AN2 — QPD2 — 2 April

Consider that the three weight vectors $\mathbf{w}_1, \mathbf{w}_2$, and $\mathbf{w}_3$ are learned for a six-dimensional dataset using a linear regression model or regularized linear regression model (Not in any particular order).

$\mathbf{w}_1 = [0.5, 0, 0.25, 0, 0, -0.14]$

$\mathbf{w}_2 = [0.8, -0.23, 0.45, 0.2, 0.31, -0.54]$

$\mathbf{w}_3 = [0.24, -0.03, 0.1, 0.02, 0.09, -0.14]$

Select the most appropriate match for these weight vectors.

Options :

6406531737028. ✖ $\mathbf{w}_1 \rightarrow$ Linear regression, $\mathbf{w}_2 \rightarrow$ Ridge regression, $\mathbf{w}_3 \rightarrow$ Lasso

6406531737029. ✖ $\mathbf{w}_1 \rightarrow$ Ridge regression, $\mathbf{w}_2 \rightarrow$ Linear regression, $\mathbf{w}_3 \rightarrow$ Lasso

6406531737030. ✖ $\mathbf{w}_1 \rightarrow$ Lasso, $\mathbf{w}_2 \rightarrow$ Ridge regression, $\mathbf{w}_3 \rightarrow$ Linear regression

6406531737031. ✔ $\mathbf{w}_1 \rightarrow$ Lasso, $\mathbf{w}_2 \rightarrow$ Linear regression, $\mathbf{w}_3 \rightarrow$ Ridge regression

ω_1 has most of its features to be zero.
$\Rightarrow$ It is likely solution of lasso regression.

ω_3 has the least values of each feature in the weight vector.
$\Rightarrow$ It is likely solution of ridge regression.

Consider a binary classification dataset (classes are 0 and 1) with two binary features $f_1, f_2 \in \{0, 1\}$. A Naive Bayes classifier is learned and the estimated parameters are given as:

$P(f_1 = 1|y = 0) = 0.2$

$P(f_2 = 1|y = 0) = 0.5$

$P(f_1 = 1|y = 1) = 0.6$

$P(f_2 = 1|y = 1) = 0.4$

If a data point $[1, 0]$ is predicted in class 0 by this classifier, what will be the possible values for the estimate of $P(y = 1)$? Assume that tie-breaking goes to class zero. Values in the options are correct to two decimal places.

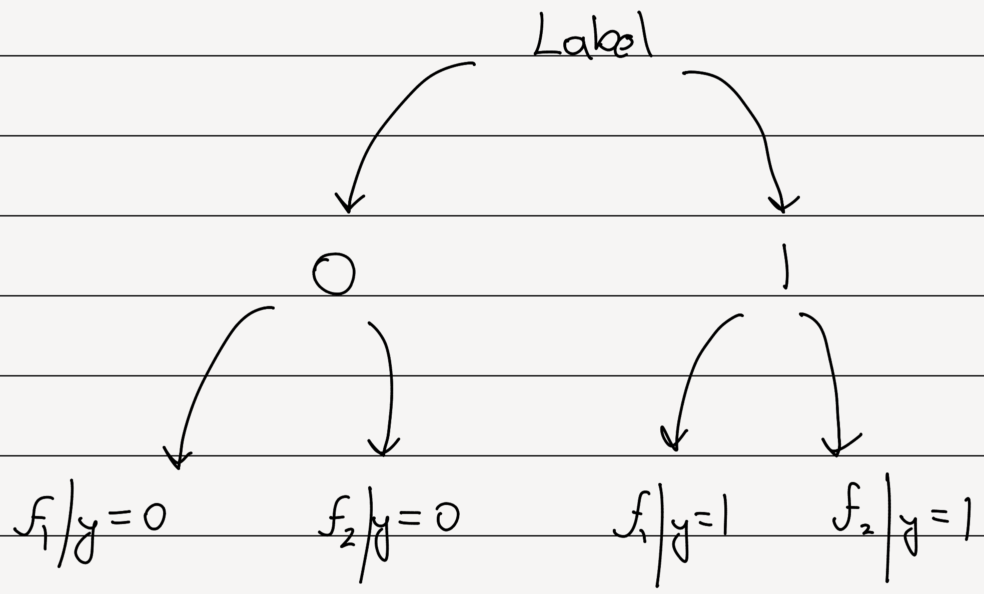
Options :

6406531737032. ✔ $(0, 0.22]$

6406531737033. ✖ $[0.22, 1)$

6406531737034. ✖ $(0, 0.29]$

6406531737035. ✖ $[0.29, 1)$



A.T.Q.

$$P(y=0 \mid [1,0]) \geq P(y=1 \mid [1,0])$$

$$\Rightarrow P(f_1=1 \mid y=0) \cdot P(f_2=0 \mid y=0) \cdot P(y) \geq P(f_1=1 \mid y=1) \cdot P(f_2=0 \mid y=1) \cdot P(y=1)$$

$$\Rightarrow (0.2)(1-0.5) \cdot P(y=0) \geq 0.6(1-0.4) \cdot P(y=1)$$

$$\Rightarrow 0.1 \cdot P(y=0) \geq 0.36 \cdot P(y=1)$$

$$\Rightarrow \frac{P(y=0)}{P(y=1)} \geq 3.6$$

$$\Rightarrow \frac{1-P(y=1)}{P(y=1)} \geq 3.6 \qquad [P(y=0)+P(y=1)=1]$$

$$\Rightarrow \frac{1}{P(y=1)} \geq 3.6+1$$

$$\Rightarrow P(y=1) \leq \frac{1}{4.6}$$

$$\Rightarrow P(y=1) \leq 0.2173$$

$$\Rightarrow P(y=1) \in [0, 0.2173]$$

Is the following statement true or false:

If $p_i^y = 0$ for $y = 0$, then $p_i^y = 1$ for $y = 1$. Here, p_j^y denotes the estimate of the probability that j^{th} feature value is 1 given that label is y ($P(f_j = 1|y)$).

Options :

6406531737036. ✖ TRUE

6406531737037. ✔ FALSE

The model is given by:

$$P(\mathbf{x} = [f_1, f_2, \dots, f_d]|y) = \prod_{i=1}^{d} (p_i^{y_i})^{f_i} (1 - p_i^{y_i})^{1-f_i}$$

The parameters to be estimated are p, $\{p_1^0, p_2^0, \dots, p_d^0\}$, and $\{p_1^1, p_2^1, \dots, p_d^1\}$. Using Maximum Likelihood Estimation, we obtain the following estimates:

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

$$\underbrace{\hat{p}_j^y = \frac{\sum_{i=1}^{n} \mathbb{1}(f_j^i = 1, y_i = y)}{\sum_{i=1}^{n} \mathbb{1}(y_i = y)}} \quad \text{for all } j \in \{1, 2, \dots, d\}, \text{ and } \forall y \in \{0, 1\}$$

This is the estimation formula for $\hat{p}_j^y$.

A linear regression model trained on a dataset $X \in \mathbb{R}^{d \times n}$ achieves zero training error for any label vector y. Which of the following options will necessarily hold true? Here I denotes an identity matrix of an appropriate size.

