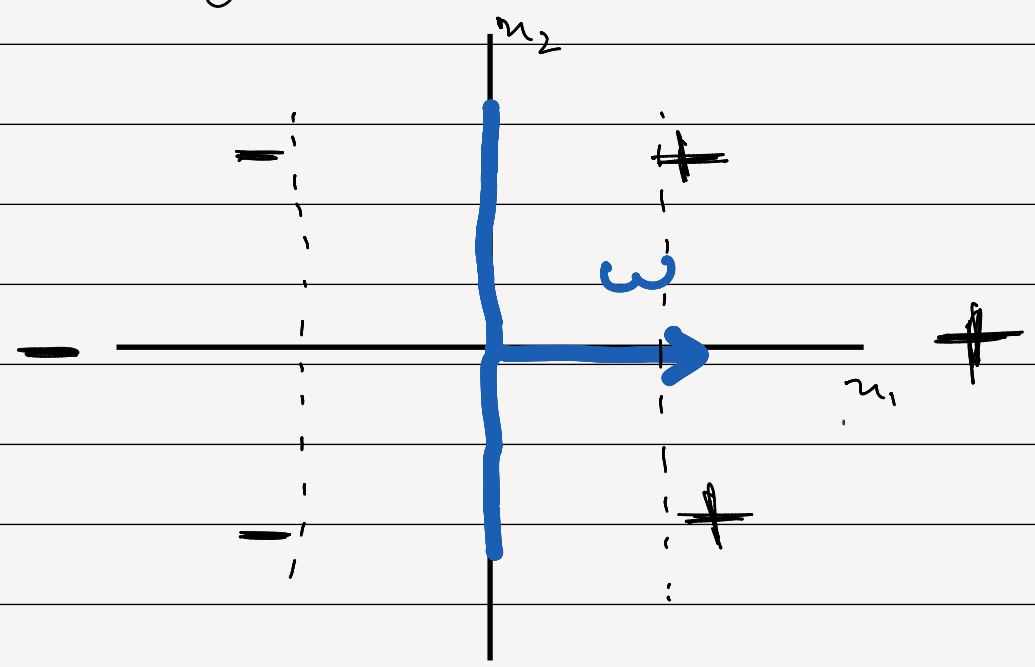


Options :

6406531884723. ✖ **w** shown in this diagram is the optimal weight vector for a hard-margin, linear-SVM

6406531884724. ✔ **w** shown in this diagram is **not** the optimal weight vector for a hard-margin, linear-SVM

w is not optimal because distance of hyperplanes/margins is very small.



This could be optimal w.

Consider a linearly separable binary classification problem in which the data-points are in $\mathbb{R}^{10}$. The training dataset has 50 data-points, 20 from the positive class and 30 from the negative class. Now, consider a hard-margin linear-SVM. How many constraints does the primal problem have? Select the most appropriate answer.

Options :

6406531884725. ✔ 50

6406531884726. ✖ 20

6406531884727. ✖ 30

6406531884728. ✖ 10

Primal Problem

$$\min_{w} \quad \frac{1}{2}||w||^2 \qquad (2)$$

$$\text{such that}$$

$$1 - (w^T x_i) y_i \leq 0 \quad \forall i$$

We are applying constraints for all datapoints.
⇒ Constraints in primal problem = 50

A clustering of 200 data points in $\mathbb{R}^2$ was done using Lloyd's algorithm with $K = 3$. You are told that the points $[1,1]^T$, $[4,1]^T$, and $[3,3]^T$ are in the same cluster. Which of the following points will definitely lie in the same cluster? (If there are any ties with this cluster, they should be assigned exclusively to this cluster.)

Options :

6406531884729. ✖ $[4,2]^T$

6406531884730. ✔ $[3,2]^T$

6406531884731. ✖ $[0,0]^T$

6406531884732. ✔ $[2,2]^T$

Consider a linear regression model that was trained on dataset X of shape (d, n). Which of the following techniques could potentially decrease the loss on the training data (assuming the loss is the squared error)?

Options :

6406531884733. ✔ Adding a dummy feature in the dataset and learning the intercept w_0 as well. (Fact)

6406531884734. ✖ Penalizing the model weights with L2 regularization.

6406531884735. ✖ Penalizing the model weights with L1 regularization.

6406531884736. ✔ Training the kernel regression model of degree 2. (Fact)

There are 100 data-points at some node in a decision tree. If the entropy of this node is 0.722, which of the following could be the number of data-points that belong to the positive class in this node? Choose all appropriate answers. Use $\log_2$ as always.

