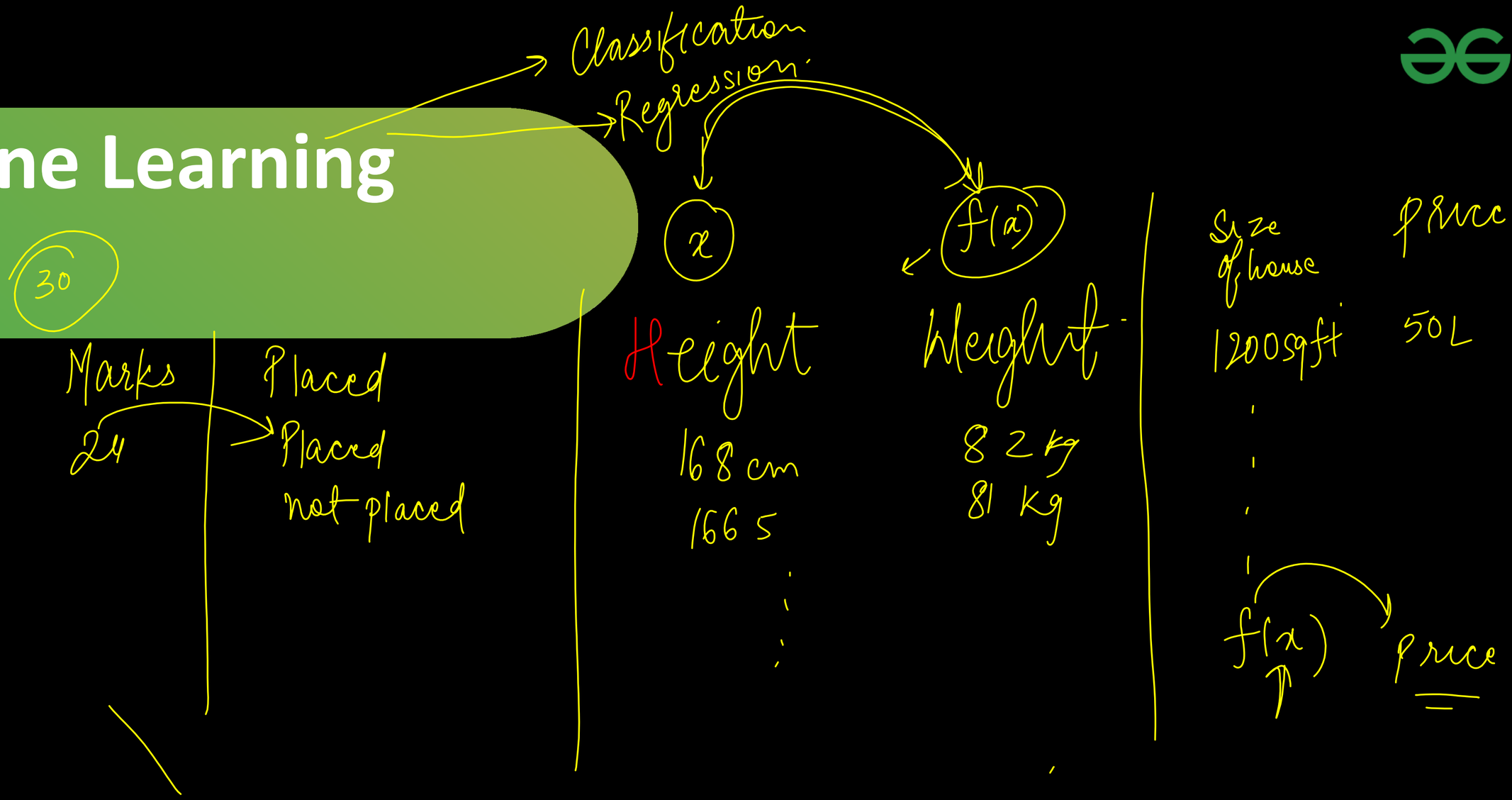
Machine Learning



Explain whether each of the following situations is a classification or regression problem:

- A company wants to launch a new product and wants to know whether it will turn out to be a success or failure. We have information on the last 100 products this company launched, including if it was a success/failure, price, weight, color, and several other variables.
- We have information on several Bay Area Tech Companies, including size, industry, revenue, average employee salary, and more. We want to know which features influence the average employee salary.
- You are given data of 100 individuals and their sequenced DNA and want to know whether these individuals will exhibit a particular disease based off their genomic mutations. We have information on 10,000 individual genomes and whether or not they exhibit the particular disease.

classification

$x_1 \; x_2 \; - \; - \; x_n \; f(x)$

Class

You are given reviews of few movies marked as positive, negative or neutral. Classifying reviews of a new movie is an example of

a. Supervised learning
b. Unsupervised learning
c. Semi-Supervised learning
d. Reinforcement learning

Imagine a newly-born starts to learn walking. It will try to find a suitable policy to learn walking after repeated falling and getting up. Specify what type of machine learning algorithm is best suited to do the same.

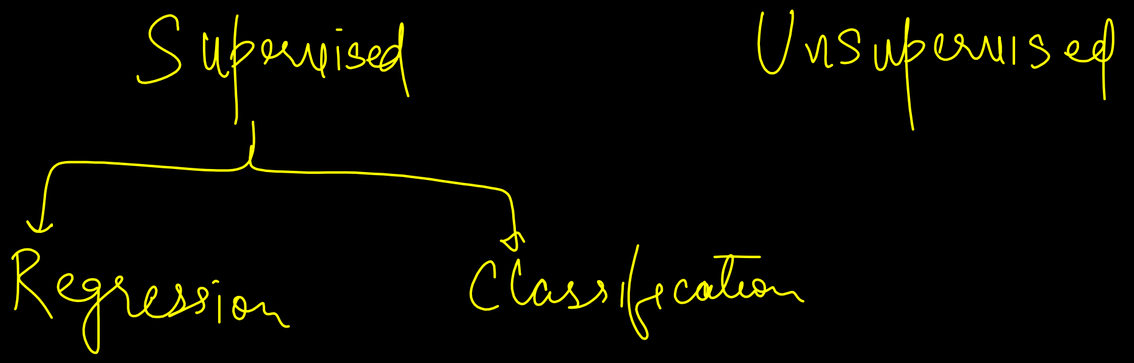
a. Supervised Learning
b. Unsupervised Learning
c. Reinforcement Learning
d. Semi-supervised Learning

Which of the following are supervised learning problems (Multiple Correct)?

A. Clustering Spotify users based on their listening history
B. Weather forecast using data collected by a satellite
C. Predicting tuberculosis using patient's chest X-Ray
D. Training a humanoid to walk using a reward system

Classify the following as regression or classification tasks

A. Predicting the outcome of an election → class
B. Predicting the weight of a giraffe based on its height → reg
C. Predicting the emotion conveyed by a sentence → class
D. Predict the temperature for the next day → reg
E. Detect pneumonia from chest X-ray image → Class

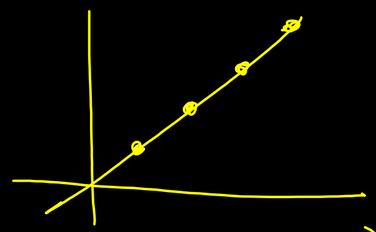


$x \neq y \} \; y = ax + b$

Linear Regression

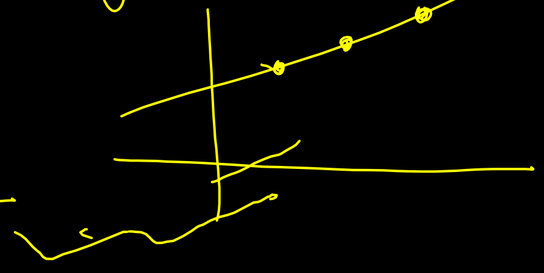
x	y
2	4
3	6
1	2
4	8

$y = 2x$



x	y	y'
2	5	5
3	7	7
1	3	3
4	9	9

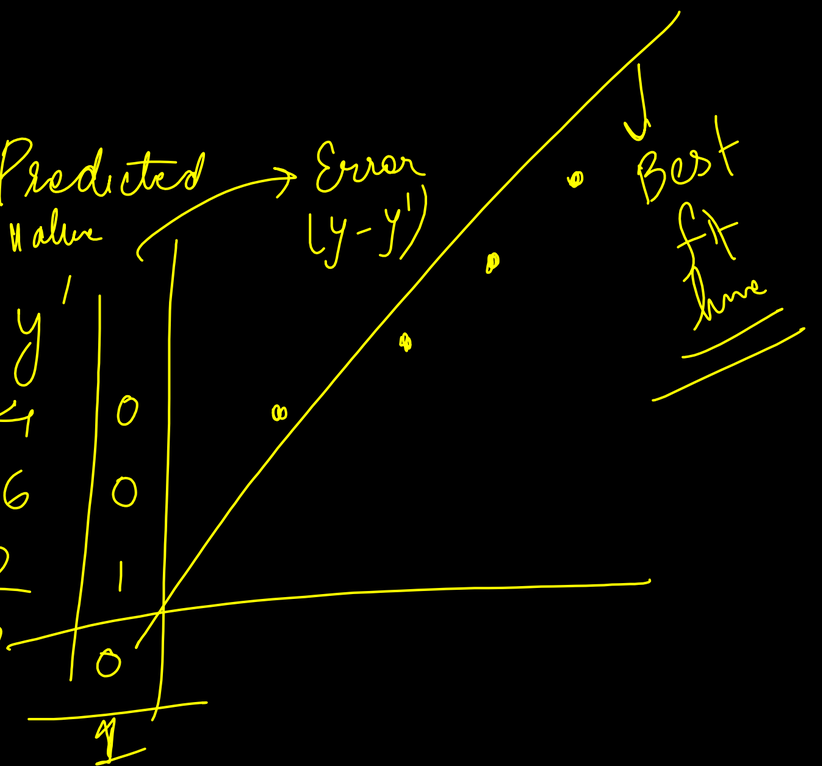
$y = 2x + 1 \quad y' = 2x + 1$



Actual value — Predicted value → Error $(y - y')$

x	y	y'	
2	4	4	0
3	6	6	0
1	3	2	1
4	8	8	0
			1

$y' = 2x \longrightarrow$ Predicted value



x	y	y'	Error $(y-y')$	Error^2	$\lvert$Error$\rvert$
1	3	2	1	1	1
2	4	4	0	0	0
3	6	6	0	0	0
4	8	8	0	0	0
5	9	10	−1	1	1
			0	2	2

$y = 2x$

Mean Squared Error $= \sum_{i=1}^{n} \frac{1}{2n}(y_i - y_i')^2$

$= \frac{1}{2n}\left[(y_1-y_1')^2 + (y_2-y_2')^2 + \cdots (y_n-y_n')^2\right]$

Mean Absolute Error

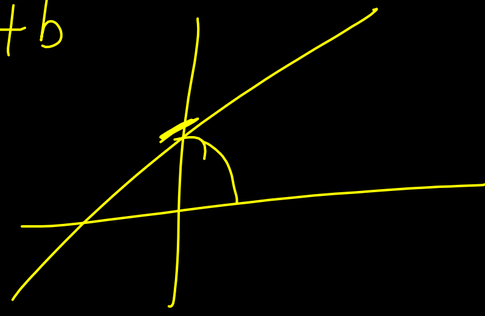
$= \sum_{i=1}^{n} \frac{1}{2n} |y_i - y_i'|$

What is Best Fit Line?

$y_i' = wx + b$

$y_i' = wx_i + b$

$y = ax + b$

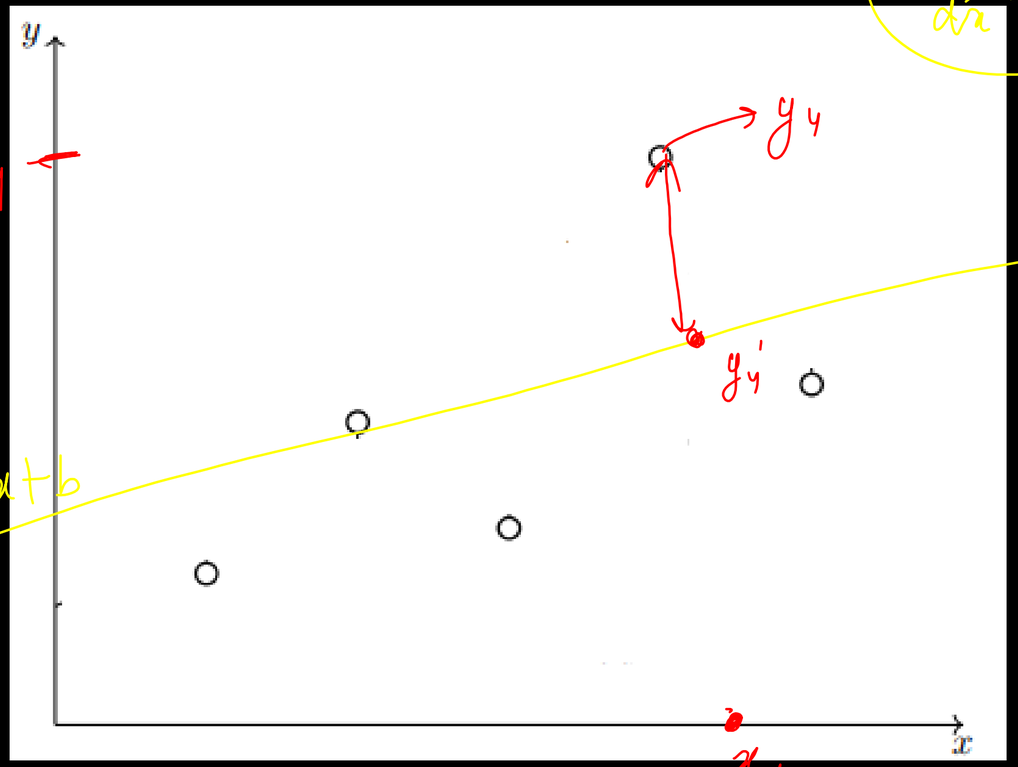


Best fit line tries to explain the variance in given data. (minimize the total residual/error)