Options :

6406531737038. ✖ $XX^T = I$

6406531737039. ✔ $X^T(XX^T)^{-1}X = I$

6406531737040. ✖ $(XX^T)^{-1}Xy$ is a vector of all ones

6406531737041. ✖ $(XX^T)^{-1}Xy$ is a vector of all zeros

Error of linear regression model is given by,

$$\Rightarrow \sum_{i=1}^{n} \left[h(x_i) - y_i\right]^2$$

$$\Rightarrow \sum_{i=1}^{n} \left[w^T x_i - y_i\right]^2$$

$$\Rightarrow \|X^T w - y\|^2 \quad \text{—— } ①$$

We know that w can be written as,

$$w = [XX^T]^{-1} Xy \quad \text{——} \quad (2)$$

Zero training error means, error is always zero for any 'y'.

$$\Rightarrow \quad ||X^T w - y||^2 = 0 \qquad [\text{From } (1)]$$

$$\Rightarrow \quad X^T w - y = 0$$

$$\Rightarrow \quad X^T[(XX^T)^{-1} Xy] - y = 0 \qquad [\text{From } (2)]$$

$$\Rightarrow \quad X^T[(XX^T)^{-1} Xy] = y$$

Dividing both sides by y

$$\Rightarrow \quad X^T (XX^T)^{-1} X = I$$

Question Label : Multiple Select Question

Consider the following three models for a one-dimensional dataset:

Model 1: $y = w_1 x_1$

Model 2: $y = w_1^2 x_1$

Model 3: $y = w_1^2 x_1 + w_2 x_1$

Select all the correct options. Assume that we have access to sufficiently large data points.

Options :

6406531737042. ✔ There may be some datasets for which model 1 performs better than model 2.

6406531737043. ✖ There may be some datasets for which model 2 performs better than model 1.

6406531737044. ✖ There may be some datasets for which model 3 performs better than model 1.

6406531737045. ✔ There may be some datasets for which model 3 performs better than model 2.

6406531737046. ✔ Model 1 and Model 3 perform equally well on all datasets.

Problem with model 2 is that it squares the weight vector.

$\Rightarrow$ −ve values will become positive.

$\Rightarrow$ Model 2 space is restricted to +ve values.

Model 3 is the min of both model 1 and model 2.
Even though theres 'w_1^2' is model 3 it can be balanced out with 'w_2'.
If w_2 is large enough it can even make the values negative.

$\Rightarrow$ Model 3 and Model 1, perform the same.

Kernel regression W5 notes.

Question Label : Multiple Select Question

Let w be the solution of the linear regression model and $\tilde{w}$ be the projection of w on the linear subspace spanned by the data points. Which of the following relationship is true?

Options :

6406531737047. ✔ training error for w = training error for $\tilde{w}$

6406531737048. ✔ $w = \tilde{w}$

6406531737049. ✖ training error for $w \neq$ training error for $\tilde{w}$

Now , lets see what's the difference between error functions of w^* and $\tilde{w}$

$$\sum_{i=1}^{n}(w^{*T}x_i - y_i) \quad \text{and} \quad \sum_{i=1}^{n}(\tilde{w}^T x_i - y_i)$$

w^* can be written as the sum of $\tilde{w}$ and the vector perpendicular to $\tilde{w}$, which is $\tilde{w}_\perp$

Note that $\tilde{w}_\perp$ is perpendicular to the (2-d) plane itself , which means it is perpendicular / orthogonal to all the points which lie in the plane (datapoints).

$$\begin{aligned} w^* &= \tilde{w} + \tilde{w}_\perp \\ w^{*T}x_i &= (\tilde{w} + \tilde{w}_\perp)^T x_i \\ &= \tilde{w}x_i + \tilde{w}_\perp^T x_i \\ &= \tilde{w}x_i \qquad \forall i \end{aligned}$$

$\tilde{w}_\perp^T x_i$ will become 0 as explained above.

Now we see that the error for both w^* and $\tilde{w}$ is exactly the same.

Bayesian modeling W6 Notes

Consider the following statement:

MAP estimate for linear regression weights w is equivalent to ridge regression.

Which of the following conditions make the above statement true?

Options :

6406531737050. ✖ Prior for w is Laplace distribution with zero mean.

6406531737051. ✔ Prior for w is $N(0, \gamma^2 I)$.

6406531737052. ✖ $y_i | x_i \sim N(0, \sigma^2 I)$

6406531737053. ✔ $y_i | x_i \sim N(w^T x_i, \sigma^2)$

Suppose you want to use a Naive Bayes classifier to predict the gender (male or female) of a person based on two features: their height (f_1) and whether their age is above 20 (f_2). Assume that the features f_1 and f_2 are conditionally independent given the gender of the person, and that the variances of the height distributions $P(f_1|y = \text{male})$ and $P(f_1|y = \text{female})$ are equal. How many parameters are required to classify a new example using this Naive Bayes classifier?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

6

Male and Female features can be thought of as binary features.

Age > 20 and Age < 20 can also be thought of as binary features. (f_2)

f_1 is a continuous feature with a gaussian distribution of $\mathcal{N}(\mu, \Sigma)$.

Label

Male

Female 1

$f_1|y = \text{male} = \mathcal{N}(\mu_1, \Sigma)$ 2

$f_2|y = \text{male} = f_2 = 0 | y = \text{male}$ 1

or

$f_2 = 1 | y = \text{male}$

$f_1|y = \text{female} = \mathcal{N}(\mu_2, \Sigma)$ 1

$f_2|y = \text{female} = f_2 = 0 | y = \text{female}$ 1

$f_2 = 1 | y = \text{female}$

$\Rightarrow$ Parameters $= \underline{\underline{6}}$

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

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

Assume that $\hat{p}$ and $\hat{p}_j^{y_i}$ are estimates for $P(y = 1)$ and $P(f_j = 1|y = y_i)$, respectively. Here, $f_i;\ i = 1, 2$ is the i^{th} feature. These parameters are estimated using MLE. If a test point has label 0, what will be the probability that the point is $[0, 0]^T$?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

0.5

$\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ Point

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

$0 \quad 1 \quad 0$ Prediction

Find $P([0,0] | y = 0)$

$\Rightarrow P([0,0] | y=0) = P(f_1 = 0 | y = 0) \cdot P(f_2 = 0 | y = 0)$

$\Rightarrow \frac{1}{2} \times \frac{2}{2}$

$\Rightarrow \underline{\underline{0.5}}$

Gaussian kernel regression with parameter $\sigma^2 = 1/2$ was applied to the following dataset with two features:

$$X = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix} \qquad y = [2.1, 1, 2, 1.2]^T$$

The weight vector can be written as $w = \phi(X)\alpha$, where ϕ is the transformation mapping corresponding to the kernel. The vector α is given by $[2.1, -2.1, 3, 0]^T$ which is obtained as $(K)^{-1}y$, where K is the kernel matrix. What will be the prediction for point $[1, 1]^T$?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