Options :

6406531884737. ✔ 80

6406531884738. ✔ 20

6406531884739. ✖ 70

6406531884740. ✖ 30

6406531884741. ✖ 50 → p = 0.5, at p = 0.5, entropy = 1

6406531884742. ✖ 100 → p = 1, at p = 1, entropy = 0

entropy for $p = 0.8$

$\Rightarrow -(0.8 \log_2(0.8) + 0.2 \log_2(0.2))$

$\Rightarrow -(-0.2575 + -0.4643)$

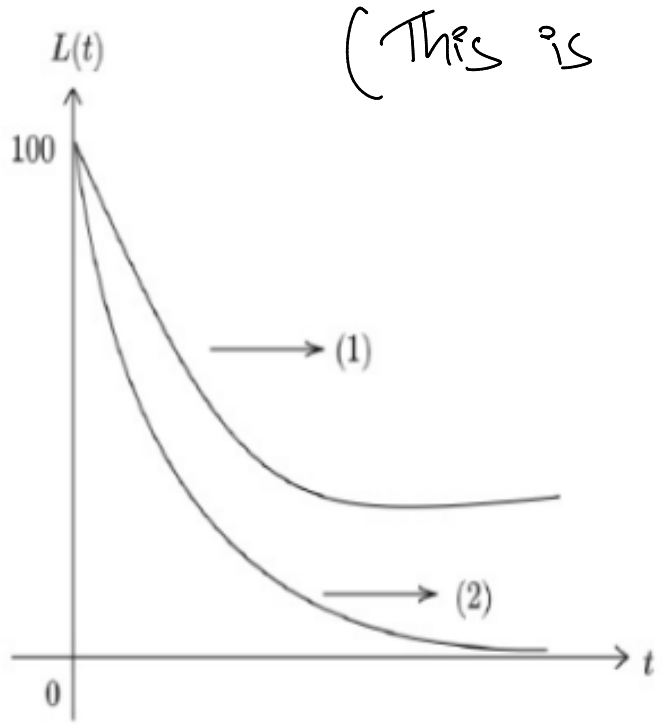
$\Rightarrow 0.7218 \simeq 0.722$

Note that entropy at p and (1−p) is always same as entropy is symmetrical function.

=> Entropy at $p = 0.2$ is also 0.7219

Consider a supervised ML problem that has a training dataset and a validation dataset. A ML model is being trained on the training dataset using gradient descent and its performance is monitored on the validation dataset. The loss function $L(t)$ at time step t is plotted against t.

(This is AdaBoost Loss Function)



One of these curves corresponds to the model's loss on the training dataset and the other corresponds to the model's loss on the validation dataset. Identify the two curves. Consider a general scenario and not an extreme instance while answering this problem. Exactly two options are correct.

Options :

6406531884751. ✔ (1) is the loss on the validation dataset ✓

6406531884752. ✔ (2) is the loss on the training dataset

6406531884753. ✖ (1) is the loss on the training dataset

6406531884754. ✖ (2) is the loss on the validation dataset

Which of the following statements are true about 'C' in SVM?

Options :

6406531884743. ✖ As C approaches 0, the soft margin SVM is equal to the hard margin SVM.

6406531884744. ✔ As C approaches ∞, the soft margin SVM is equal to the hard margin SVM.

6406531884745. ✔ A smaller value of C tends to create a larger margin.

6406531884746. ✖ C can be negative, as long as the bribe(ξ) each data point pays is nonnegative.

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/are true? (d is the total number of features)

Options :

6406531884747. ✖ Increasing p reduces the correlation between any two trees in the forest.

6406531884748. ✔ Decreasing p reduces the correlation between any two trees in the forest. ✓

6406531884749. ✔ Increasing p increases the performance of individual trees in the forest. ✓

6406531884750.

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

p is the probability of sampling features.

Consider the following dataset of 4 points in $\mathbb{R}^5$:

$$\{-2\mathbf{u},\ -\mathbf{u},\ \mathbf{u},\ 2\mathbf{u}\}$$

where $\mathbf{u} = [0.5 \quad 0 \quad 0.5 \quad -0.5 \quad -0.5]^T$. What is the variance along the first principal component after performing standard PCA on this dataset?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

2.5

Here,

$X = [-2u \quad -u \quad u \quad 2u]$

We know,

$$C = \frac{1}{n} XX^T$$

$$\Rightarrow C = \frac{1}{4} 10\, u u^T$$

$$\Rightarrow C = 2.5\, u u^T$$

$$= 2.5\begin{bmatrix} 0.25 & 0 & 0.25 & -0.25 & -0.25 \\ 0 & 0 & 0 & 0 & 0 \\ 0.25 & 0 & 0.25 & -0.25 & -0.25 \\ -0.25 & 0 & -0.25 & 0.25 & 0.25 \\ -0.25 & 0 & -0.25 & 0.25 & 0.25 \end{bmatrix}$$

$$2.5\begin{bmatrix} 0.25 & 0 & 0.25 & -0.25 & -0.25 \\ 0 & 0 & 0 & 0 & 0 \\ 0.25 & 0 & 0.25 & -0.25 & -0.25 \\ -0.25 & 0 & -0.25 & 0.25 & 0.25 \\ -0.25 & 0 & -0.25 & 0.25 & 0.25 \end{bmatrix}\begin{bmatrix} 1 \\ 0 \\ 1 \\ -1 \\ -1 \end{bmatrix} = 2.5\begin{bmatrix} 1 \\ 0 \\ 1 \\ -1 \\ -1 \end{bmatrix}$$

This has only one principle component because the rank of the C is 1, and the eigenvalue is 2.5.