$y = e^x$

$\frac{dy}{dx} = e^x = 0$

$y = wx + b$



$x_i \rightarrow y_i$

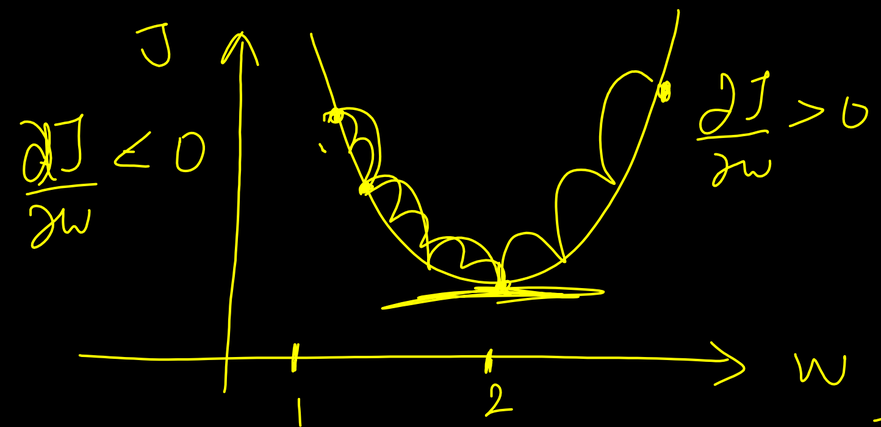
$\frac{1}{2n} \Sigma (y_i - y_i')$

$\downarrow$ Min

$J(w,b)$ Cost function $= \frac{1}{2n} \Sigma (y_i - (wx_i + b))^2$

Goal $\rightarrow w, b$

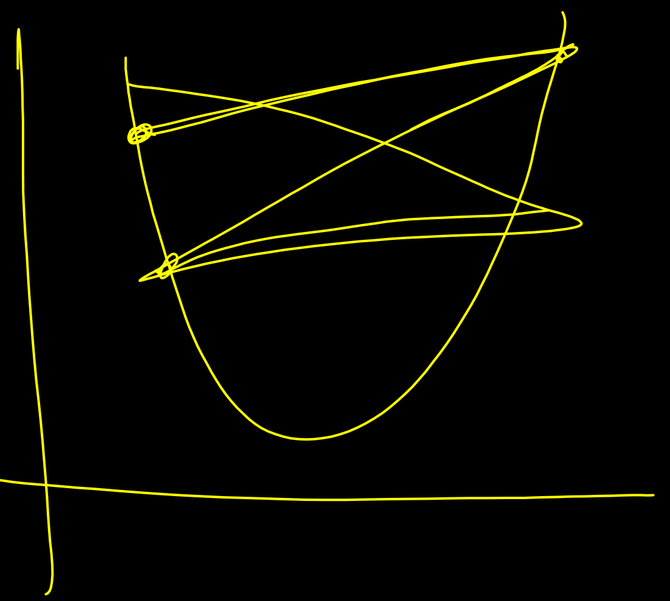
Gradient Descent



$\frac{\partial J}{\partial w} < 0$

$\frac{\partial J}{\partial w} > 0$

$$W = W - \alpha \frac{\partial J}{\partial w}$$



$$W = W - \alpha \frac{\partial J}{\partial w}$$

$$= 1 - \alpha\left(-\frac{2}{2}\right)$$

$$= 1 + 0.1 \times 2$$

$$= 1.2$$

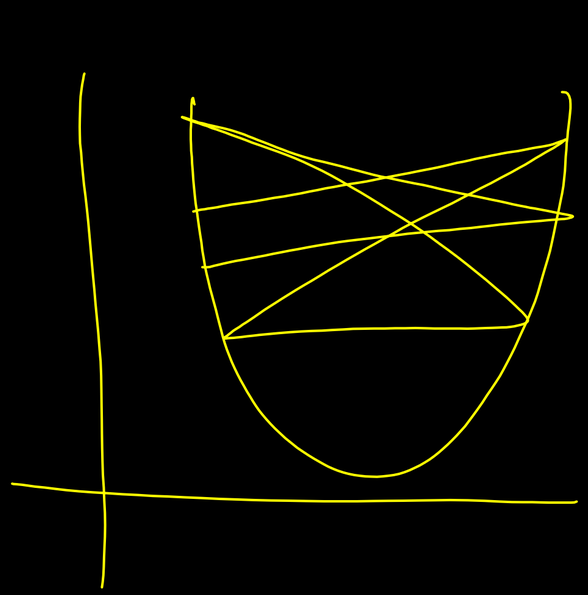
$$W = 1.2 - 0.1\,(-1.2)$$

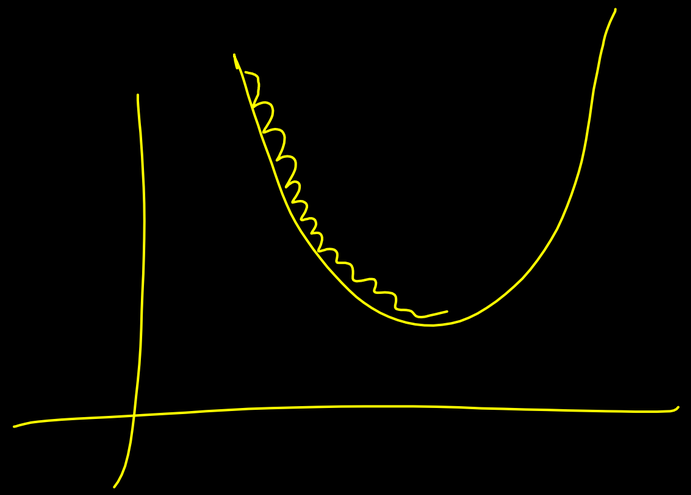
$$= 1.32$$

$$b = b - \alpha \frac{\partial J}{\partial b}$$

Learning Rate

$$w = w - \alpha \frac{\partial J}{\partial w}, \quad b = b - \alpha \frac{\partial J}{\partial b}$$





$\alpha \rightarrow$ too large.

$\alpha \rightarrow$ too small

Consider the following training set of m = 4 training examples

x	y
1	0.5
2	1
4	2
0	0

$y = \frac{1}{2}x + 0$

Consider the linear regression model y = wx + b. What are the values of w and b, that you would expect to obtain upon running gradient descent on this model?

(a) b = 0.5 and w = 0

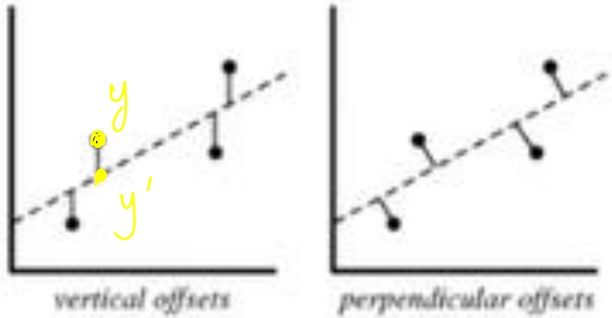
(b) b= 0.5 and w = 0.5

(c) b = 0 and w = 0.5

(d) b= 0 and w = 0

Answer: (c)

Which of the following offsets, do we use in linear regression's least square line fit? Assume the horizontal axis is the independent variable and vertical axis is dependent variable.



A) Vertical offset
B) Perpendicular offset
C) Both, depending on the situation
D) None of above

Which of the following if any is a valid cost function in a regression setting and why?

a. $J(w) = \frac{1}{2m}\sum_{i=1}^{m}(f_w(x^{(i)}) - y^{(i)})^2$

b. $J(w) = \frac{1}{2m}\sum_{i=1}^{m}(f_w(x^{(i)}) - y^{(i)})$

c. $J(w) = \frac{1}{2m}\sum_{i=1}^{m}|f_w(x^{(i)}) - y^{(i)}|$

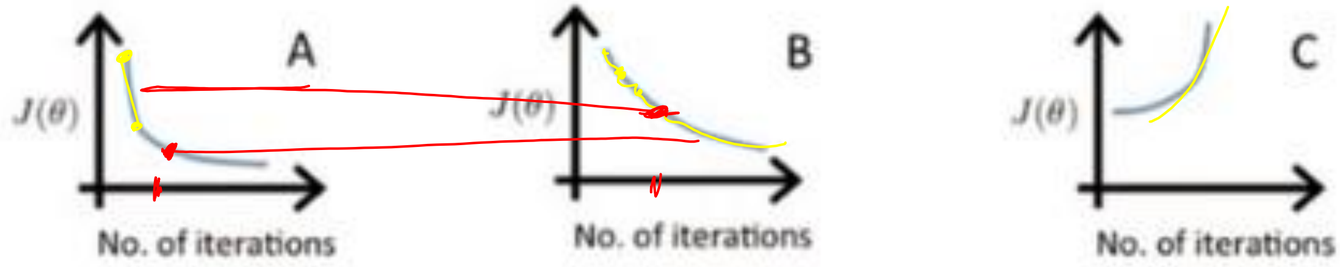
Which statement is true about outliers in Linear regression?

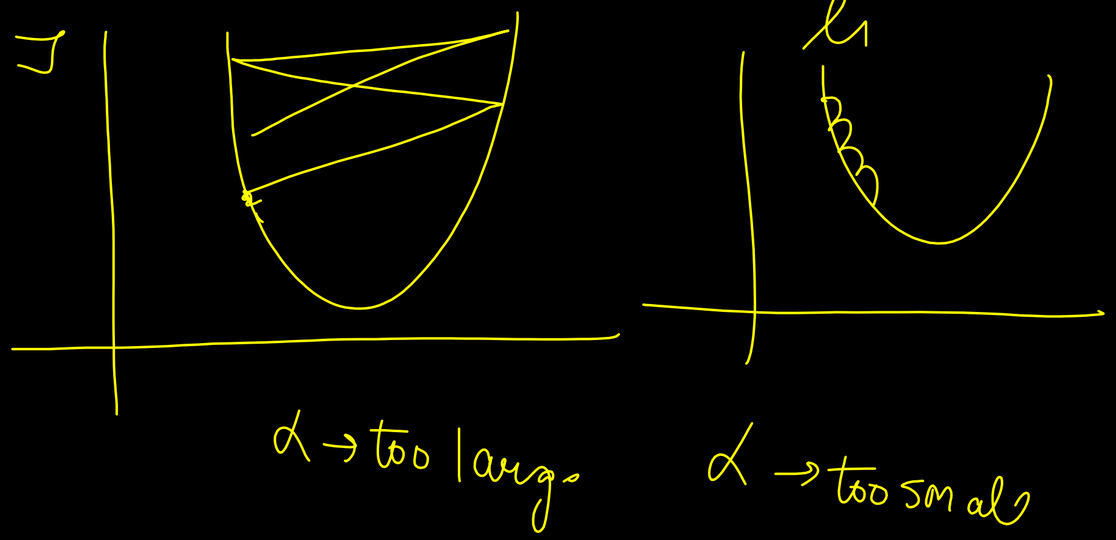
A) Linear regression is sensitive to outliers
B) Linear regression is not sensitive to outliers
C) Can't say
D) None of these

1, 2, 3, 4, 1000

Suppose l1, l2 and l3 are the three learning rates for A,B,C respectively. Which of the following is true about l1, l2 and l3?

A) l2 < l1 < l3
B) l1 > l2 > l3
C) l1 = l2 = l3
D) None of these



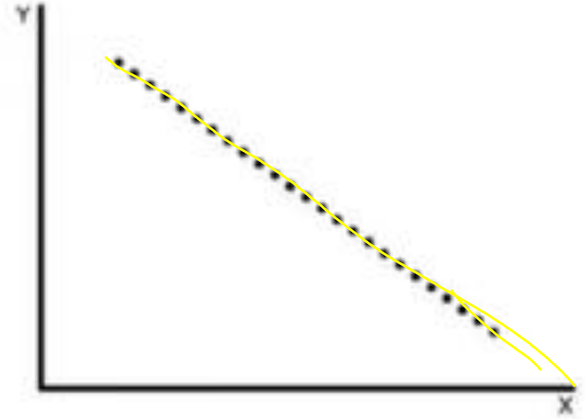


l_2 small

$l_2 < l_1 < l_3$

l_3 → too large

Consider the following data where one input(X) and one output(Y) is given. What would be the cost for this data if you run a Linear Regression model of the form (Y = w_1 * x_1 + b)?



A) Less than 0
B) Greater than zero
C) Equal to 0
D) None of these

The selling price of a house depends on the following factors. For example, it depends on the number of bedrooms, number of kitchen, number of bathrooms, the year the house was built and the square footage of the lot. Given these factors, predicting the selling price of the house is an example of _____ task.

a. Binary Classification
b. Multilabel Classification
c. Simple Linear Regression
d. Multiple Linear Regression

$x_1 \quad x_2 \quad \cdots \quad x_k \quad y$

$y = w_1x_1 + w_2x_2 + \cdots + w_kx_k + b$

Multiple regression

Allows a response variable Y to be modeled as a linear function of multidimensional feature vector

x_1 no of room, x_2 location, x_3 size, y price

$y' = w_1x_1 + w_2x_2 + w_3x_3 + b$ ⟶ Goal → w_1, w_2, w_3, b

$y' = w^T x$

$y' = w_0x_0 + w_1x_1 + w_2x_2 + w_3x_3$

$= [w_0\ w_1\ w_2\ w_3] \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_3 \end{bmatrix}$

$= w^T X$

$w_i = w_i - \alpha \frac{\partial J}{\partial w_i}$

$b = b - \alpha \frac{\partial J}{\partial b}$

$J = \frac{1}{2n} \sum_{i=1}^{n} (y_i - y')^2$

In the context of machine, **sparsity** refers to a model where many of the feature weights (coefficients) are exactly zero. In simple terms:

Sparse models are models where only a small number of features (input variables) have non-zero coefficients, meaning that the model effectively ignores most of the features.

$w_1, w_2, \ldots, w_k \rightarrow w_1 = 0, w_2 = 0, w_3 \neq 0, w_4 \neq 0 \ldots$

Dense models, on the other hand, use most or all of the features with non-zero coefficients.

$w_1 \neq 0, w_2 \neq 0 \ldots$

$\downarrow$

$w_1 \approx 0$

0.0001

L1 Regularization (Lasso)

$$\text{L1 Regularization Term} = \lambda \sum_{i=1}^{n} |w_i|$$

hyperparameter

$$y' = w_1 x_1 + w_2 x_2 + + w_3 x_3 + b$$

$$\text{Error} = \frac{1}{2n}(y - y') + \lambda \sum |w_1|$$

Min

w_1, b

$\lambda = 1000$

$1000 \times$

$= 1000$

$w_1 \approx 0$

Sparse

1

L2 Regularization (Ridge):

$$\text{L2 Regularization Term} = \lambda \sum_{i=1}^{n} w_i^2$$

$$\text{Error} = \frac{1}{2n} \sum (y - y_i')^2 + \lambda \sum_{i=1}^{n} w_i^2$$

$w_1 \neq 0$

$w_1 = 0.0001$

$$y = 100x_1 + 200x_2 + 300x_3 + 6$$

$w_1 \quad w_2 \quad w_3$

Combined L1 and L2 Regularization (Elastic Net):

$$\text{Elastic Net Regularization Term} = \lambda_1 \sum_{i=1}^{n} |w_i| + \lambda_2 \sum_{i=1}^{n} w_i^2$$

Here, λ_1 and λ_2 control the strengths of L1 and L2 regularization, respectively.

Select all the following reasons why we would use LASSO over Ridge?

A. It can help us identify which features are important
B. It is faster to learn the weights for LASSO than for Ridge
C. LASSO usually achieves lower generalization error than Ridge
D. If there are many features, the model learned using LASSO can make predictions more efficient

Answer: (A), (C)

With Lasso Regression the influence of the hyper parameter lambda, as lambda tends to zero the solution approaches to _____ .