3

Given,

$w = \phi(X)\alpha$

$x_{test} = [1, 1]$

We know that prediction is,

$\Rightarrow w^T x_{test}$

$\Rightarrow (\phi(X)\alpha)^T (x_{test})$

$\Rightarrow (\alpha^T \cdot \phi(X)^T)(\phi(x_{test}))$

$\Rightarrow K(X, x_{test})\alpha$

$[K(X, x_{test})\alpha]$

Gaussian kernel is given by,

$$K(x_i, x_j) = \exp\left(\frac{-\|x_i - x_j\|^2}{2\sigma^2}\right)$$

$$\Rightarrow \exp(-\|x_i - x_j\|^2) \quad (\sigma^2 = 1/2)$$

$\Rightarrow [e^{-1}, e^{-1}, 1, e^{-2}]$

$\Rightarrow [e^{-1}, e^{-1}, 1, e^{-2}][2.1, -2.1, 3, 0]^T$

$\Rightarrow \underline{\underline{3}}$

Suppose we have a binary classification dataset with 1000 data points, consisting of 600 points belonging to class 0 and 400 points belonging to class 1. If we use a k-nearest neighbor (k-NN) model with $k = 900$ to predict the class labels of the data points, how many data points will be classified correctly?

Response Type : Numeric

[redacted]

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

600

We predict the majority of the 900 points using KNN here.

Even if 400 class 1 points are all closest neighbours, there are still 500 class 0 points [900 − 400 = 500] left as the closest neighbours.

Majority (400 class 1, 500 class 0) = class 0

$\Rightarrow$ Prediction will always be class 0 for any test datapoint.

$\Rightarrow$ 600 points have class 0.

Suppose we have 1000 training examples and want to compute the 10-fold Cross-Validation error. This error is calculated as the average of the errors obtained from n_1 iterations of the Cross-Validation process. Each iteration involves training a model on a subset of size n_2 of the training data and evaluating its performance on a disjoint subset of size n_3.

What is the appropriate value of n_1 ?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

10

n_1 is the number of iterations.

K-fold cross-validation has K iterations,

$\Rightarrow$ 10-fold cross-validation will have 10 iterations.

What is the appropriate value of n_2?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

900

n_2 is training set.

$\Rightarrow$ K-1 parts are used for training.

size of 1 part $= \frac{\text{no. of datapoints}}{\text{no. of folds}} = \frac{1000}{10} = 100$

$\Rightarrow 1000 - 100 = 900$ points are used for training.

Question Label : Short Answer Question

What is the appropriate value of n_3 ?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

100

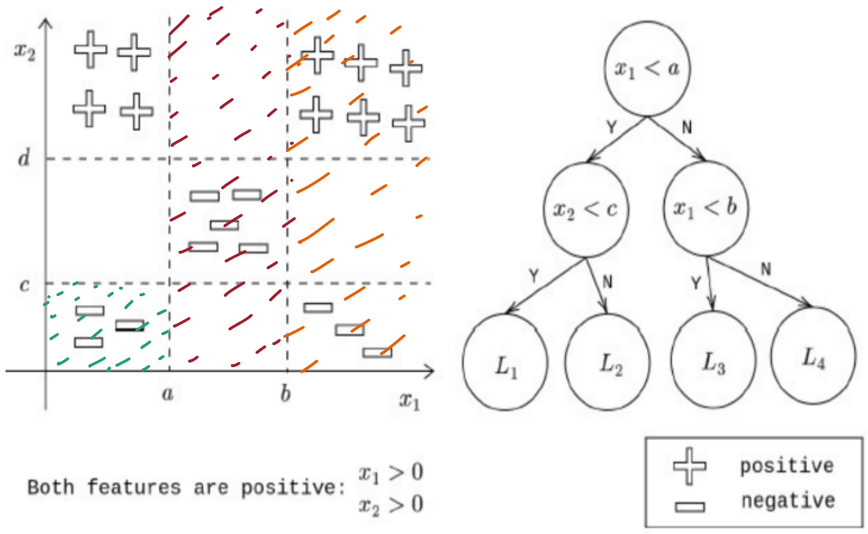
n_3 is validation set.

$\Rightarrow$ 1 part is used for validation

$\Rightarrow$ 100 points will be used for validation.

2. **K-Fold Cross Validation**: The training set is partitioned into K equally-sized parts. The model is trained K times, each time using K-1 parts as the training set and the remaining part as the validation set. The λ value that leads to the lowest average error is chosen.

Consider the following training dataset for a binary classification problem on the left and some decision tree for it on the right. The labels lie in the set $\{+1, -1\}$.



L_1, L_2, L_3, L_4 are leaves. The four dotted lines $x_1 = a, x_1 = b, x_2 = c, x_2 = d$ are drawn for your reference. Both features x_1 and x_2 are positive. Our focus will only be on the first quadrant. Use $\log_2$ for all entropy calculations. Calculate all intermediate quantities upto three decimal places.

The $+$, $-$ signs represent the datapoints in that region. (This was not given in the question, they should have given this info tho).

Question Label : Short Answer Question

What is the label of leaf L_2? 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

$x_1 < a \xrightarrow{Yes} x_2 < c \xrightarrow{Yes} \underline{\underline{L_1}}$

On the graph green dotted area is L_1.

We can see that L_1 is -1.

$\Rightarrow L_2 = +1$

What is the label of leaf L_4? 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

$x_1 < a \xrightarrow{No} x_1 < b \xrightarrow{Yes} \underline{\underline{L_3}}$

$\Rightarrow a < x_1 < b$

$\Rightarrow$ x_1 lies between a and b,

on the graph red dotted area is L_3.

$\Rightarrow$ Same of that L_3 is -1.

$\Rightarrow L_4 = +1$

Select all true statements regarding the decision boundary of the decision tree.

Options :

6406531737063. ✔ The dotted line $x_2 = d$ is **not** a part of the decision boundary. That is, not even a single point on $x_2 = d$ is a part of the decision boundary.

6406531737064. ✔ The entirety of the dotted line $x_1 = a$ is a part of the decision boundary. That is, every single point on the dotted line is a part of the decision boundary.

6406531737065. ✖ The entirety of the dotted line $x_2 = c$ is a part of the decision boundary. That is, every single point on the dotted line is a part of the decision boundary.

6406531737066. ✖ Only a finite segment of the dotted line $x_1 = b$ is a part of the decision boundary. That is, there are some points on the dotted line that are **not** a part of the decision boundary.

See the graph above.

True because in our decision tree we arent checking for $x_2 < d$ or $x_2 > d$.

True because our first decision is based on $x_1 < a$.

Only the 1st part of the 3 sections is part of decision boundary.

Full $x_1 = b$ is part of decision boundry as in our decision tree, we are checking for $x_1 < 5$

What is the entropy of the leaf L_3? Enter your answer correct to three decimal places.

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

0

L_3 is the red dotted area,

$p = \frac{5}{5} = 1$

We know entropy$(p=0)$ = entropy$(p=1) = 0$

$\Rightarrow \underline{\underline{0}}$

What is the entropy of the leaf L_4? Enter your answer correct to three decimal places.