Consider a binary classification problem for a linearly separable dataset in $\mathbb{R}^4$. The optimal weight vector for a hard-margin linear-SVM classifier is given to be $\mathbf{w}^*$. The data-point $(\mathbf{x}, y)$ belongs to the training dataset:

$$\mathbf{w}^* = \begin{bmatrix} 1 \\ 0 \\ -1 \\ 2 \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} 1 \\ -5 \\ -3 \\ 0 \end{bmatrix}, \quad y = 1$$

What is the value of α^* corresponding to this data-point? If you think the answer cannot be determined with this information, enter -1. If you think it can be determined, enter the correct value of α^*. Note that α^* is the Lagrange multiplier corresponding to this data-point.

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

0

$$[1 \quad 0 \quad -1 \quad 2]\begin{bmatrix} 1 \\ -5 \\ -3 \\ 0 \end{bmatrix} = 1+0+3 \Rightarrow 4$$

$$\Rightarrow w^T x_i > 1$$

$\Rightarrow$ Point does not lie on supporting

hyperplane

=> α_i^* : depicts importance of points

If point does not lie on hyperplane, then it does not contribute to w^*.

=> Point is not important.

=> $\alpha_i^* = 0$

Consider the following architecture for a neural network:

Layer	Neurons
Input	10
Hidden Layer-1	20
Hidden layer-2	30
Output layer	1

How many weights does this network have? Assume that there is no bias associated with any neuron.

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

830

weights/parameters = $10 \times 20 + 20 \times 30 + 30 \times 1$

=> $200 + 600 + 30$

=> 830

If a data-point $(\mathbf{x}, y)$ is correctly classified by a logistic regression model with weight vector $\mathbf{w}$, what could be the maximum value of the logistic loss associated with it?

Assume that the model predicts the label 1 if $P(y = 1 \mid \mathbf{x}) \geq 0.5$ and -1 otherwise. Note that $y \in \{-1, 1\}$ as we are looking at the logistic loss. Use $\log_e$, the natural logarithm, while calculating the logistic loss. Enter your answer correct to three decimal places.

Hint: Geometry leads, algebra follows

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Range

Text Areas : PlainText

Possible Answers :

0.68 to 0.70

Consider the binary classification problem with 5 binary features. The training dataset contains the following five points belonging to $\{0,1\}^5$:

$$[1,0,1,0,1]^T, [1,1,1,0,1]^T, [0,0,1,1,0]^T, [0,1,0,1,0]^T, [1,1,0,0,0]^T$$

The labels of the points are $0, 1, 1, 1$ and 0, respectively. Assume that the naive condition holds true. Do not apply any smoothing on the dataset.

How many parameters need to be estimated to make a prediction for a data point using a naive Bayes algorithm?