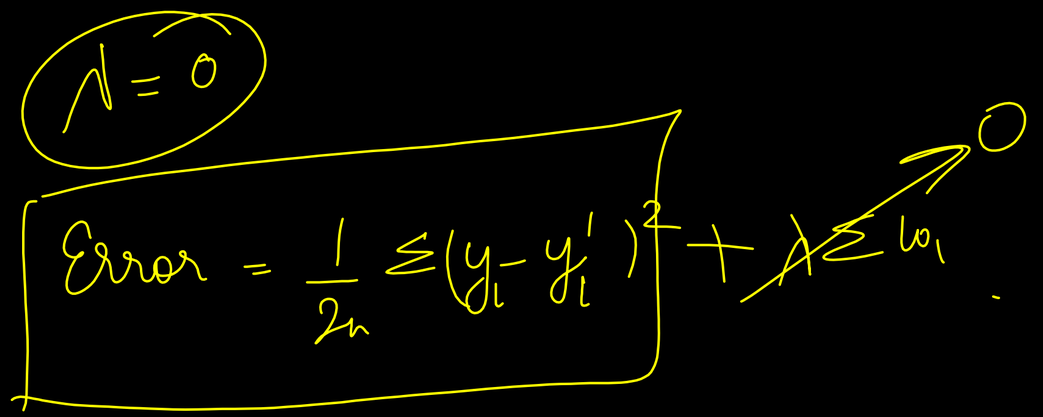
a) Zero.

b) One.

c) Linear regression.

d) Infinity.

Answer: (c)



In this Lasso and Ridge regression as alpha value increases, the impact on slope is

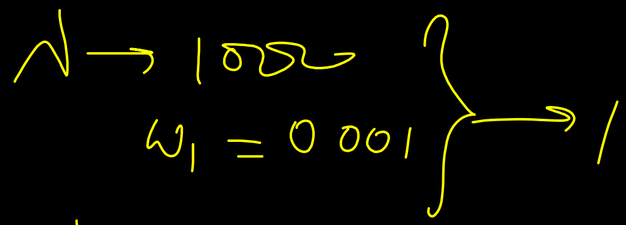
a) Slope is fixed for whatever maybe the value.

b) the slope of the regression line reduces and becomes horizontal.

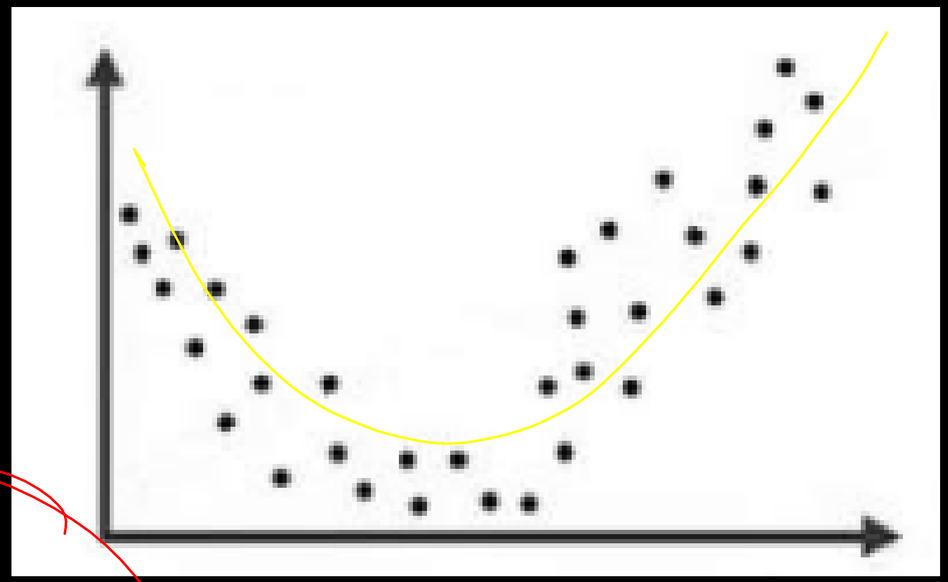
c) the slope of the regression line increases and becomes vertical

d) None of the above.

Answer: (b)



Polynomial regression

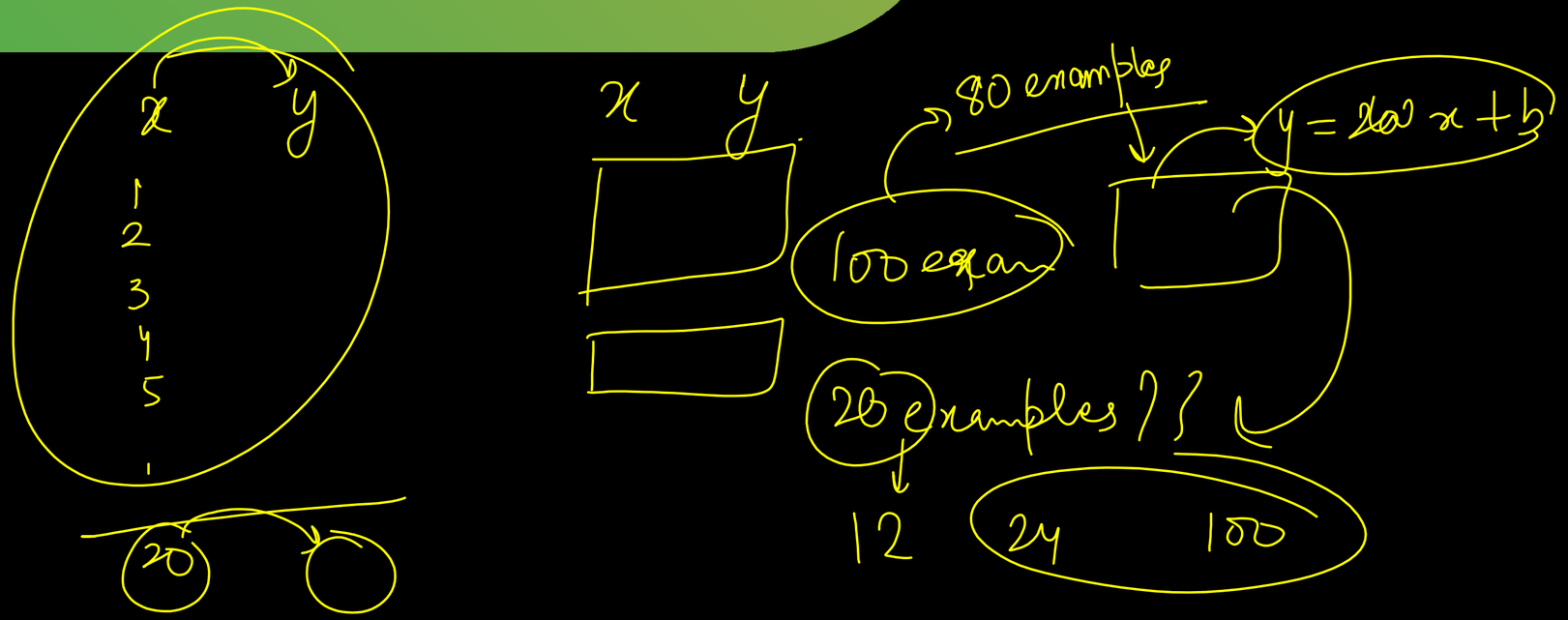


$y = w_1 x + w_2 x^2 + b$

$y = w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 + w_5 x^5 + w_6 x^6 + w_7 x^7 + b$

$y = wx + b$

Bias Variance Tradeoff



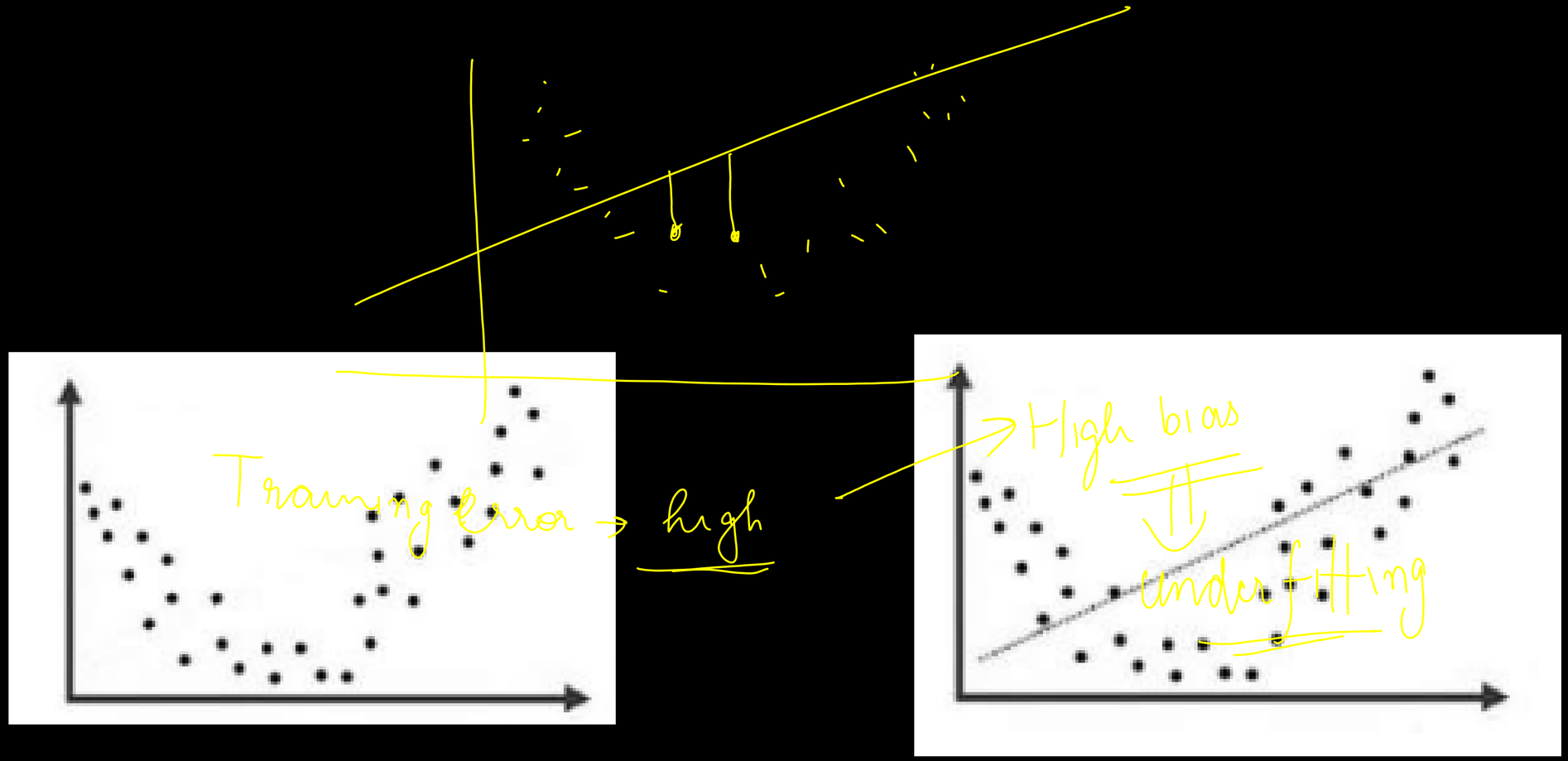
Training & Test Data

In machine learning, data is split into training data and test data.

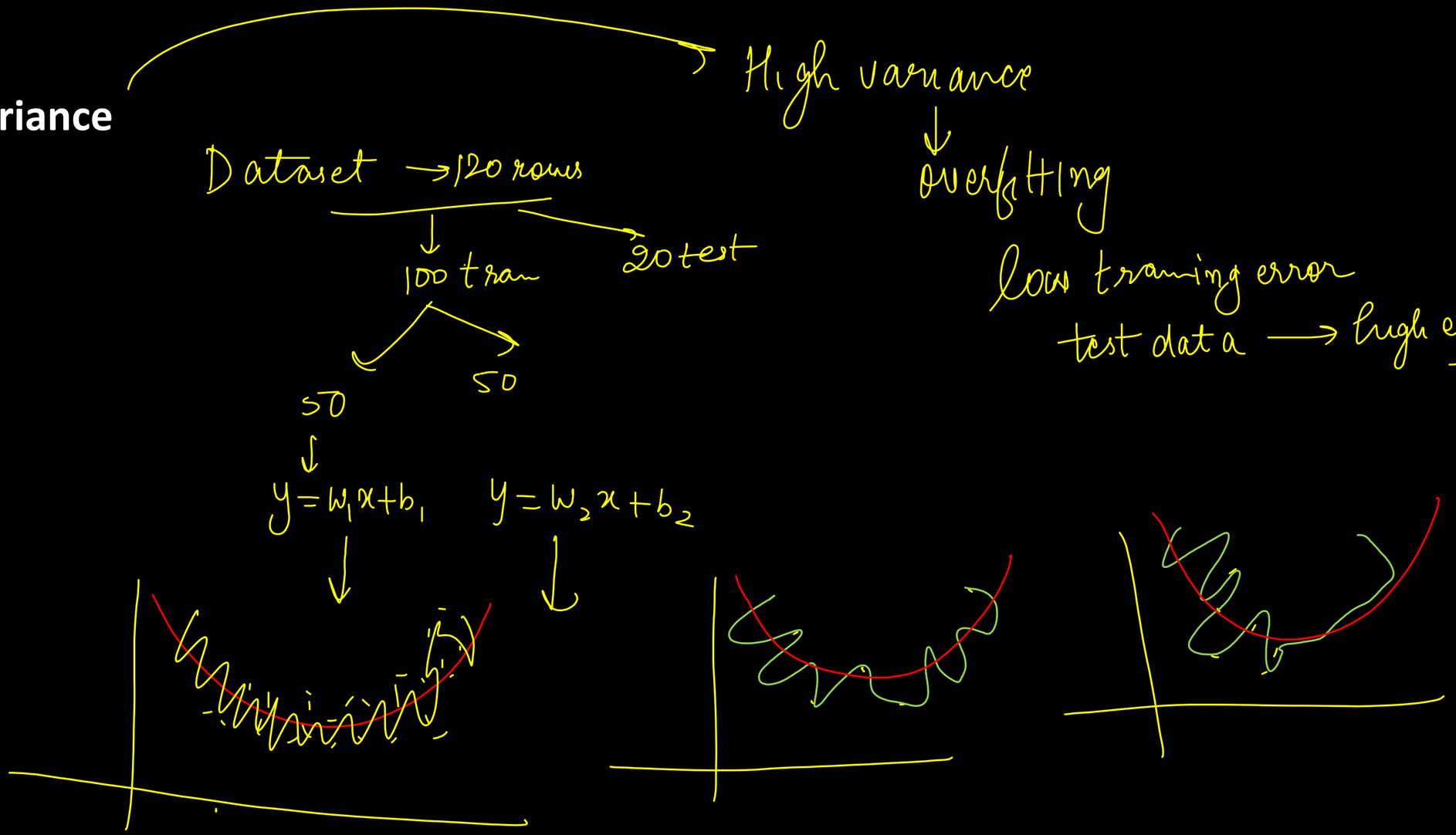
The first split of data, i.e. the initial reserve of data you use to develop your model, provides the training data.