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Range

Text Areas : PlainText

Possible Answers :

0.90 to 0.93

L_4 is orange dotted area,

$p = \frac{6}{9} = \frac{1}{3}$

$\Rightarrow \text{Entropy}(1/3) = -\left(\frac{1}{3}\log_2\left(\frac{1}{3}\right) + \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right)$

$\Rightarrow -\left(0.33 \times (-1.584) + 0.66(-0.584)\right)$

$\Rightarrow \underline{\underline{0.908}}$

Information gain = Entropy at root − Weighted entropy of leaves

Enter your answer correct to three decimal places.

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

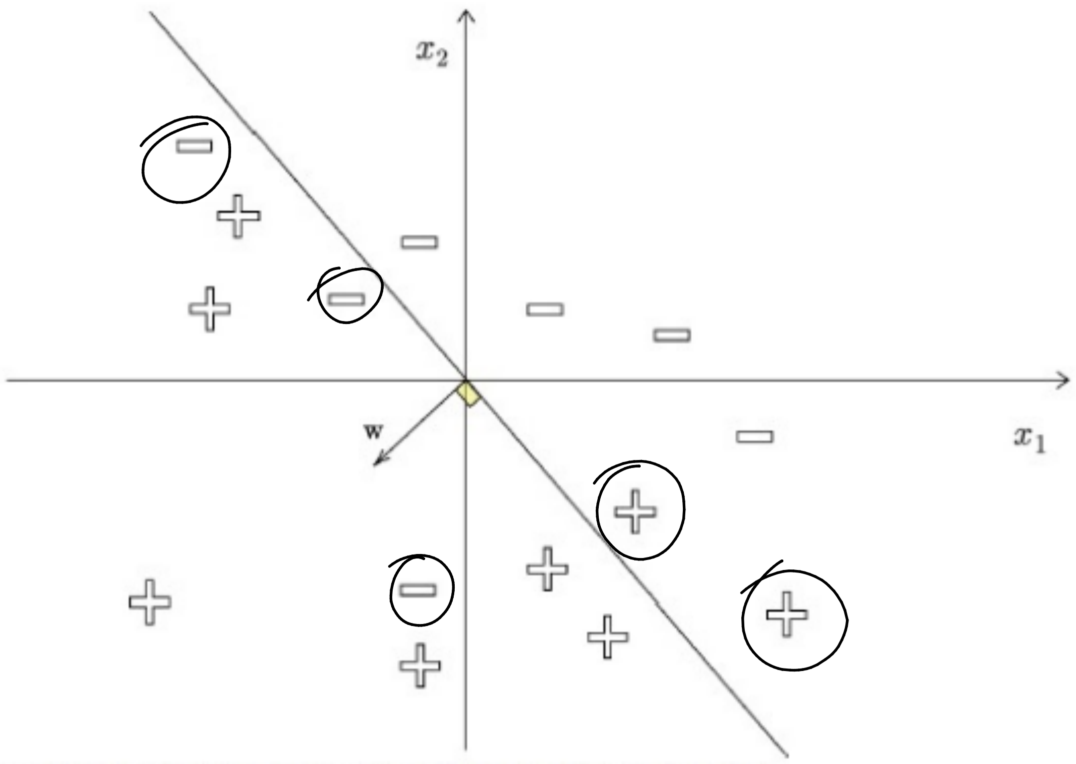
Answers Type : Range

Text Areas : PlainText

Possible Answers :

0.58 to 0.62

Consider the following training dataset for a binary classification problem that has 15 data-points. The labels are in the set $\{+1, -1\}$. The symbol **+** is a data-point with label $+1$ and **−** is a data-point with label -1.



w is the weight-vector corresponding to a linear classifier.

How many points are misclassified by the classifier? Circled points are misclassified

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

5

Consider another linear classifier with $\mathbf{w}' = 3\mathbf{w}$. How many points are misclassified by this new classifier?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

5

w' is just a scaled up weight vector.
⇒ The direction w' hasnt changed.
⇒ Linear classification line is still the same.

AN4 Exam QPI 4 6-Aug-2023

Consider a training dataset of n points for a regression problem. Assume that the model is linear. Let $\mathbf{w}_1$ and $\mathbf{w}_2$ be the optimal weight vectors obtained from solving the following optimization problems.

$$\mathbf{w}_1 = \arg\min_{\mathbf{w}} \sum_{i=1}^{n} (\mathbf{w}^T \mathbf{x}_i - y_i)^2$$

$$\mathbf{w}_2 = \arg\min_{\mathbf{w}} \sum_{i=1}^{n} (\mathbf{w}^T \mathbf{x}_i - y_i)^3$$

Cubic values can be negative. Sum of all those values might become 0, which wouldn't help us in drawing any conclusions.

Choose the most appropriate answer.

Options :

6406531963440. ✔ $\mathbf{w}_1$ will generalize better than $\mathbf{w}_2$ on the test dataset.

6406531963441. ✖ $\mathbf{w}_2$ will generalize better than $\mathbf{w}_1$ on the test dataset.

6406531963442. ✖ Both models will show identical performance on the test dataset.

The training dataset for a binary classification problem is as follows:

$$\{\,(\mathbf{u}, 1),\quad (-\mathbf{u}, 0),\quad (2\mathbf{u}, 1),\quad (-2\mathbf{u}, 0)\,\}$$

where, $\mathbf{u} \in \mathbb{R}^d$ is a non zero constant and each element in the set given above is a data-point of the form $(\mathbf{x}_i, y_i)$. The labels lie in $\{0, 1\}$. Consider a linear classifier with weight vector $\mathbf{w}$. What condition should the weight vector satisfy for the zero-one loss to be zero on this dataset?

Options :

We can see that all +ve 'u' values have 1 as label, while -ve values have 0 label.

Conventionally we assign label 1 if condition is satisfied.

6406531963443. ✖ $\mathbf{w}^T\mathbf{u} < 0$

6406531963444. ✔ $\mathbf{w}^T\mathbf{u} > 0$

6406531963445. ✖ $\mathbf{w}^T\mathbf{u} = 0$

6406531963446. ✖ We can never find a $\mathbf{w}$ for which the zero-one loss becomes zero on this dataset.

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 :

6406531963447. ✔ Adding a dummy feature in the dataset and learning the intercept w_0 as well.

If the dataset is not centered, we add intercept w_0, which pushes our line towards the datapoints.

6406531963448. ✖ Penalizing the model weights with L2 regularization.

6406531963449. ✖ Penalizing the model weights with L1 regularization.

6406531963450.

✔ Training the kernel regression model of degree 2.

Sometimes the relationship between datapoints might not be linear, degree 2 kernel regression helps overcome that.

Which of the following statements are true about the decision tree algorithm?

Options :

6406531963455. ✖ Decision trees are prone to overfit if the maximum depth is set too low.

6406531963456. ✔ Decision trees are prone to underfit if the maximum depth is set too low.

6406531963457. ✔ Decision trees are sensitive to small perturbations in the dataset and can result in different tree structures.

6406531963458. ✔ Decision trees can handle both numerical and categorical features.

Which of the following statements is/are true regarding solution of Ridge regression problem?

Options :

6406531963451. ✔ If there are multiple w solutions for minimizing mean square error, then w_R will be the one with least norm.