Response Type : Numeric

Evaluation Required For SA : Yes

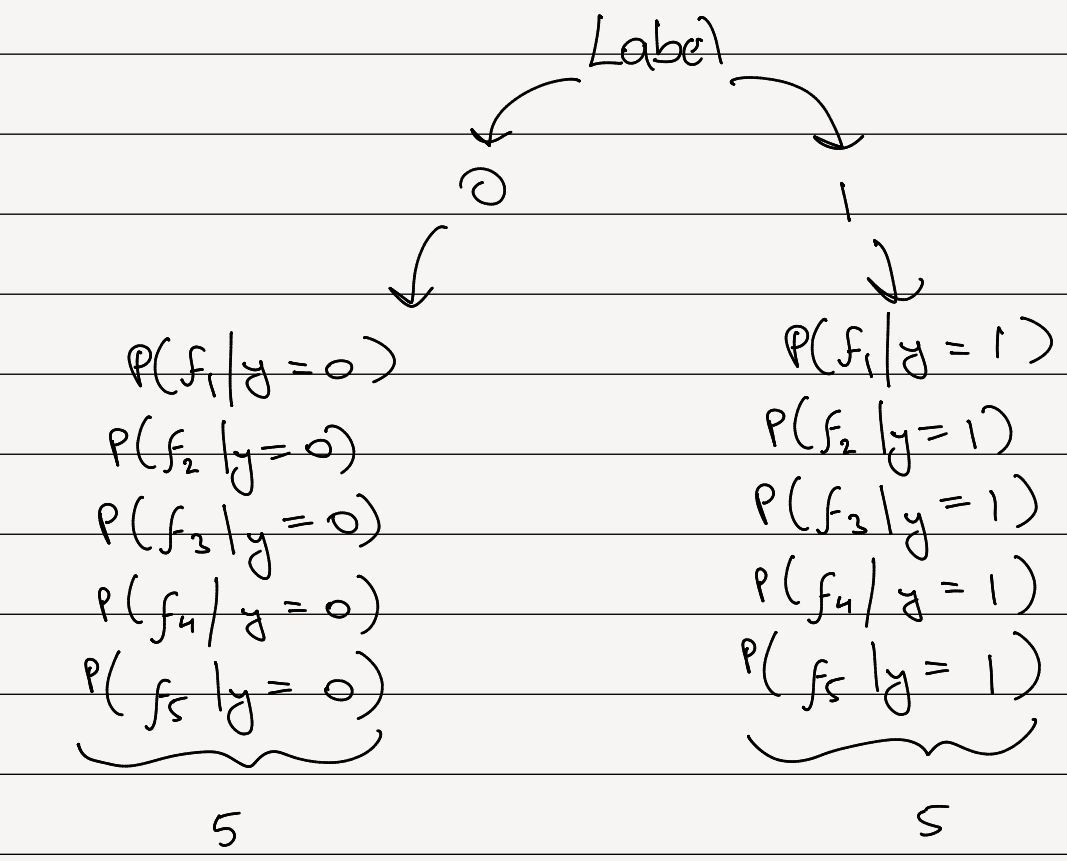
Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

11



$\Rightarrow$ Parameters $= 5 + 5 +$ $\underbrace{1}_{P(y=1)}$

$\Rightarrow 11$

With what probability does the first feature of a point take the value 0 given that the point is labeled 1? Enter your answer correct to two decimal places.

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Range

Text Areas : PlainText

Possible Answers :

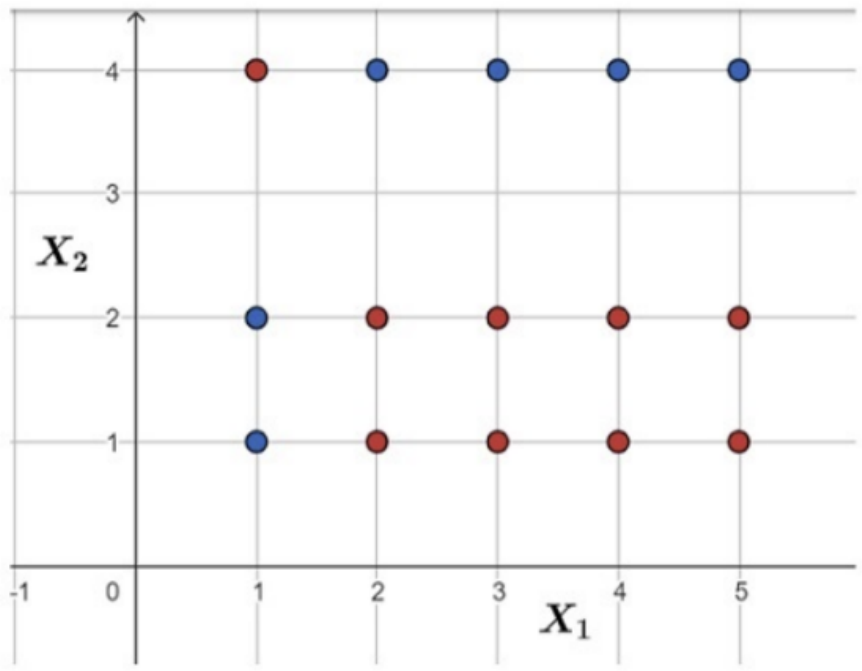
0.64 to 0.68

$P(f_1 = 0 \mid y = 1)$

$$\Rightarrow \frac{\sum_{i=1}^{n} 1(f_1 = 0 \mid y_i = 1)}{\sum_{i=1}^{n} 1(y_i = 1)}$$

$\Rightarrow \frac{2}{3} = 0.66$

Consider the following two-dimensional dataset with two classes: $+1$ for blue points and -1 for red points. An AdaBoost algorithm was run on this dataset using decision stumps as weak learners.



When training the new weak learner $h_t(x)$ (decision stump at t^{th} iteration), we choose the split that minimizes the weighted misclassification error with respect to current weights D_t i.e. choose h_t that minimizes $\sum_{i=1}^{n} D_t(i)\mathbb{1}(h_t(x_i) \neq y_i)$.

What will be the misclassification error incurred by the first decision stump?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

0.2