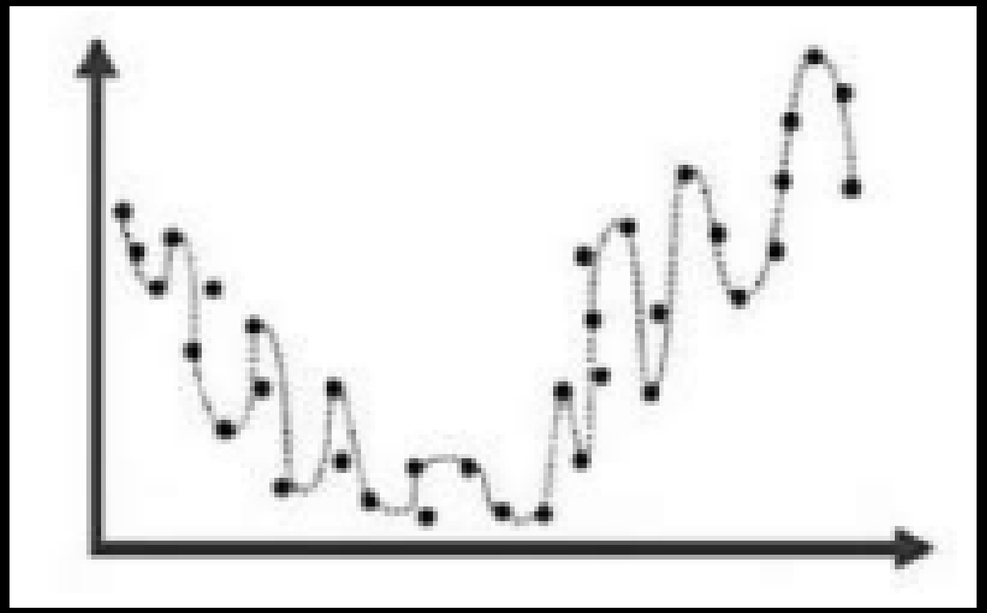
After you have successfully developed a model based on the training data and are satisfied with its accuracy, you can then test the model on the remaining data, known as the test data.

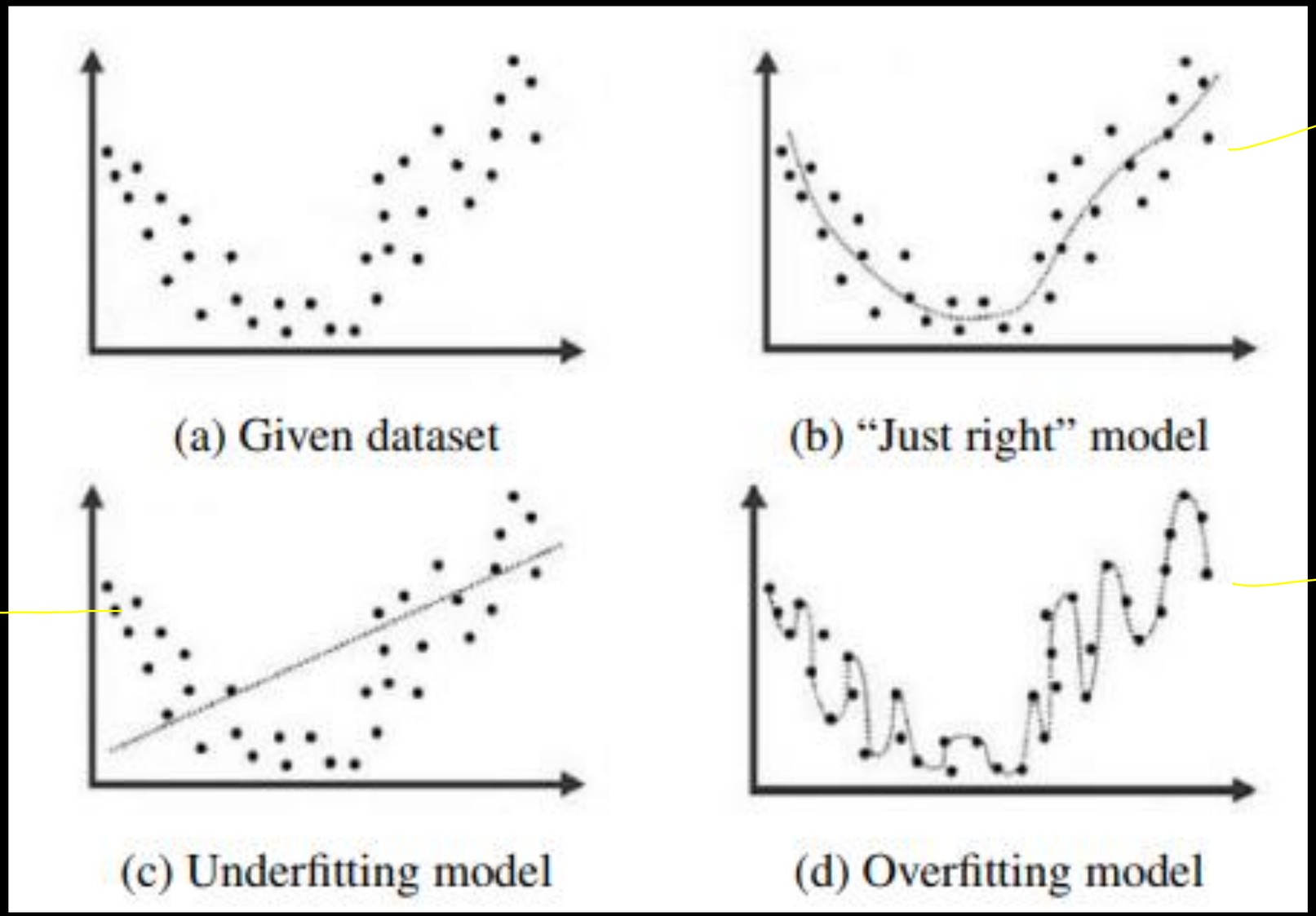
Bias

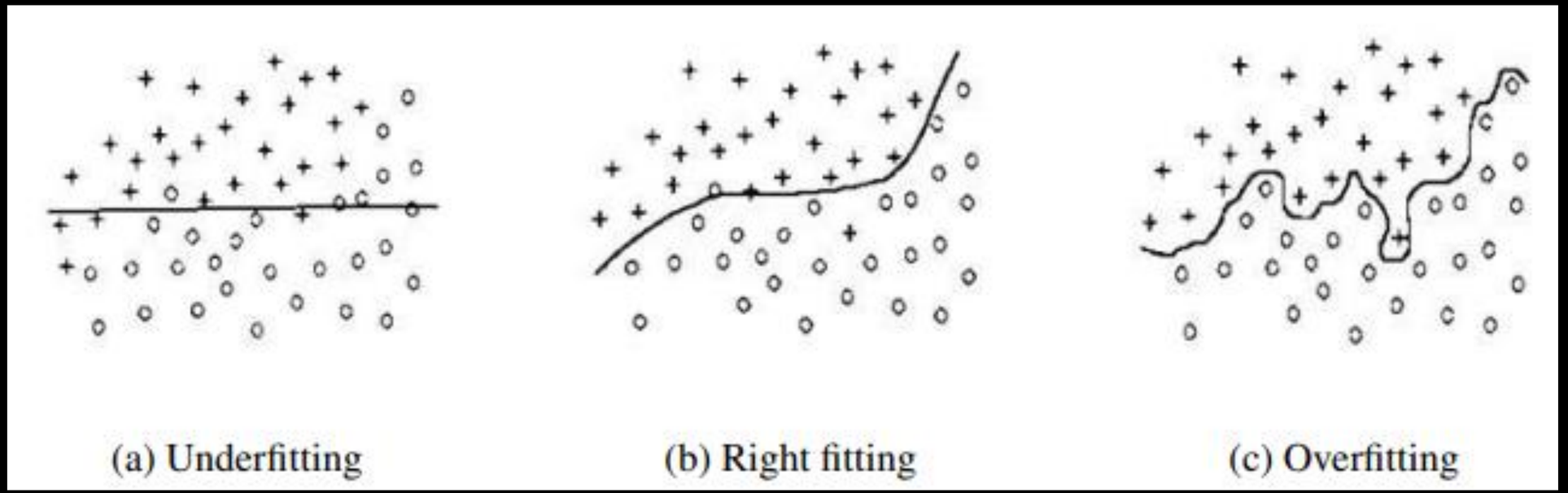


Variance


High variance
Dataset → 120 rows
100 train
20 test
50
50
$y = w_1 x + b_1$
$y = w_2 x + b_2$
Overfitting
low training error
test data → high error




(a) Given dataset
(b) "Just right" model
(c) Underfitting model
(d) Overfitting model

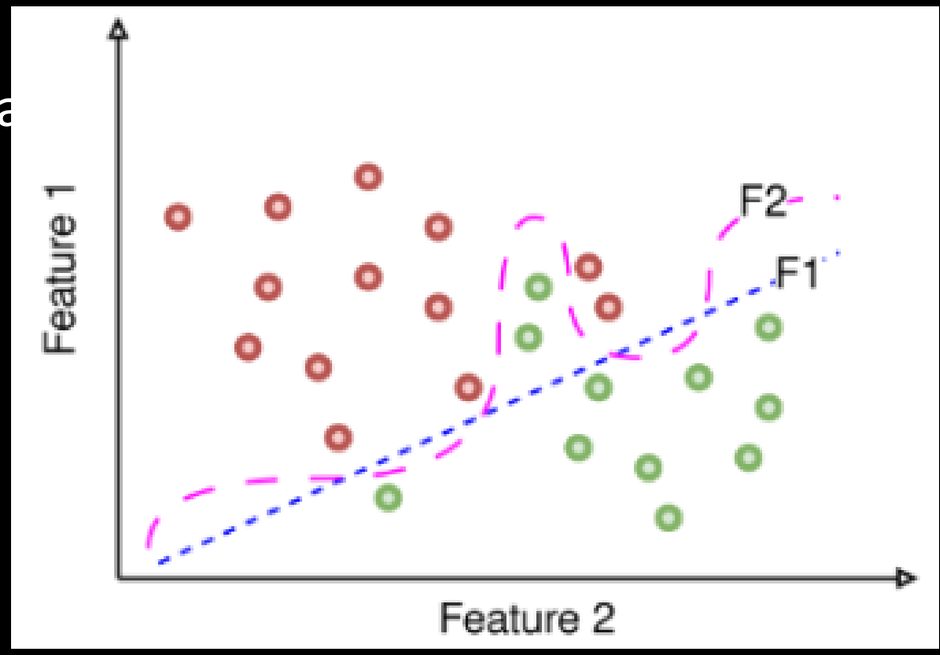

(a) Underfitting
(b) Right fitting
(c) Overfitting

Tradeoff:

•The bias-variance tradeoff suggests that there is a balance to be struck between bias and variance. As you decrease bias (by increasing model complexity), you typically increase variance, and vice versa.

•The goal is to find the right level of model complexity that minimizes the combined error due to bias and variance. The objective is to achieve a model that generalizes well to new, unseen data.

Here is a 2-dimensional plot showing two functions that classify data points into two classes. The red points belong to one class, and the green points belong to another. The dotted blue line (F1) and dashed pink line (F2) represent the two trained functions.



Which of the two functions overfit the training data

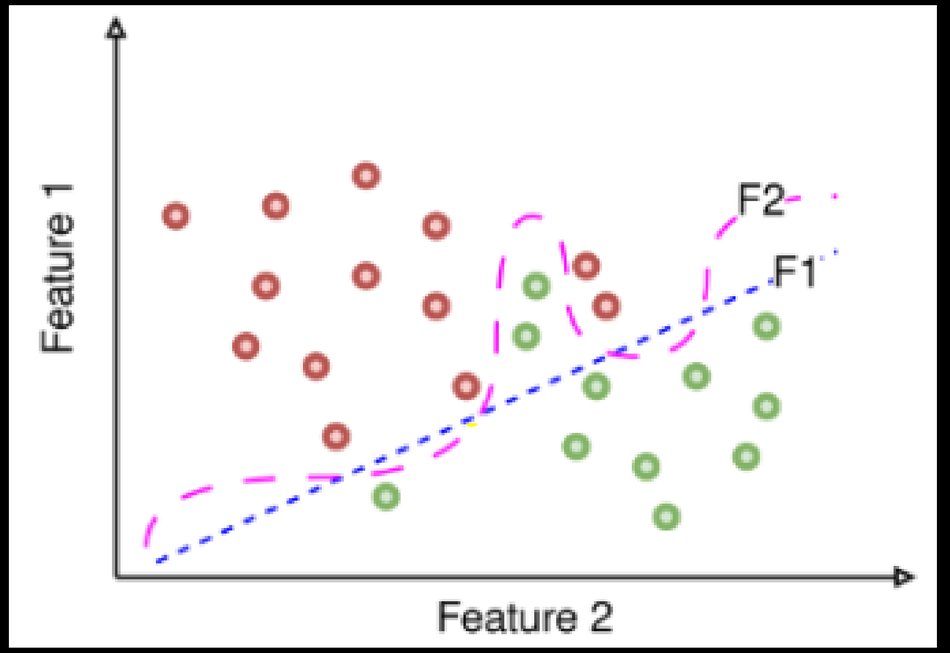
A. Both functions F1 & F2
B. Function F1
C. Function F2
D. None of them

Answer: (C)

Which of the following 2 functions will yield higher training error?

A. Function F1
B. Function F2
C. Both functions F1 & F2 will have the same training error
D. Cannot be determined

Answer: (A)



Which of the following techniques can help reduce overfitting in a machine learning model?

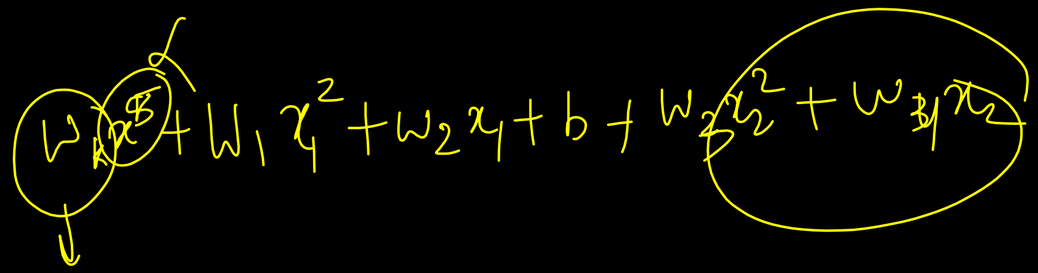
a) Increasing the model complexity

b) Decreasing the amount of training data

c) Adding more features to the model

d) Applying regularization techniques

Answer: (d)



After training a model, you observe a significant gap between the training and test performance metrics. Which of the following techniques is most likely to reduce this gap?

a) Increasing the training dataset size
b) Decreasing the model complexity
c) Adding more features to the model
d) Fine-tuning hyperparameters

Answer: (b),(d)

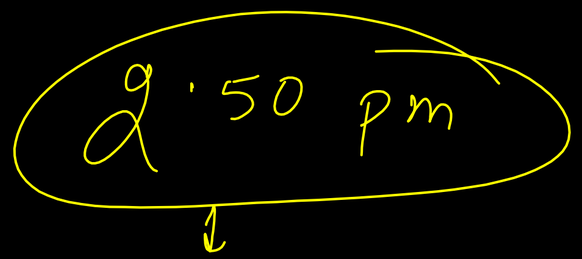
Suppose, you got a situation where you find that your linear regression model is under fitting the data. In such situation which of the following options would you consider?

a. You will add more features
b. You will start introducing higher degree features
c. You will remove some features
d. None of the above.