W6 Notes,

- Bayesian Modeling for Linear Regression.
- Ridge Regression.

6406531963452. ✖ If there are multiple w solutions for minimizing mean square error, then w_R will be the one with highest norm.

6406531963453. ✔ Prior for w is $\mathrm{N}(0, \gamma^2 I)$ and $y_i | x_i \sim \mathrm{N}(w^T x_i, \sigma^2)$

6406531963454. ✖ Prior for w is $\mathrm{N}(1, \gamma^2 I)$ and $y_i | x_i \sim \mathrm{N}(0, \sigma^2)$

Consider kernel regression with the kernel function $(\mathbf{x}_1^T\mathbf{x}_2 + 2)^2$ applied on the following dataset.

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

The optimal weight vector $\mathbf{w}^*$ is given by:

$$\mathbf{w}^* = \phi(X)[0.1, 2, 3.9, 5, 6, 8]^T$$

where ϕ is transformation mapping corresponding to the given kernel. What will be the prediction for the data point $[0, 0, 1]^T$?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

100

$$K(x_1, x_2) = (x_1^T x_2 + 2)^2$$

$$\omega^* = \phi(X)[0.1, 2, 3.9, 5, 6, 8]^T$$

We know,

$$\text{Prediction} = \omega^T x_{test}$$

here,

$$\omega = \omega^*$$

$$x_{test} = [0, 0, 1]^T$$

Using Prediction formula,

$$\Rightarrow \omega^{*T} x_{test}$$

$$\Rightarrow \left(\phi(X)[0.1, 2, 3.9, 5, 6, 8]^T\right)^T \left([0, 0, 1]^T\right)$$

$$\Rightarrow \left(\phi(X)^T [0, 0, 1]^T\right)\left([0.1, 2, 3.9, 5, 6, 8]\right)$$

$$\Rightarrow (\phi(X))^T (\phi[0, 0, 1]^T)$$

$$\Rightarrow K(X, [0, 0, 1]^T)$$

[From here I'll be solving for transformation only, will come to this vector later]

$\Rightarrow$ After solving for kernel mapping, we get

$$\Rightarrow [4, 4, 4, 4, 4, 4]^T$$

$$\Rightarrow 4[1, 1, 1, 1, 1, 1]^T$$

Multiplying this vector with the above vector we get

$$\Rightarrow 4(25) = \underline{\underline{100}}$$

Consider a ridge regression model with the loss $L(\mathbf{w}) = ||X^T\mathbf{w} - \mathbf{y}||^2 + \lambda||\mathbf{w}||^2$ is trained on a given dataset with $\lambda = 0.1, 0, 1, 10, 100$. Which of the following value of λ is more likely to underfit the model?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

100

Higher Penalty (λ) = Lower weights of w = underfitting

Highest Penalty here is $\lambda = 100$

Consider the following data set:

$X = [8, 6, 10]$

Assuming a ridge penalty $\lambda = 100$, what will be the value of $\frac{\hat{w}_{ridge}}{\hat{w}_{MLE}}$?

Here $\hat{w}_{ridge}$ and $\hat{w}_{MLE}$ are the Ridge and MLE estimates of the weight vectors, respectively. Assume that the label vector y of shape $(3, 1)$ is known. 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 :

0.65 to 0.70

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

$$\hat{\omega}_{Ridge} = (XX^T + \lambda I)^{-1} Xy$$

$$\hat{\omega}_{MLE} = (XX^T)^{-1} Xy$$

$$XX^T = [8,6,10]\begin{bmatrix}8\\6\\10\end{bmatrix} = 64+36+100 = 200$$

$$\lambda I = 100 \times 1 = 100$$

$$\frac{\hat{\omega}_{Ridge}}{\hat{\omega}_{MLE}} = \frac{(XX^T + \lambda I)^{-1} \cancel{Xy}}{(XX^T)^{-1} \cancel{Xy}}$$

$$\Rightarrow \frac{(200+100)^{-1}}{(200)^{-1}} = \frac{200}{300} = \frac{2}{3}$$

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

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

What will be the prediction for the point $x = 1$? Assume that $\hat{p}$, an estimate for $P(y=1)$, is 0.5.

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

0

To predict for a datapoint we find,

$$P(y=0|x=1) \geq P(y=1|x=1)$$

$$\Rightarrow \frac{P(x=1|y=0).\cancel{P(y=0)}}{\cancel{P(x=1)}} \geq \frac{P(x=1|y=1).\cancel{P(y=1)}}{\cancel{P(x=1)}}$$

$$\Rightarrow P(x=1|y=0) \geq P(x=1|y=1)$$

$P(y=1) = 0.5$

$\Rightarrow P(y=0) = 1 - P(y=1)$

$\Rightarrow 1 - 0.5$

$\Rightarrow 0.5$

$$\Rightarrow \frac{\cancel{1}}{\cancel{\sqrt{2\pi}}(1)} \exp\left(-\frac{1}{2}\left(\frac{1-0}{1}\right)^2\right) \geq \frac{\cancel{1}}{\cancel{\sqrt{2\pi}}\sqrt{2}} \exp\left(-\frac{1}{2}\left(\frac{1-2}{\sqrt{2}}\right)^2\right)$$

$$\Rightarrow e^{-1/2} \geq \frac{1}{\sqrt{2}} e^{-1/4}$$

$$\Rightarrow 1 \geq \frac{1}{\sqrt{2}} e^{-\frac{1}{4}+\frac{1}{2}}$$

$$\Rightarrow 1 \geq \frac{1}{\sqrt{2}} e^{1/4}$$

$$\Rightarrow 1 \geq 0.9$$

$\Rightarrow$ Label will be predicted as zero.

Consider a binary classification problem and a decision tree that is being trained to classify the points. In one of the internal nodes in this tree, 75% of the data-points belong to one of the two classes and the rest belong to the other class. You are not given the information about which class is more numerous in this node.

Based on the above data, answer the given subquestions.

Do you have enough information to find the entropy of this node?

Options :

6406531963463. ✔ Yes

6406531963464. ✖ No

As 75% points belong to a node.

$\Rightarrow p = \frac{75}{100} = 0.75$

Note that entropy is a symmetrical function.

$\Rightarrow$ Entropy$(p=0.75)$ = Entropy$(p=0.25)$

If the answer to the previous questions is "Yes", find the entropy of the node. Use $\log_2$ and enter your answer correct to three decimal places.

If the answer to the previous question is "No", enter -1 as your answer.

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Range

Text Areas : PlainText

Possible Answers :

0.79 to 0.83

$$\text{Entropy} = -\left(p\log_2 p + (1-p)\log_2(1-p)\right)$$

$$\Rightarrow -\left(0.25\log_2(0.25) + 0.75\log_2(0.75)\right)$$

$$\Rightarrow -\left(0.25(-2) + 0.75(-0.415)\right)$$

$$\Rightarrow 0.5 + 0.31125$$

$$\Rightarrow 0.811$$

Consider a probability distribution over (X, y) where features are one-dimensional and $y \in \{+1, -1\}$. Let $X|(y=1)$ follow a uniform distribution over $[0,4]$ and $X|(y=-1)$ follows a uniform distribution over $[2,4]$.