You have generated data from a 3-degree polynomial with some noise. What do you expect of the model that was trained on this data using a 5-degree polynomial as function class?

a. Low bias, high variance
b. High bias, low variance.
c. Low bias, low variance.
d. High bias, low variance.

Cross Validation



Types of Cross Validation Set

1. Leave-One-Out Cross-Validation (LOOCV)
2. Hold-out cross-validation
3. K-Fold Cross-Validation
4. Stratified K-Fold Cross-Validation
5. Time Series Cross-Validation

How cross validation helps in bias variance tradeoff?

Bias Estimation:
By performing cross-validation, If the average performance is consistently poor across all folds, it may indicate that the model has high bias.

Cross-validation helps in identifying underfitting by revealing consistent errors in different portions of the dataset.

Variance Estimation:
If there is a significant difference in performance between folds, it suggests that the model is sensitive to the specific data used for training and testing, indicating potential overfitting.

Hyperparameter Tuning:
Cross-validation is commonly used for hyperparameter tuning.

Models with different hyperparameter configurations are trained and evaluated on multiple folds.
This helps in selecting hyperparameter values that balance bias and variance, optimizing the model for better generalization.

Cross validation is a model evaluation method. **Leave-one-out cross validation**(LOOCV) is K-fold cross validation taken to its logical extreme, with K equal to N, the number of data points in the set. That means that N separate times, the function approximator is trained on all the data except for one point and a prediction is made for that point. Thus, it iterates over the other datapoints keeping the rest of the dataset fixed. What can be the major issues in LOOCV?

a. low variance
b. high variance
c. faster run time compared to K-fold cross validation
d. slower run time compared to K-fold cross validation

Which of the following cross validation strategies cannot be stratified?

a) k-fold cross validation

b) hold out cross validation

c) leave one out cross validation

d) shuffle split cross validation

Which of the following cross validation versions may not be suitable for very large datasets with hundreds of thousands of samples?

a) k-fold cross-validation

b) Leave-one-out cross-validation

c) Holdout method

d) All of the above

As k increases in k-fold cross-validation method?

a) The variance of the resulting estimate is reduced as k is increased.

b) The variance of the resulting estimate is reduced as k is decreased.

c) None of the above

Which of the following is a disadvantage of k-fold cross-validation method?

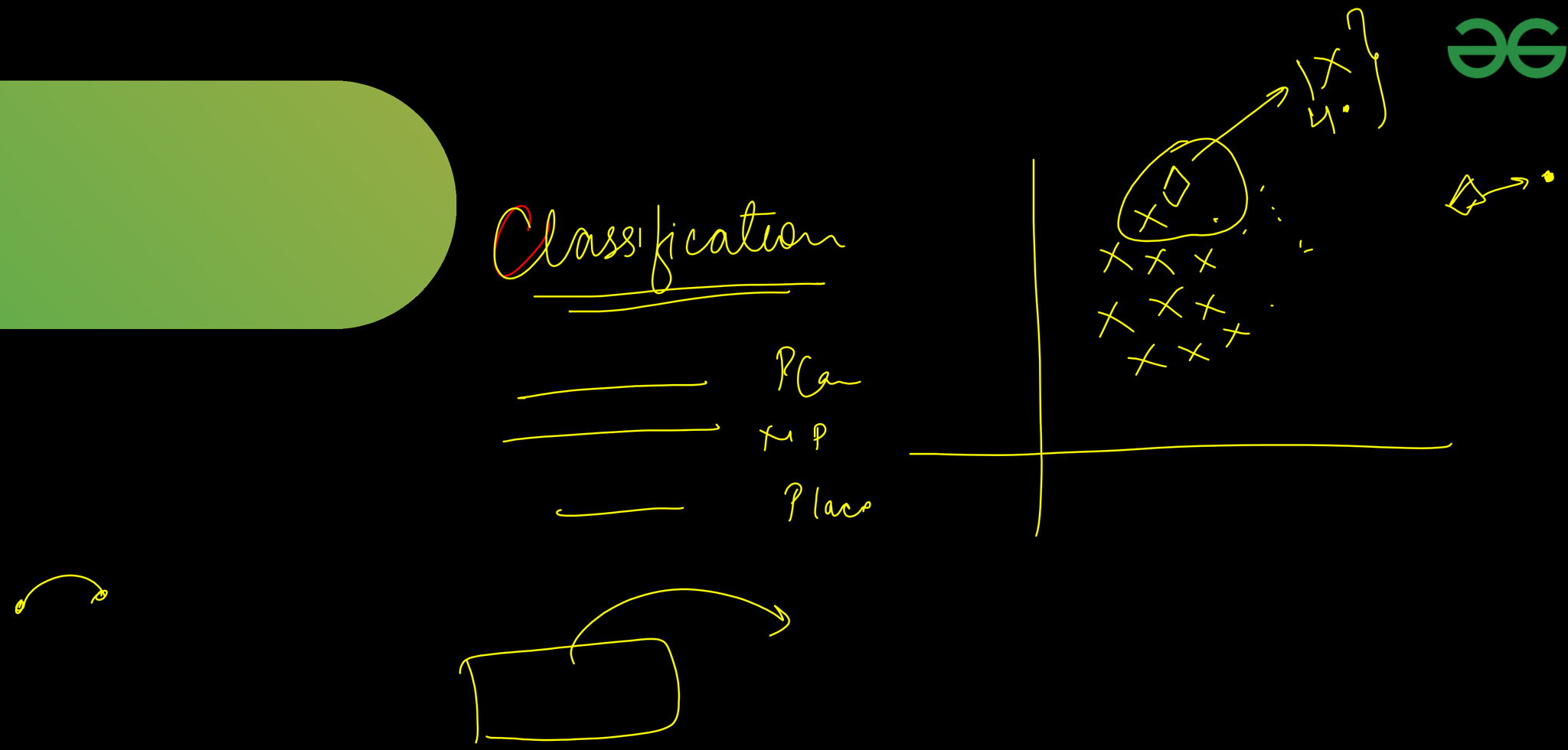
a) The variance of the resulting estimate is reduced as k is increased.

b) This usually does not take longer time to compute

c) Reduced bias

d) The training algorithm has to rerun from scratch k times

KNN



KNN → K=3

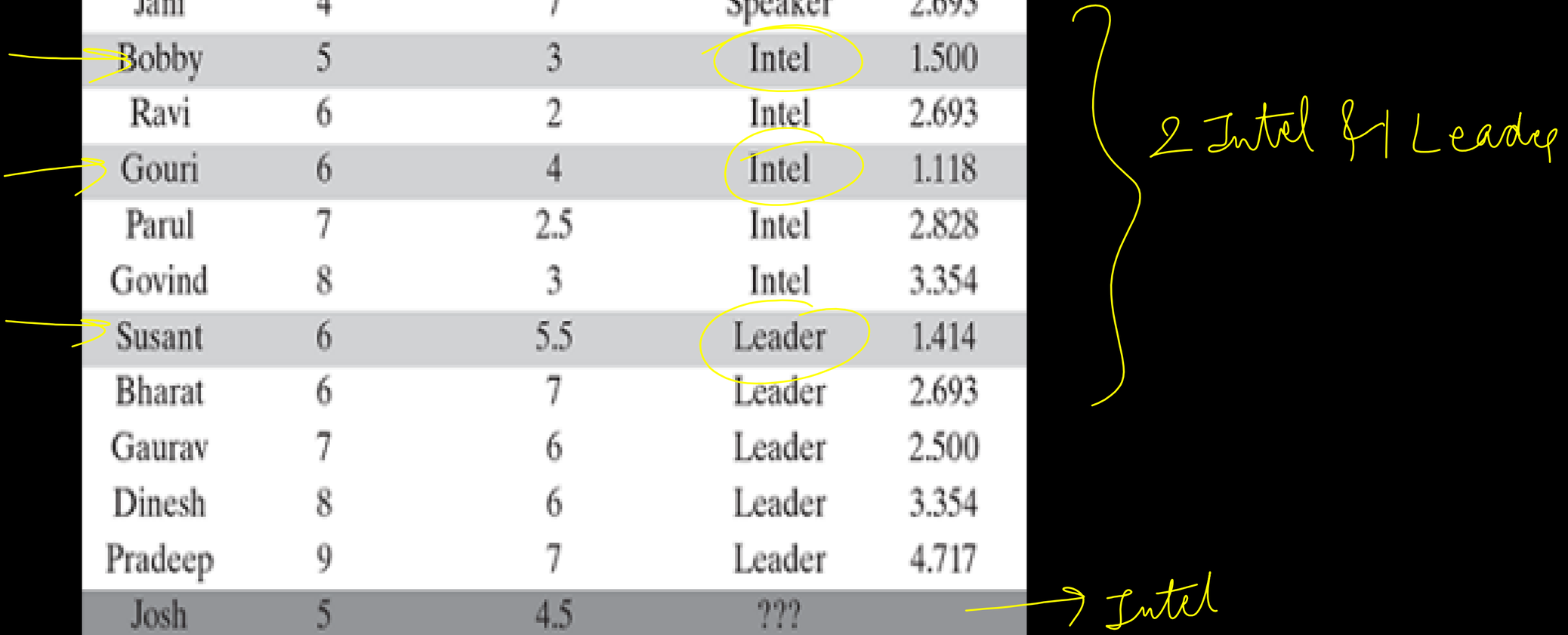
Name	Aptitude	Communication	Class
Karuna	2	5	Speaker
Bhuvna	2	6	Speaker
Gaurav	7	6	Leader
Parul	7	2.5	Intel
Dinesh	8	6	Leader
Jani	4	7	Speaker
Bobby	5	3	Intel
Parimal	3	5.5	Speaker
Govind	8	3	Intel
Susant	6	5.5	Leader
Gouri	6	4	Intel
Bharat	6	7	Leader
Ravi	6	2	Intel
Pradeep	9	7	Leader
Josh	5	4.5	???

$\sqrt{(5-2)^2+(4.5-5)^2}$ =

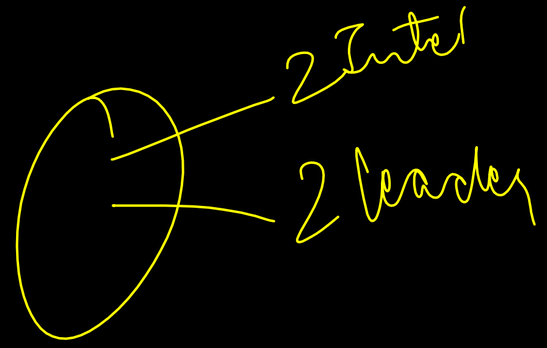
(x_1, y_1) & (x_2, y_2)

$\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$

Name	Aptitude	Communication	Class	Distance
Karuna	2	5	Speaker	3.041
Bhuvna	2	6	Speaker	3.354
Parimal	3	5.5	Speaker	2.236
Jani	4	7	Speaker	2.693
Bobby	5	3	Intel	1.500
Ravi	6	2	Intel	2.693
Gouri	6	4	Intel	1.118
Parul	7	2.5	Intel	2.828
Govind	8	3	Intel	3.354
Susant	6	5.5	Leader	1.414
Bharat	6	7	Leader	2.693
Gaurav	7	6	Leader	2.500
Dinesh	8	6	Leader	3.354
Pradeep	9	7	Leader	4.717
Josh	5	4.5	???	

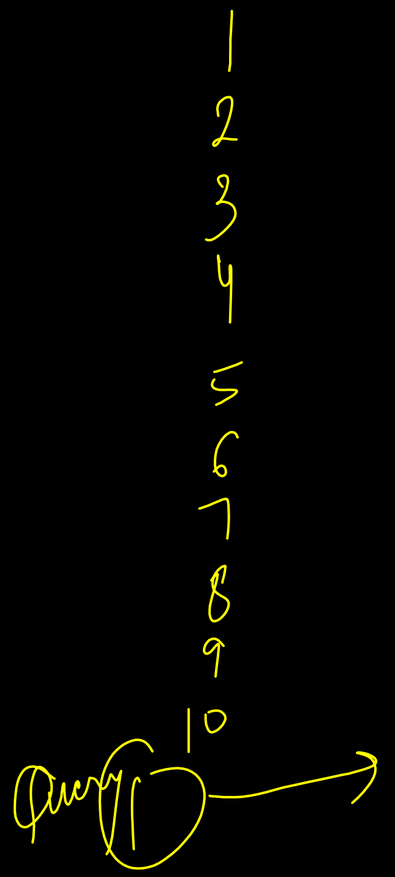


Why is the odd value of ‘k’ preferred over an even value in the k-NN algorithm?



K-Nearest Neighbor is a _____, _____ algorithm

a. Non-parametric, eager
b. Parametric, eager
c. Non-parametric, lazy
d. Parametric, lazy



State whether the statement is True/False:

k-NN algorithm does more computation on test time rather than train time.

1. True
2. False

Consider a dataset with the following three data points in a two-dimensional space:

- Data point A: (2, 3), Class: 1
- Data point B: (4, 6), Class: 1
- Data point C: (5, 2), Class: 2

If K=2, what is the predicted class of a new data point at coordinates (3, 4) using the Euclidean distance metric?

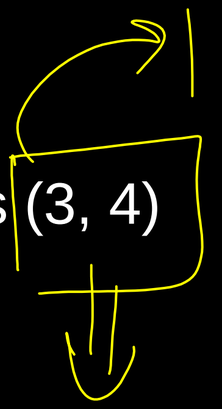
$(3,4)$ & $A = \sqrt{(3-2)^2+(4-3)^2} = \sqrt{2}$

$(3,4)$ & $B = \sqrt{(3-4)^2+(4-6)^2} = \sqrt{5}$

$(3,4)$ & $C = \sqrt{(3-5)^2+(4-2)^2} = \sqrt{8}$

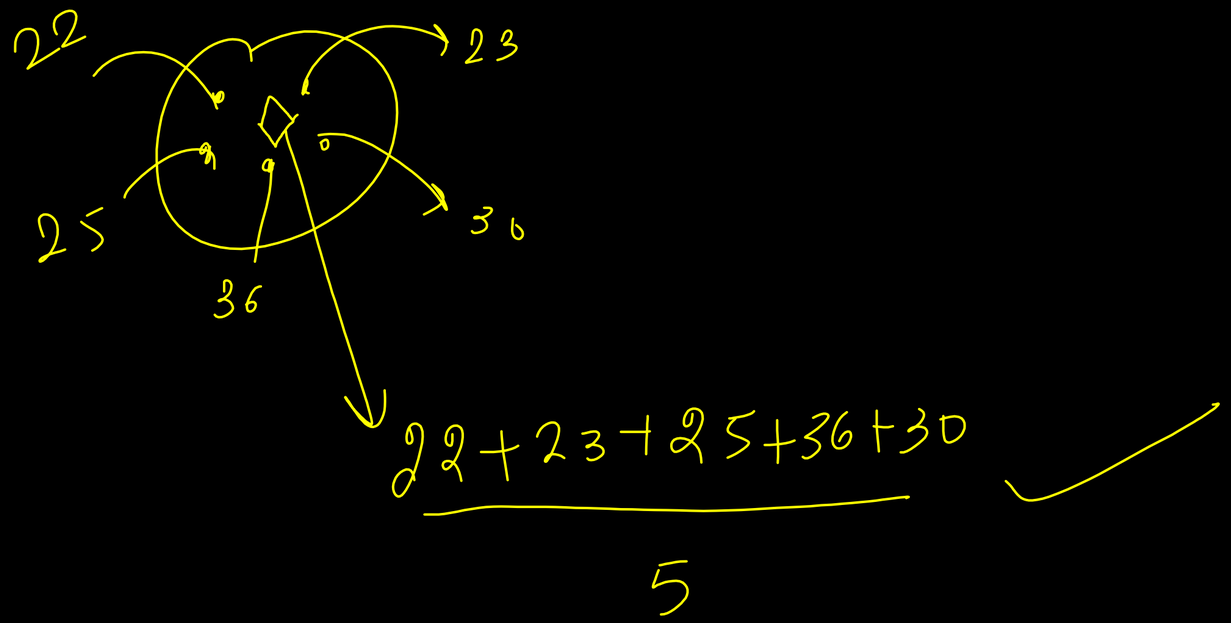
A & B

1



Can k-NN algorithm be used for a regression problem?

Yes



What would be the relationship between the training time taken by 1- NN, 2-NN, and 3-NN?

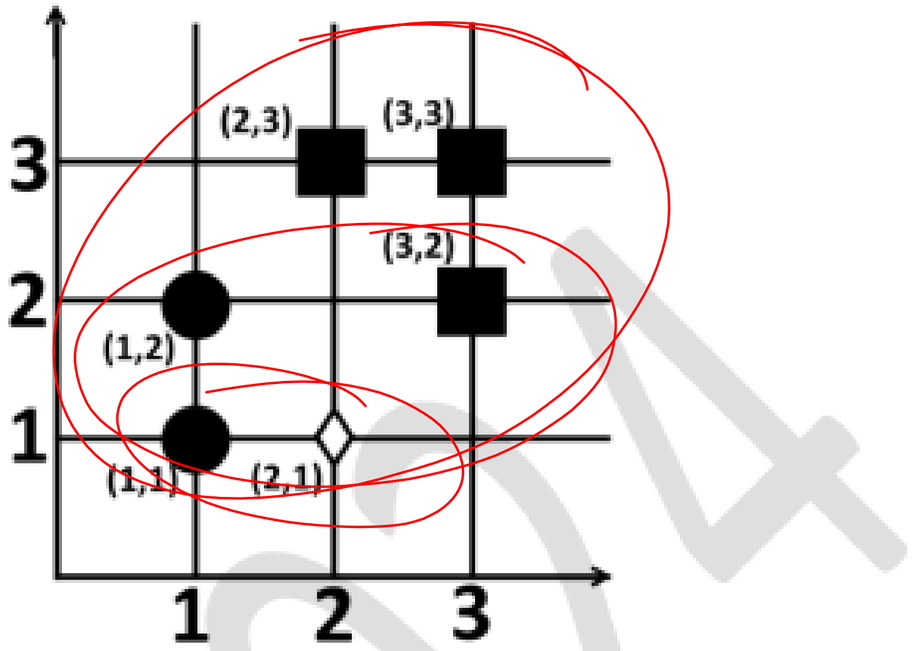
1. 1-NN > 2-NN > 3-NN
2. 1-NN < 2-NN < 3-NN
3. 1-NN ~ 2-NN ~ 3-NN
4. None of these

Given the two-dimensional dataset consisting of 5 data points from two classes (circles and squares) and assume that the Euclidean distance is used to measure the distance between two points. The minimum odd value of k in k-nearest neighbor algorithm for which the diamond (◇) shaped data point is assigned the label square is _____.

k=1

k=3

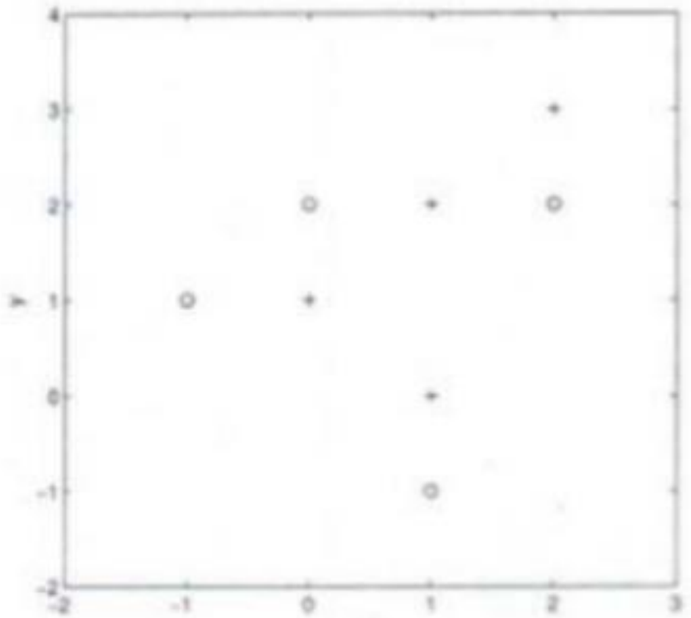
k=5



Suppose, you have given the following data where x and y are the 2 input variables and Class is the dependent variable.

x	y	Class
-1	1	$-$
0	1	$+$
0	2	$-$
1	-1	$-$
1	0	$+$
1	2	$+$
2	2	$-$
2	3	$+$

Below is a scatter plot which shows the above data in 2D space.



Suppose, you want to predict the class of new data point x=1 and y=1 using Eucledian distance in 3-NN. In which class this data point belong to?

a. + Class
b. – Class
c. Can't Say
d. None of these

Consider a set of five training examples given as $((x_i,y_i), c_i)$ values, where x_i and y_i are the two attribute values (positive integers) and c_i is the binary class label:

((1,1,),-1) → 2+5 = 7

((1,7),+1) → 2+1 = 3

((3,3),+1) → 0+3 = 3

((5,4),-1) → 2+2 = 4

((2,5),-1) → 1+1 = 2

(x_1,y_1) (x_2,y_2)

$|y_2-y_1| + |x_2-x_1|$

Classify a test example at coordinates (3,6) using a k-NN classifier with k= 3 and Manhattan distance defined by d((u,v),(p,q)) = |u-p| + |v-q| → +1

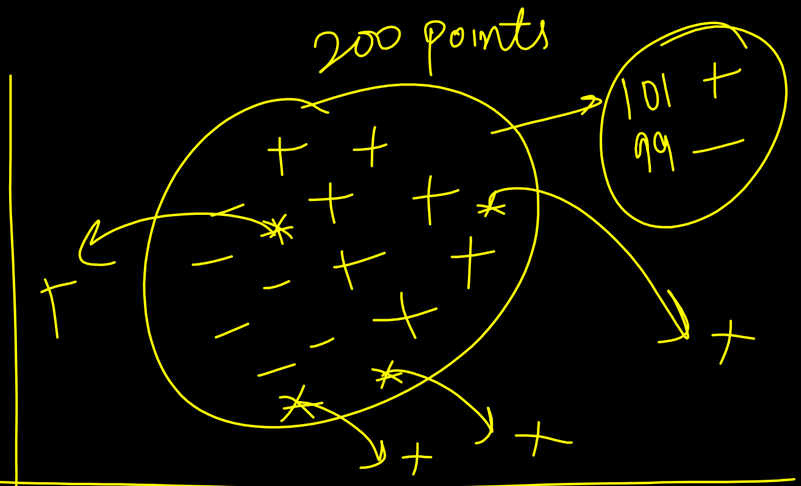
Large value of K

↓

underfitting

↓

High bias



Small value of K

↓

Overfitting

↓

high variance & low bias

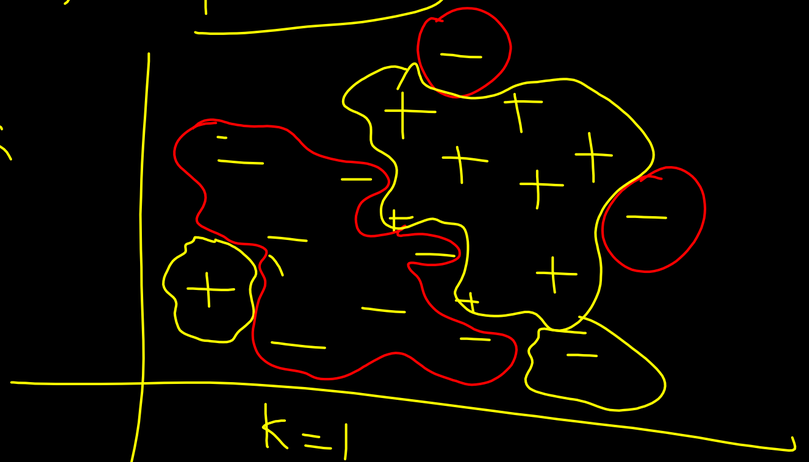
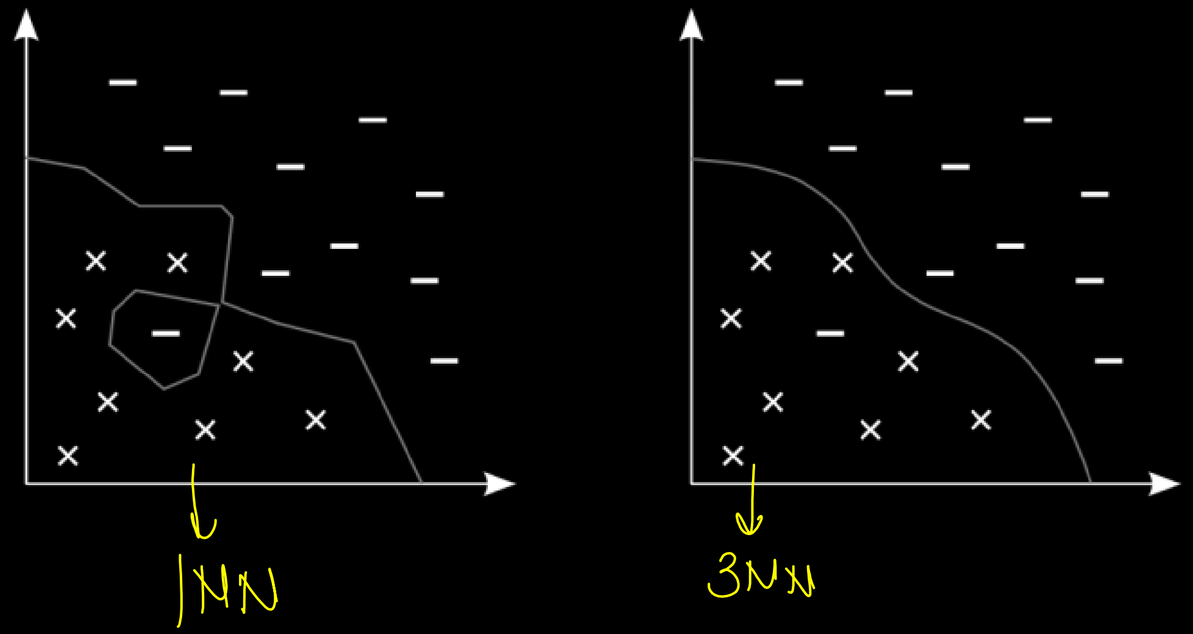


Figure illustrates decision boundaries for two nearest – neighbor classifiers. Determine which one of the boundaries belongs to the 1-NN and which one belong to 3-NN?



You have been given the following 2 statements. Find out which of these options is/are true in case of k-NN?

1. In case of very large value of k, we may include points from other classes into the neighborhood.
2. In case of too small value of k, the algorithm is very sensitive to noise.

a. 1 is True and 2 is False
b. 1 is False and 2 is True
c. Both are True
d. Both are False

What is the effect of increasing the value of K in KNN on the bias and variance of the model?

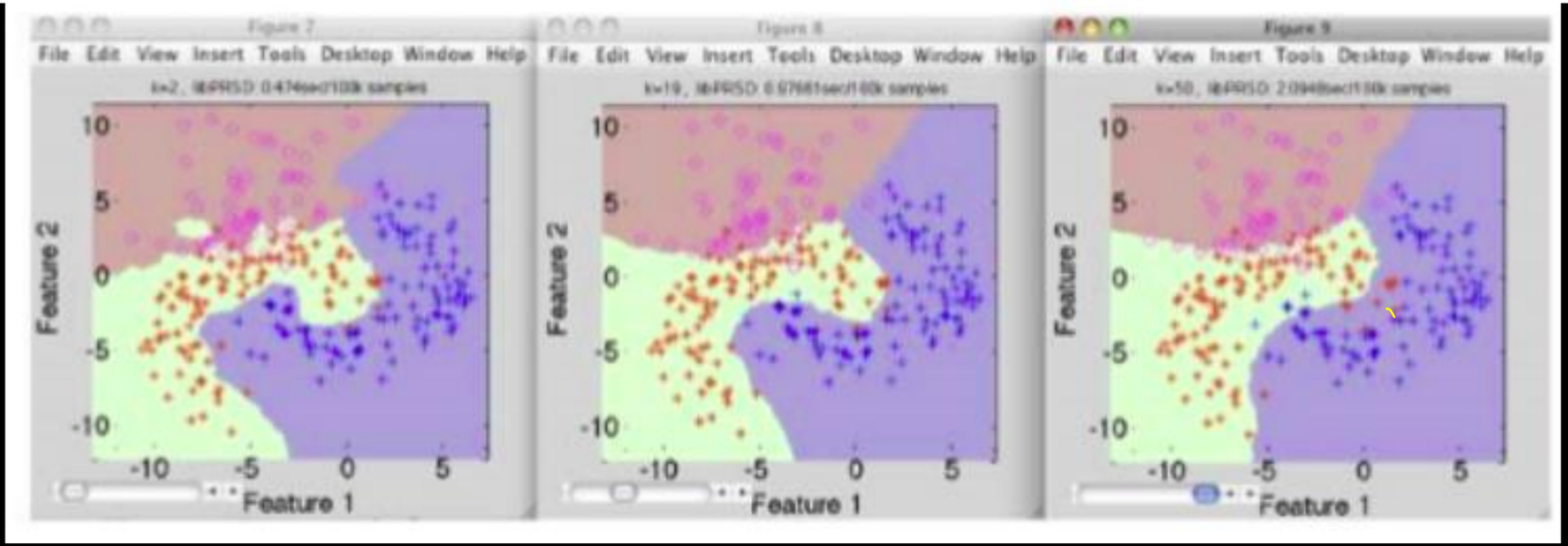
a) Increasing K increases bias and decreases variance.
b) Increasing K decreases bias and increases variance.
c) Increasing K has no effect on bias or variance.
d) Increasing K decreases both bias and variance.

Overfitting ⟶ underfitting

When you find many noises in data, which of the following options would you consider in kNN?