Given,

$X|(y=1) \sim \text{Uniform}(0,4)$

$X|(y=-1) \sim \text{Uniform}(2,4)$

Based on the above data, answer the given subquestions.

If $p = P(y=1)$ is estimated to be 0.4, what will be the prediction for the point $x = 3$ using the Bayes classifier? 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

Given,

$P(y=1) = 0.4$

$\Rightarrow P(y=-1) = 0.6$

Find: $P(y|x=3)$

$$\Rightarrow P(y=1|x=3) = \frac{P(x=3|y=1) \cdot P(y=1)}{P(x=3)} \quad \text{—} \quad 1$$

$$P(y=-1|x=3) = \frac{P(x=3|y=-1) \cdot P(y=-1)}{P(x=3)} \quad \text{—} \quad 2$$

When comparing 1 and 2 we can ignore the denominator $P(x=3)$ as its common in both.

For 1

$\Rightarrow \left(\frac{1}{4-0}\right) \times \frac{4}{10}$

$\Rightarrow \frac{1}{10}$

$\Rightarrow 0.1$

For 2

$\left(\frac{1}{4-2}\right) \times \left(\frac{6}{10}\right)$

$\frac{3}{10}$

0.3

$0.3 > 0.1$

$\Rightarrow$ Predicted label will be -1.

Let $x = 2$ and let $\hat{p}$ be the estimate for $p = P(y = 1)$. Find conditions on $\hat{p}$ such that the Bayes classifier predicts 1 for this x. Consider that the tie-breaker is predicted in class 1.

Options :

6406531963467. ✖ $\hat{p} \leq \frac{1}{4}$

6406531963468. ✖ $\hat{p} \geq \frac{1}{4}$

6406531963469. ✖ $\hat{p} \leq \frac{2}{3}$

6406531963470. ✔ $\hat{p} \geq \frac{2}{3}$

Our Bayes prediction equation for $\hat{p}$ will be

$$\Rightarrow \underbrace{\frac{1}{4} \cdot \hat{p}}_{\text{label } 1} \geqslant \underbrace{\frac{1}{2} \cdot (1 - \hat{p})}_{\text{label } -1}$$

$$\Rightarrow \cancel{\frac{1}{2}}\left(\frac{1}{2}\hat{p}\right) \geqslant \cancel{\frac{1}{2}}(1 - \hat{p})$$

$$\Rightarrow \frac{1}{2}\hat{p} \geqslant 1 - \hat{p}$$

Adding $\hat{p}$ to both sides

$$\Rightarrow \frac{3}{2}\hat{p} \geqslant 1$$

$$\Rightarrow \hat{p} \geqslant \frac{2}{3}$$

If $\hat{p} \geqslant \frac{2}{3}$ then our equation will predict label 1.

If $p = P(y = 1)$ is estimated to be 0.5 using MLE on a given training dataset, what will be the training error of the Bayes classifier for this problem?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

0.5

?

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

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

Assume that $\hat{p}$ and $\hat{p}_j^{y_i}$ are estimates for $P(y=1)$ and $P(f_j = 1 | y = y_i)$, respectively. Here, $f_i;\ i = 1, 2, 3$ is the i^{th} feature. These parameters are estimated using MLE. Do not apply any smoothing on the dataset.

Based on the above data, answer the given subquestions.

Calculate the value of $\hat{p}_2^0$.

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

0.5

$$\hat{p}_2^0 = \frac{\sum_{i=1}^{n} 1(f_2 = 1 | y = 0)}{\sum_{i=1}^{n} 1(y=0)}$$

$$\Rightarrow \quad 1/2 = \underline{\underline{0.5}}$$

Calculate the value of $\hat{p}_2^1$.

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

0

$$\hat{p}_2^1 = \frac{\sum_{i=1}^{n} 1(f_2 = 1 | y_i = 1)}{\sum_{i=1}^{n} 1(y_i = 1)}$$

$$\Rightarrow \quad \frac{0}{2} = \underline{\underline{0}}$$

AN2 Exam QPDI 20-Nov-2022

Let X be the data matrix of shape (d, n) and y be the corresponding label vector. A linear regression model of the form $\hat{y}_i = w^T x_i$ is fit using the squared error on the same dataset. If the solution w^* to the optimization problem is orthogonal to the subspace spanned by the data points (columns of matrix X), what will be the squared error?

Options :

6406531484863. ✖ 0

6406531484864. ✖ 1

6406531484865. ✔ $||y||^2$

6406531484866. ✖ Insufficient information to answer

$$MSE = \sum_{i=1}^{n} (w^T x_i - y_i)^2$$

Here w is orthogonal to x_i

$\Rightarrow w^T x_i = 0$

$$\Rightarrow MSE = \sum_{i=1}^{n} (-y_i)^2$$

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

$$\Rightarrow ||y||^2$$

Which of the following regression model will certainly achieve zero training error on a given training dataset where the error is defined as the sum of squared error? Assume that $x_i \in \mathbb{R}^d$ is the i^{th} data point and $y_i \in \mathbb{R}$ is the corresponding label.

Options :

6406531484867. ✖ $h(x_i) = \overline{y}\ \forall i$, where $\overline{y}$ is the average of all the labels.

6406531484868. ✖ $h(x_i) = w^T x_i\ \forall i$, where $w \in \mathbb{R}^d$

6406531484869. ✖ $h(x_i) = c$ where c is a constant.

6406531484870. ✔ $h(x_i) = y_i\ \forall i$

Let w^*, w^g, and w^{sg} be the weight vectors obtained using analytical, gradient descent, and stochastic gradient descent approaches, respectively, on the same linear regression model. The following expression holds true for these weight vectors:

$$||w^g - w^*|| < ||w^{sg} - w^*||$$

The model obtained by the analytical solution gives a training error of 0.5. Which of the following approaches is more likely to give less training error? Assume that the loss function is a convex function.

Options :

6406531484871. ✔ Gradient descent

6406531484872. ✖ Stochastic gradient descent

weight features of w^{sg} are greater than, weight features of w^g.

$$MSE = \sum_{i=1}^{n} (w^T x_i - y_i)^2$$

Higher weights will give higher MSE.

Question Label : Multiple Choice Question

Consider the following data set:

$X = [-3, 5, 4]$

$y = [-10, 20, 20]$

Assuming a ridge penalty $\lambda = 50$, what will be the value of $\frac{\hat{w}_{ridge}}{\hat{w}_{MLE}}$?

Here $\hat{w}_{ridge}$ and $\hat{w}_{MLE}$ are the Ridge and MLE estimates of the weight vectors, respectively.

Options :

6406531484873. ✖ 2

Similar ques done above

6406531484874. ✖ 1

6406531484875. ✖ 0.666

6406531484876. ✔ 0.5

6406531484877. ✖ 0.25

Question Label : Multiple Choice Question

Consider the following data $\{(x_1, y_1), (x_2, y_2), (x_3, y_3)\}$:

x	y
0	2
2	2
3	1

Assume that Leave one out cross validation is applied on this data.