1. Increase the value of k
2. Decrease the value of k
3. Noise does not depend on k
4. K = 0

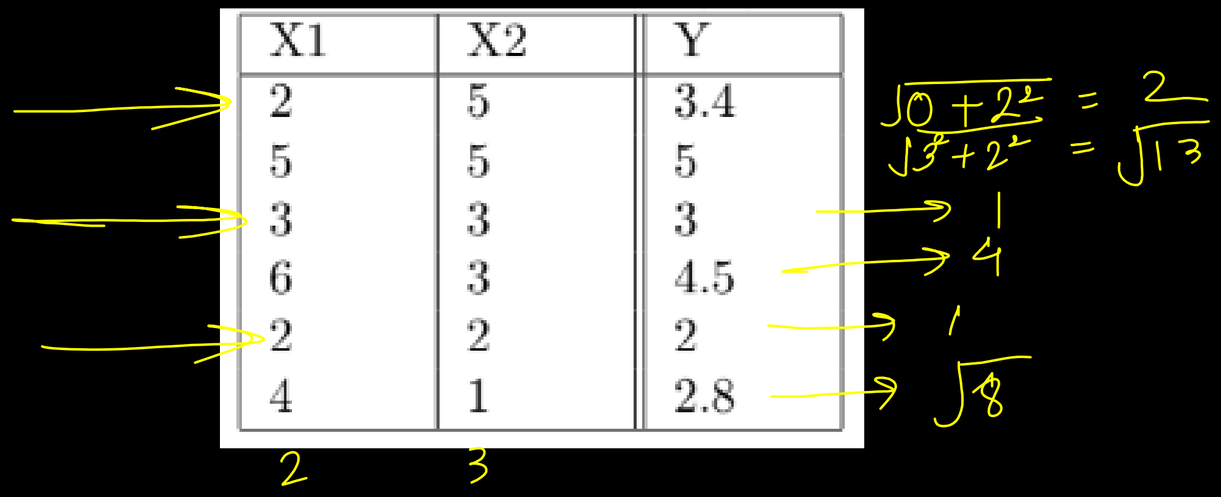
Suppose you are given the following images (1 represents the left image, 2 represents the middle and 3 represents the right). Now your task is to find out the value of k in k-NN in each of the images shown below. Here k1 is for 1st, k2 is for 2nd and k3 is for 3rd figure.



a. k1 > k2> k3
b. k1 < k2> k3
c. k1 < k2 < k3
d. None of these

Given the following dataset, for k = 3, use KNN regression to find the prediction for a new data-point (2,3) *(Use Euclidean distance measure for finding closest points)*

X1	X2	Y
2	5	3.4
5	5	5
3	3	3
6	3	4.5
2	2	2
4	1	2.8



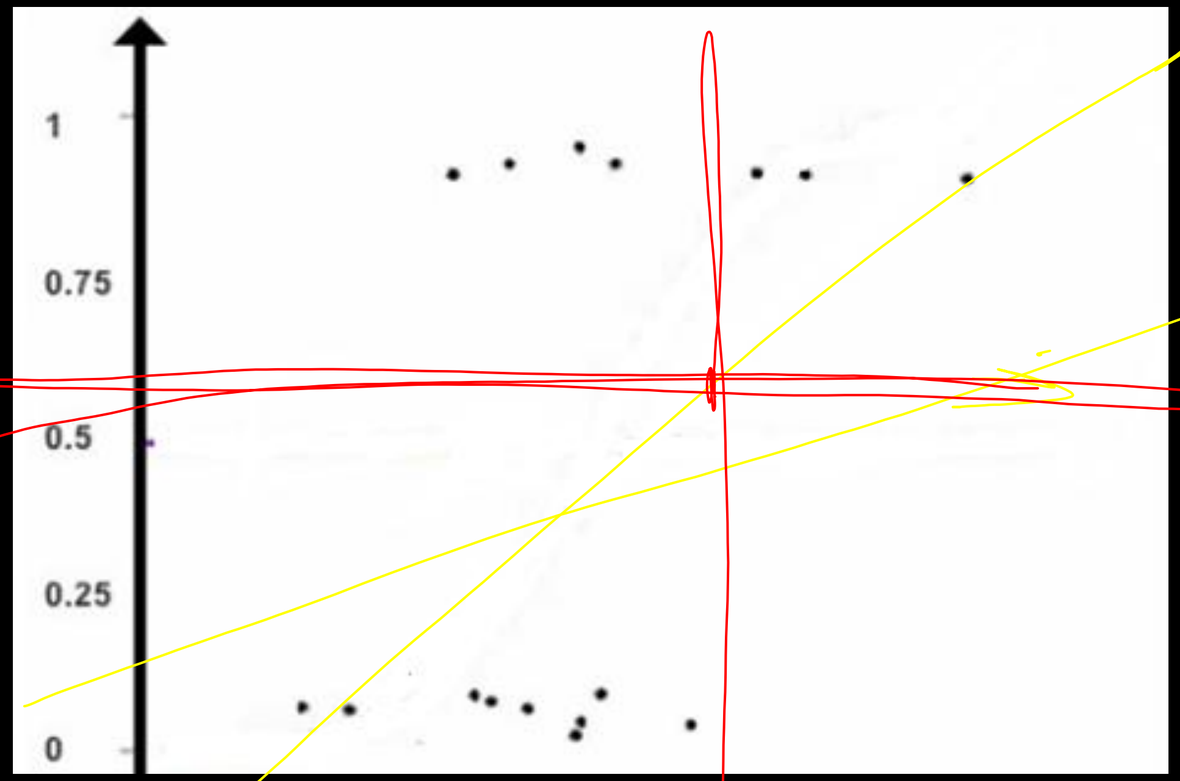
A. 2.0
B. 2.6
C. 2.8
D. 3.2

$$\frac{3+4+3+2}{3} = \frac{8.4}{2} = 2.8$$

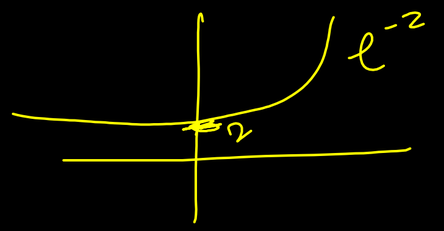
Answer: (C)

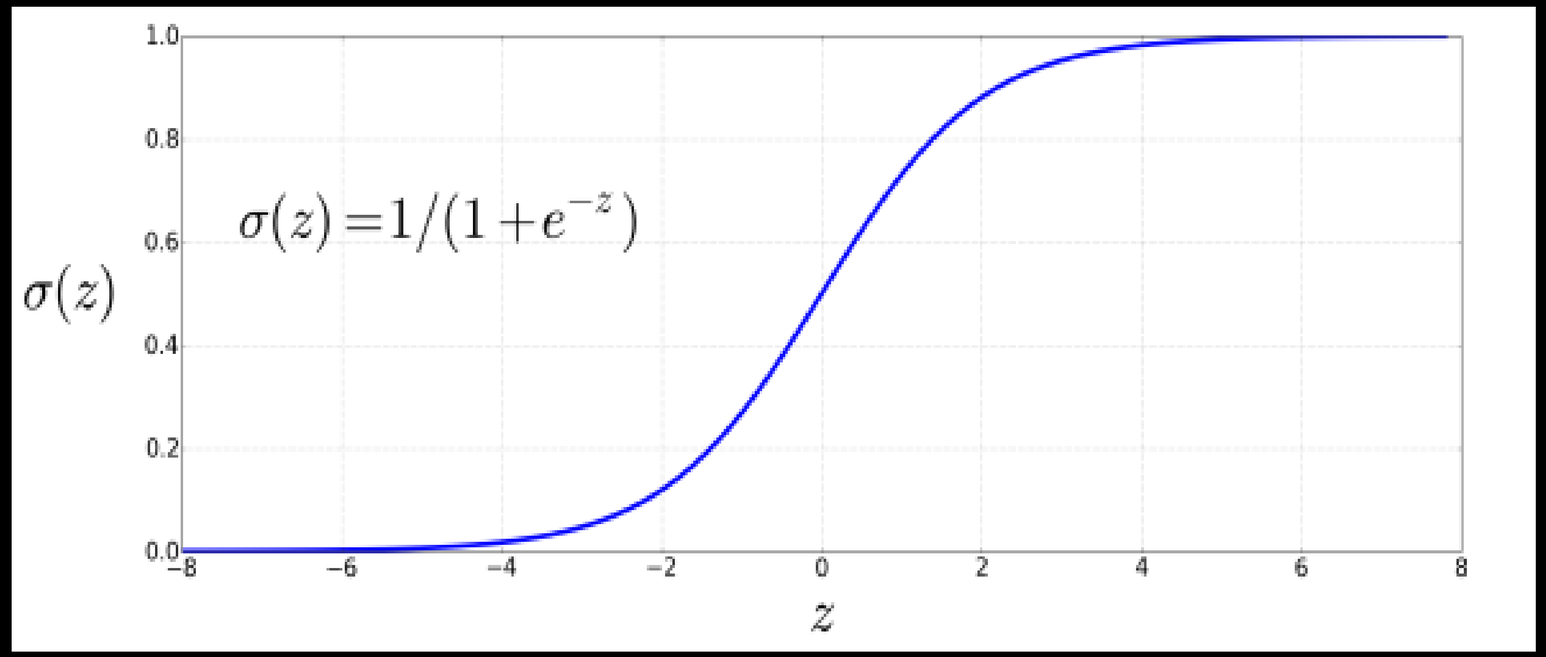
Classification

Logistic Regression



Sigmoid Function





$z = wx + b$

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

$$= \frac{1}{1+e^{-(wx+b)}}$$

$\sigma(z) = P(y=1)$

$1 - \sigma(z) = P(y=0)$

$w = 2$, $b = 1$

$$\sigma(z) = \frac{1}{1+e^{-2x+1}}$$

$$= \frac{1}{1+e^{-1}}$$

$0<\sigma(z)<1$ (Output between 0 and 1)

$$
\begin{aligned}
P(y=1) &= \sigma(\mathbf{w}\cdot\mathbf{x}+b) \\
&= \frac{1}{1+\exp(-(\mathbf{w}\cdot\mathbf{x}+b))}
\end{aligned}
$$

$\Rightarrow \sigma(z)$

$$
\begin{aligned}
P(y=0) &= 1-\sigma(\mathbf{w}\cdot\mathbf{x}+b) \\
&= 1-\frac{1}{1+\exp(-(\mathbf{w}\cdot\mathbf{x}+b))} \\
&= \frac{\exp(-(\mathbf{w}\cdot\mathbf{x}+b))}{1+\exp(-(\mathbf{w}\cdot\mathbf{x}+b))}
\end{aligned}
$$

$\Rightarrow 1-\sigma(z)$

Goal w & b

$(y-y')^2$

1	0
0	1

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

$$\sigma'(z) = \frac{e^{-z}}{(1+e^{-z})^2} = \sigma(z)\left[1-\sigma(z)\right]$$

Decision boundary

$$\text{decision}(x) = \begin{cases} 1 & \text{if } P(y=1|x) > 0.5 \\ 0 & \text{otherwise} \end{cases}$$

Loss Function

$L(\hat{y},y) = -[y \log \hat{y} + (1-y)\log(1-\hat{y})]$

$$L(\hat{y},y) = -[y \log \sigma(\mathbf{w}\cdot\mathbf{x}+b) + (1-y)\log(1-\sigma(\mathbf{w}\cdot\mathbf{x}+b))]$$

$\hat{y} = \frac{1}{1+e^{-z}}$.

$-\sum y_i \log_2 y_i$

$y=1$, $y=0$

$L(\hat{y},y) = -\log \hat{y}$ $\quad L(\hat{y},y) = -\log(1-\hat{y})$

$J = \frac{1}{2n}\sum L(y,\hat{y})$

$z = wx + b$

$w = w - \alpha \frac{\partial L}{\partial w}$

$b = b - \alpha \frac{\partial L}{\partial b}$

Gradient Descent Implementation

$$L(\hat{y}, y) = -[y \log \sigma(\mathbf{w} \cdot \mathbf{x} + b) + (1 - y) \log (1 - \sigma(\mathbf{w} \cdot \mathbf{x} + b))]$$

$\sigma(z)$

$$\frac{\partial L}{\partial w} = -\left[y \frac{1}{\sigma(z)} \times \sigma(z)(1-\sigma(z))\, x + (1-y) \frac{1}{1-\sigma(z)} (-\sigma(z))(1-\sigma(z))\, x\right]$$

$$= -\left[xy - xy\sigma(z) - x\sigma(z) + xy\sigma(z)\right]$$

$$= -x(y - \sigma(z))$$

$$= (y - \sigma(z))x = (y - \hat{y})x$$

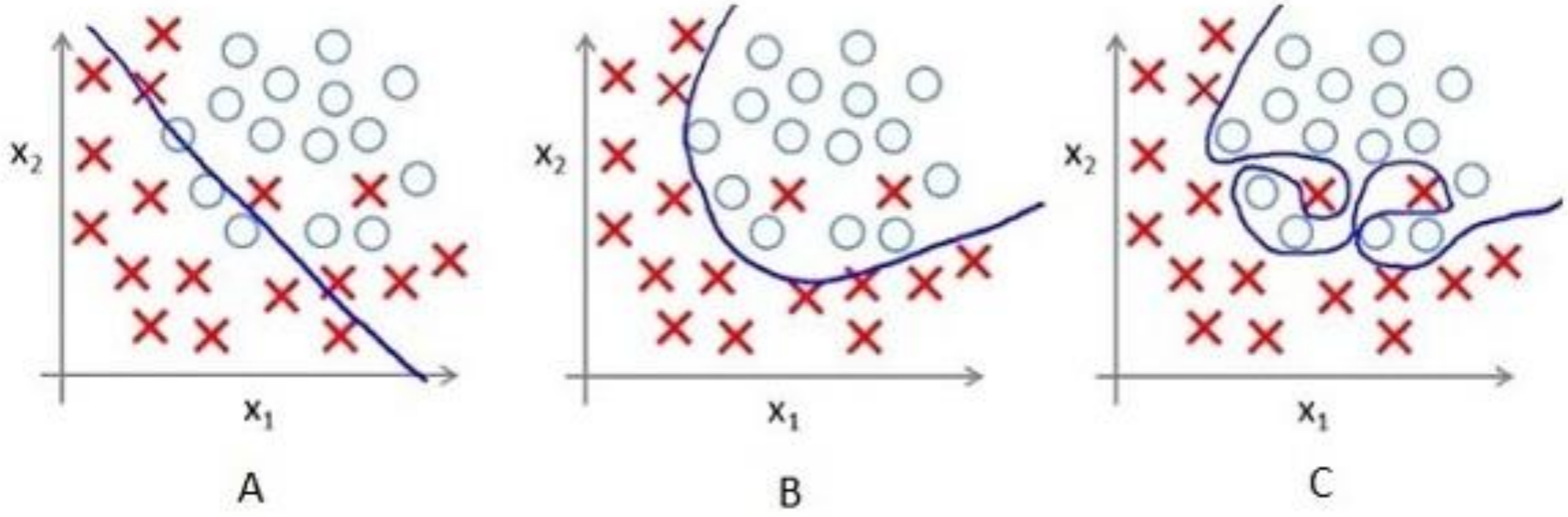
$$\frac{\partial L}{\partial b} = (y - \hat{y})$$

If g(z) is the sigmoid function, then its derivative with respect to z may be written in term

of g(z) as

a. $g(z)(g(z)-1)$
b. $g(z)(1+g(z))$
c. $-g(z)(1+g(z))$
d. $g(z)(1-g(z))$

Below are the three scatter plot(A,B,C left to right) and hand drawn decision boundaries for logistic regression.