Note: The model to be used is $y = w_0 + w_1 x$.

What will be the weights obtained when (x_2, y_2) is used in the validation set?

→ Leave this out

Options :

6406531484878. ✖ $\{w_0 : 4, w_1 : -1\}$

6406531484879. ✖ $\{w_0 : 2/5, w_1 : 0\}$

6406531484880. ✖ $\{w_0 : 4, w_1 : -2/5\}$

6406531484881. ✔ $\{w_0 : 2, w_1 : -1/3\}$

$y = w_0 + w_1 x$

When $(x=0, y=2)$

$\Rightarrow 2 = w_0 + w_1(0)$

$\Rightarrow 2 = w_0$

When $(x=3, y=1)$

$\Rightarrow 1 = w_0 + w_1(3)$

$\Rightarrow 1-2 = 3w_1$

$\Rightarrow w = \frac{-1}{3}$

A Gaussian naive Bayes model is trained on a given dataset. For an unseen data point x, the following two values are calculated as

$$P(x|y=0) = 0.4$$
$$P(x|y=1) = 0.6$$

What will be the predicted label for x?

Options :

6406531484886. ✖ 0

6406531484887. ✖ 1

6406531484888. ✔ Insufficient information to make a prediction

$P(y=0)$ or $P(y=1)$ is not given.

The training dataset for a binary classification problem has 100 points, 50 of which belong to class $+1$. Consider a k-NN algorithm with $k=1$ that is used to predict the labels of the training data-points. A point is considered as its own neighbor. Based on this setup, study the following statements:

S1: The number of points that are **misclassified** by the classifier is zero.

S2: Since the training error is zero, we have found a very good classifier for this problem.

Options :

6406531484882. ✔ S1 is true but S2 is false

6406531484883. ✖ S1 is false but S2 is true

6406531484884. ✖ Both S1 and S2 are true

6406531484885. ✖ Both S1 and S2 are false

There is no majority label here, its a tie.
50 for +1, 50 for −1

Even if training error is zero $k=1$ underfits data.
⇒ We havent found good classifier.

You know the distribution $P(X, y)$ for a given dataset $\{X, y\}$. Can you always find the distribution $P(y|X)$ for the same dataset $\{X, y\}$?

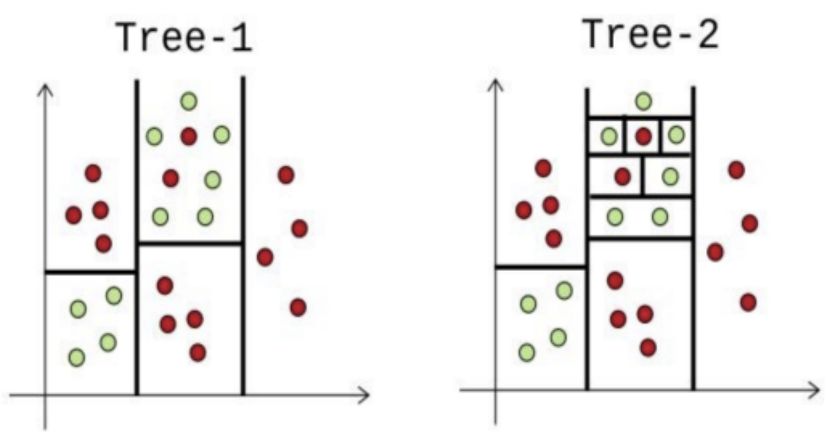
Options :

6406531484889. ✔ Yes

6406531484890. ✖ No

?

Consider a training dataset for a binary classification problem in $\mathbb{R}^2$. Two decision trees are trained on the same dataset. The decision regions obtained are plotted for both the trees:



Which of these two trees is likely to perform better on test data?

Options :

6406531484900. ✔ Tree-1

6406531484901. ✖ Tree-2

Tree −2 is over fit.

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

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

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

Choose all linear classifiers that result in zero misclassifications on this dataset. Here, $\mathbf{w}$ is the weight vector for the linear classifier.

Options :

6406531484891. ✓ $\mathbf{w} = \begin{bmatrix}1\\2\end{bmatrix}$

6406531484892. ✓ $\mathbf{w} = \begin{bmatrix}10\\19\end{bmatrix}$

6406531484893. ✖ $\mathbf{w} = \begin{bmatrix}-1\\-4\end{bmatrix}$

6406531484894. ✖ $\mathbf{w} = \begin{bmatrix}-5\\3\end{bmatrix}$

6406531484895. ✖ $\mathbf{w} = \begin{bmatrix}5\\-3\end{bmatrix}$

Calc $w^T x$ for all w and x.
if $w^T x > 0$, prediction = 1
else, prediction = -1

Only $w = \begin{bmatrix}1\\2\end{bmatrix}$ and $w = \begin{bmatrix}10\\19\end{bmatrix}$ predict all points correctly.

Kernel regression with a polynomial kernel of degree three is applied on a data set $\{X, y\}$. Let the weight vector be given by

$$w = \phi(X)[1.3, 0.6, -0.2, -0.7]^T$$

Here $\phi(X)$ is the transformed data matrix whose i^{th} column is $\phi(x_i)$. What will be the prediction for the data point $[0, 0, 0, 0]^T$?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

1

$K(x_i, x_j) = (1 + x_i x_j)^3$

As prediction datapoint $= [0, 0, 0, 0]$

Kernel function will always give

$\Rightarrow K(x_i, [0,0,0,0]) = (1+0)^3$
$\Rightarrow 1$

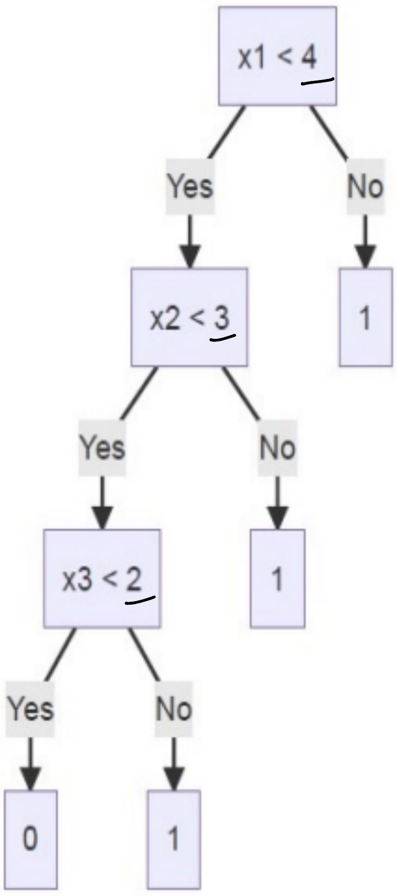
$\Rightarrow$ Our datapoint will be $[1, 1, 1, 1]^T$

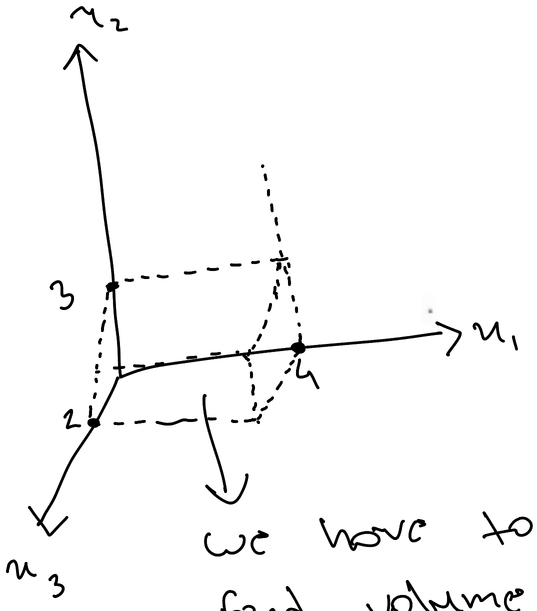
$\Rightarrow [1.3, 0.6, -0.2, -0.7]\,[1, 1, 1, 1]^T$

$\Rightarrow 1.9 - 0.9 = 1$

Consider a dataset in $\mathbb{R}^3$. Each data-point is represented by $\mathbf{x} = [x_1 \quad x_2 \quad x_3]^T$. The features in this problem are all positive. That is, $x_1, x_2, x_3 > 0$ for all data-points. Consider the following decision tree trained on this dataset. The features are represented without the subscript in the nodes: for example, x_1 is represented as $x1$.

Volume = 4x3x2 = 24





we have to find volume of this cuboid as this cuboid represents set S.

Consider only those points for which x_1, x_2, and x_3 are all positive.

Let S be the set of all points in $\mathbb{R}^3$ that are predicted as 0 by this decision tree. What is the volume of the region S?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

24 → 24

Suppose you have a three-class classification problem where class label $y \in \{0, 1, 2\}$ and each training example x_i has 3 binary features $f_1, f_2, f_3 \in \{0, 1\}$. How many parameters do you need to know to classify an example using the Naive Bayes classifier?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

11

Label

0 1 2

$f_1 \mid y = 0$	$f_1 \mid y = 1$	$f_1 \mid y = 2$
$f_2 \mid y = 0$	$f_2 \mid y = 1$	$f_2 \mid y = 2$
$f_3 \mid y = 0$	$f_3 \mid y = 1$	$f_3 \mid y = 2$

$3 + 3 + 3$

$\Rightarrow 9 + 2 = 11$

$P(y=0)$ and $P(y=1)$

Consider fitting a linear regression model (as stated below) for the following data:

x	y	$\hat{y}$
-1	1	β_0
0	-1	β_0
2	1	β_0

Fit $y_i = \beta_0$. Find β_0.

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Range

Text Areas : PlainText

Possible Answers :

0.3 to 0.4

Let $\hat{y}$ denote predicted datapoints.

$\Rightarrow \text{Error} = \sum_{i=1}^{n} (\beta_0 - y_i)^2$

$\Rightarrow \text{Error} = (\beta_0 - 1)^2 + (\beta_0 - -1)^2 + (\beta_0 - 1)^2$

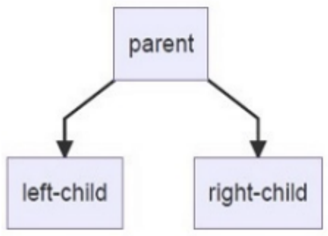
To optimize this function, we will take derivative

$\Rightarrow 0 = 2(\beta_0 - 1) + 2(\beta_0 + 1) + 2(\beta_0 - 1)$

$\Rightarrow 4(\beta_0 - 1) + 2(\beta_0 + 1) = 0$

$\Rightarrow 6\beta_0 - 2 = 0$

$\Rightarrow \beta_0 = \frac{2}{6} = \frac{1}{3} = 0.33$



Consider a decision stump for a binary classification problem that has 500 data points at the parent node, out of which 200 data points go into the left child. The number of data points that belong to class 1 in the parent node is 300. The number of data points that belong to class 1 in the left child is 50. The labels are in $\{1, 0\}$.

Note for calculations: Use $\log_2$ for all calculations that involve logarithms. For all questions, enter your answer correct to three decimal places. Use three decimal places even while calculating intermediate quantities.

Based on the above data, answer the given subquestions.

What is the label assigned to the left child? Enter 1 or 0 .

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

0

Left child has 200 data points.
50 of them have label 1.
$\Rightarrow$ 150 of them have label 0.

What is the entropy of the parent?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Range

Text Areas : PlainText

Possible Answers :

0.92 to 1

Parent has 500 data points.
300 of them belong to class 1.

$$\Rightarrow p = \frac{300}{500} = \frac{3}{5}$$

$$\Rightarrow \text{Entropy} = -\left(\frac{3}{5}\log_2\left(\frac{3}{5}\right) + \frac{2}{5}\log_2\left(\frac{2}{5}\right)\right)$$

$$\Rightarrow -\left(\frac{3}{5}(-0.736) + \frac{2}{5}(-1.321)\right)$$

$$\Rightarrow \underline{0.97}$$

What is the entropy of the left child?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Range

Text Areas : PlainText

Possible Answers :

0.80 to 0.83

$$p = \frac{5\cancel{0}}{20\cancel{0}} = \frac{1}{4}$$

$$\Rightarrow \text{Entropy} = -\left(\frac{1}{4}\log_2\left(\frac{1}{4}\right) + \frac{3}{4}\log_2\left(\frac{3}{4}\right)\right)$$

$$\Rightarrow -\left(\frac{1}{4}(-2) + \frac{3}{4}(-0.4150)\right)$$

$$\Rightarrow 0.5 + 0.311$$

$$\Rightarrow \underline{0.811}$$

What is the entropy of the right child?

Response Type : Numeric

Evaluation Required For SA : Yes

[redacted]

Show Word Count : Yes

Answers Type : Range

Text Areas : PlainText

Possible Answers :

0.62 to 0.68

Right child has 300 datapoints.

$\Rightarrow$ Class 1 datapoints $= \underbrace{300}_{\text{class 1 From parent}} - \underbrace{50}_{\text{class 1 From child}} = 250$

$$\Rightarrow p = \frac{\cancel{250}^{5}}{\cancel{300}_{6}} = \frac{5}{6}$$

$$\Rightarrow \text{Entropy} = -\left(\frac{5}{6}\log_2\left(\frac{5}{6}\right) + \frac{1}{6}\log_2\left(\frac{1}{6}\right)\right)$$

$$\Rightarrow -\left(\frac{5}{6}(-0.263) + \frac{1}{6}(-2.584)\right)$$

$$\Rightarrow \underline{\underline{0.649}}$$

What is the information gain corresponding to the question at the parent node?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Range

Text Areas : PlainText

Possible Answers :

0.21 to 0.29

$$\Rightarrow \text{Entropy}(D) - [\gamma\,\text{Entropy}(D_{yes}) + (1-\gamma)\,\text{Entropy}(D_{no})]$$

$$\Rightarrow 0.97 - \left[\frac{3}{5}\times 0.649 + \frac{2}{5}\times 0.811\right]$$

$$\Rightarrow 0.97 - 0.7138$$

$$\Rightarrow \underline{\underline{0.257}}$$