Which of the following above figure shows that the decision boundary is overfitting the training data?

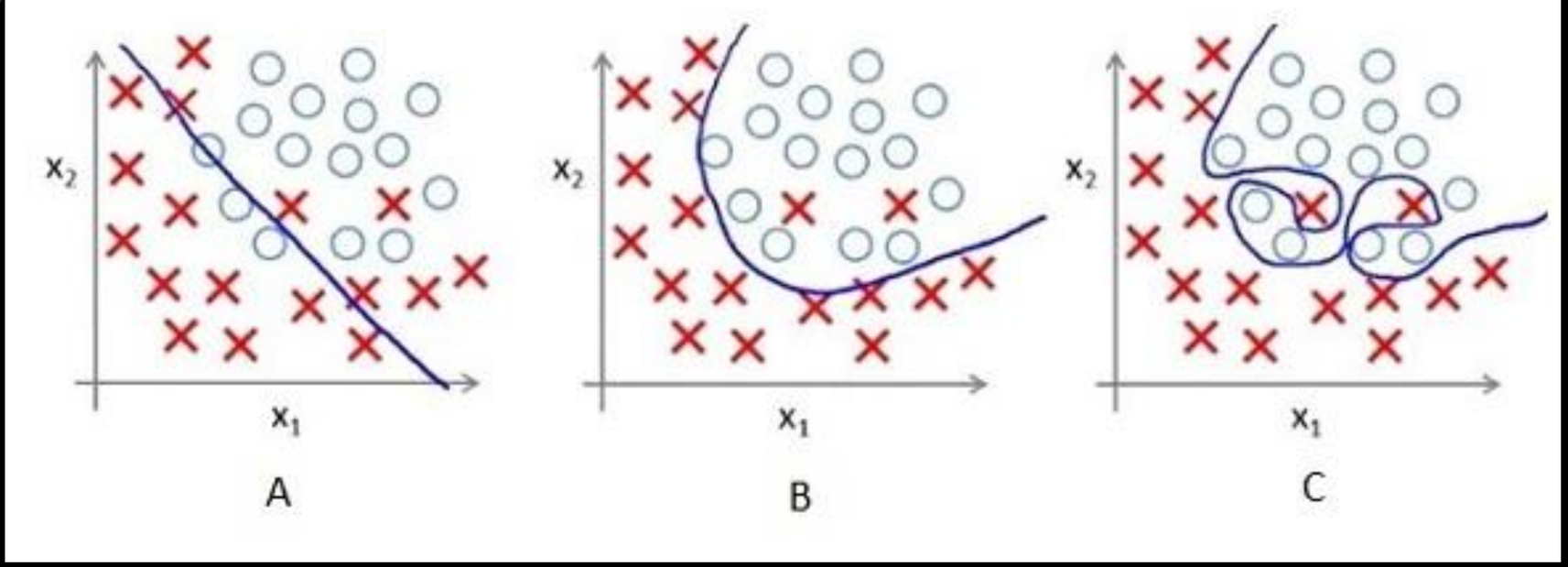
A) A

B) B

C) C

D)None of these

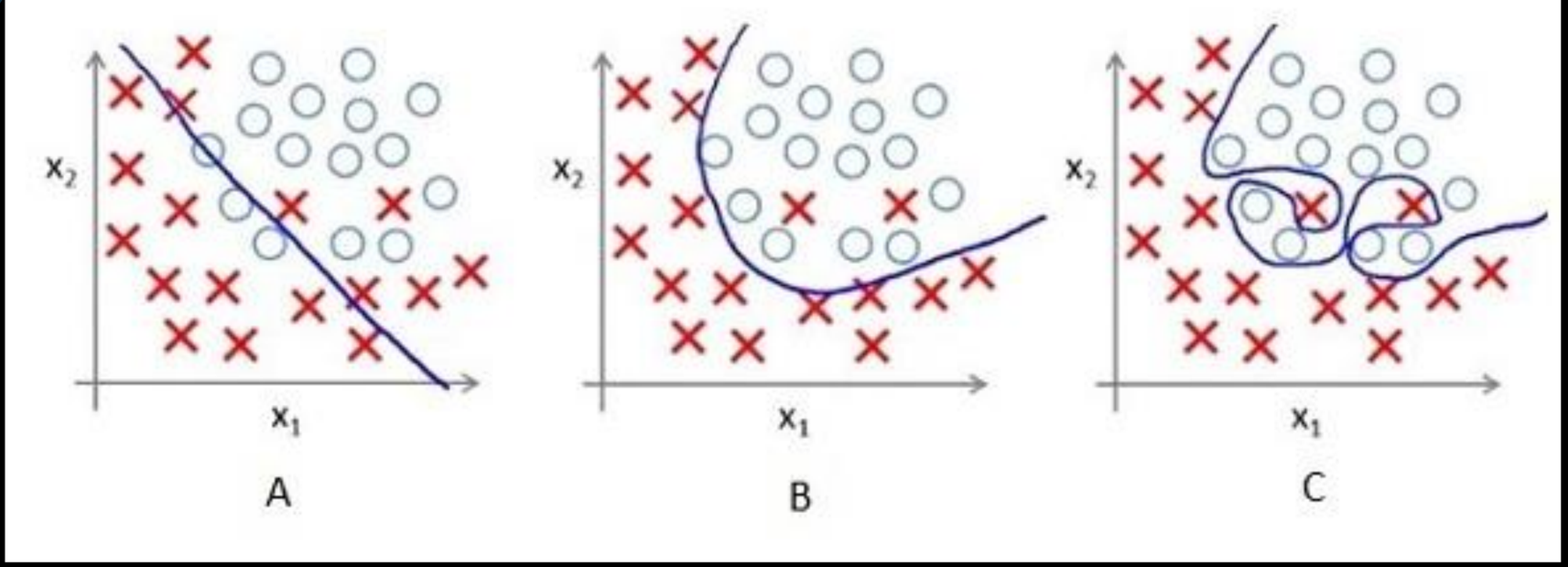
Answer: (C)



What do you conclude after seeing this visualization?

1.The training error in first plot is maximum as compare to second and third plot.
2.The best model for this regression problem is the last (third) plot because it has minimum training error (zero).
3.The second model is more robust than first and third because it will perform best on unseen data.
4.The third model is overfitting more as compare to first and second.
5.All will perform same because we have not seen the testing data.

Answer: (1,3 &4)



Suppose, above decision boundaries were generated for the different value of regularization. Which of the above decision boundary shows the maximum regularization?

A) A

B) B

C) C

D) All have equal regularization

Answer: (A)

Which of the following are symptoms of a logistic regression model being overfit? Select all that apply

(a) Large estimated coefficients
(b) Good generalization to unseen data
(c) Simple decision boundary
(d) Complex decision boundary

Answer: (a), (d)

How will the bias change on using high(infinite) regularization?

A)Bias will be high
B) Bias will be low
C)Can’t say
D)None of these

Answer: (A)

Suppose you are using a Logistic Regression model on a huge dataset. One of the problem you may face on such huge data is that Logistic regression will take very long time to train.

A) Decrease the learning rate and decrease the number of iteration

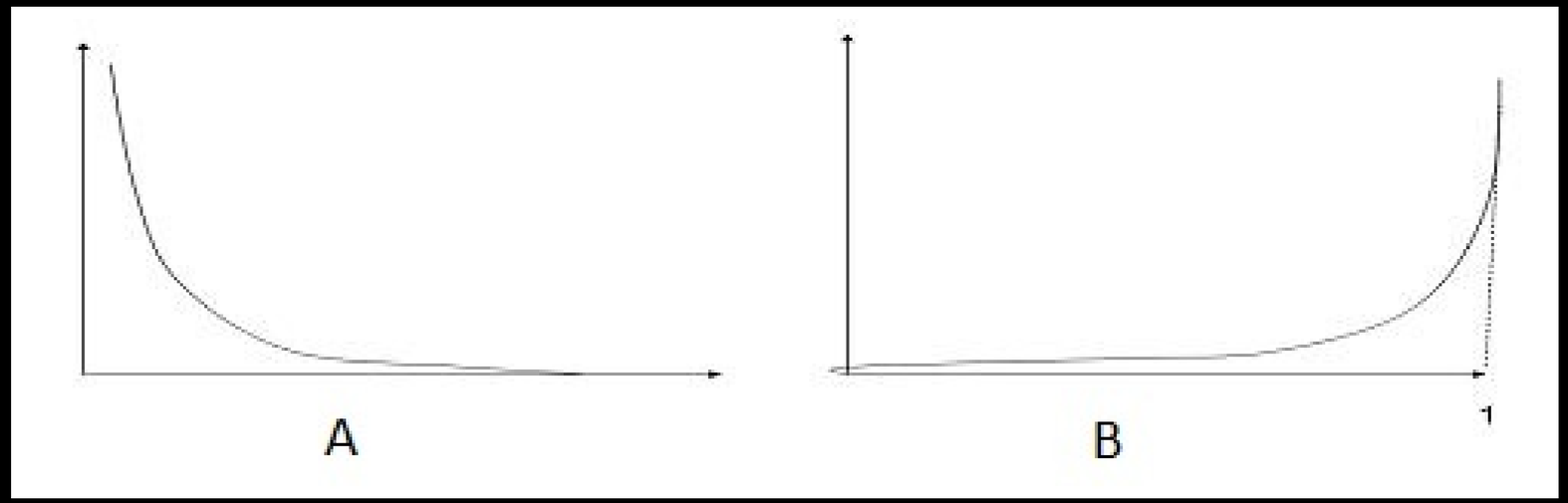
B) Decrease the learning rate and increase the number of iteration

C) Increase the learning rate and increase the number of iteration

D) Increase the learning rate and decrease the number of iteration

Solution: D

Which of the following image is showing the cost function for y =1.



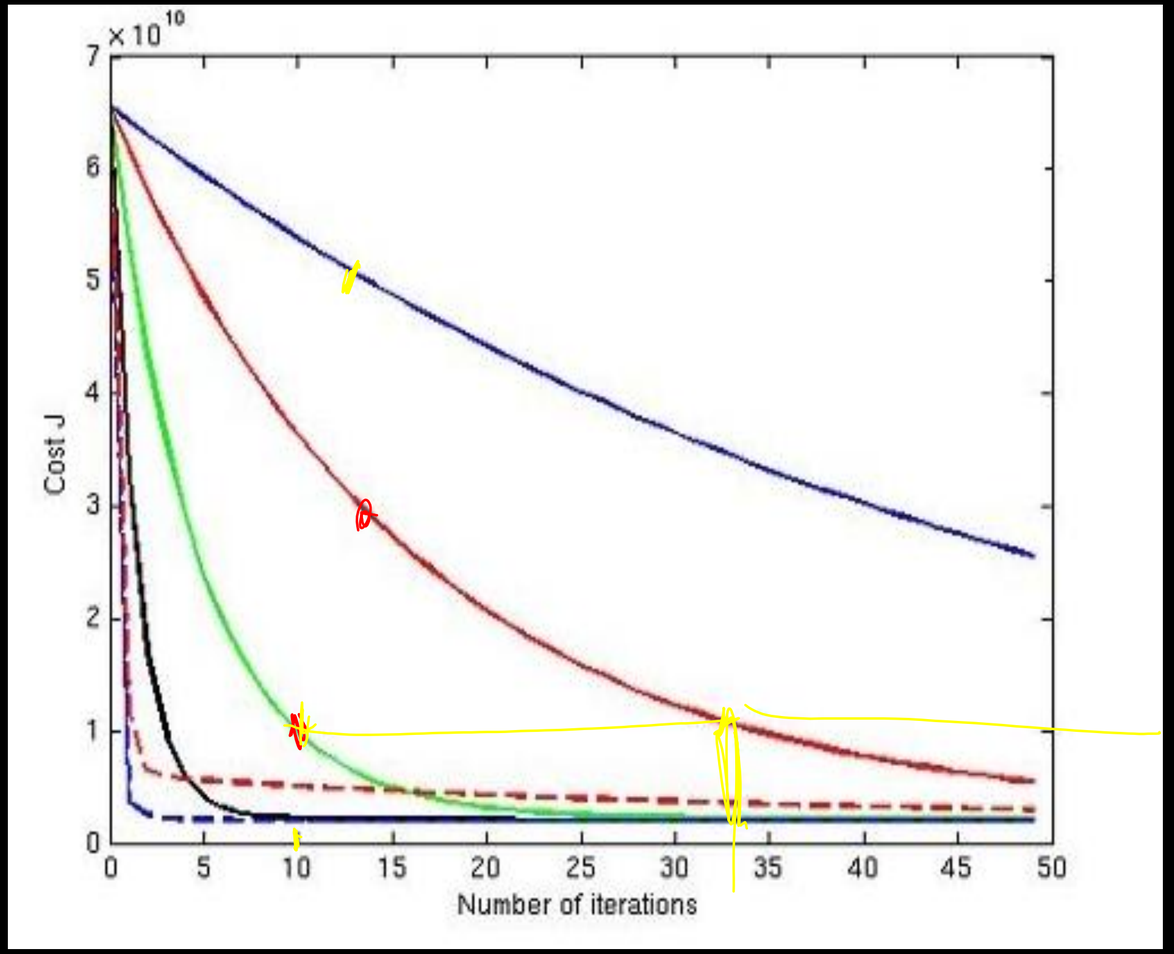
A) A
B) B
C) Both
D) None of these

Solution: A

Imagine, you have given the below graph of logistic regression which is shows the relationships between cost function and number of iteration for 3 different learning rate values (different colors are showing different curves at different learning rates).

1.The learning rate for blue is l1
2.The learning rate for red is l2
3.The learning rate for green is l3

A) l1>l2>l3
B) l1 = l2 = l3
C) l1 < l2 < l3
D) None of these



Solution: C

t me/Sakshi_singhal936

Naïve Bayes

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

$X = (x_1, x_2, x_3)$

$$P(C_1|X) = \frac{P(X|C_1)\ P(C_1)}{P(X)}$$

$$P(C_2|X) = \frac{P(X|C_2)\ P(C_2)}{P(X)}$$

$$P(C_3|X) = \frac{P(X|C_3) \cdot P(C_3)}{P(X)}$$

$$\propto P(X|C_1)\ P(C_1) = P(x_1, x_2, x_3|C_1)\ P(C_1)$$
$$= P(x_1|C_1)\ P(x_2|C_1)\ P(x_3|C_1)\ P(C_1)$$

$$\propto P(X|C_2)\ P(C_2)$$
$$= P(x_1|C_2)\ P(x_2|C_2)\ P(x_3|C_2)\ P(C_2)$$

$$\propto P(X|C_3)\ P(C_3) = P(x_1|C_3)\ P(x_2|C_3)\ P(x_3|C_3)\ P(C_3)$$

C_1 = Fish, C_2 = Animal, C_3 = Bird

Sl. No.	Swim	Fly	Crawl	Class
1	Fast	No	No	Fish
2	Fast	No	Yes	Animal
3	Slow	No	No	Animal
4	Fast	No	No	Animal
5	No	Short	No	Bird
6	No	Short	No	Bird
7	No	Rarely	No	Animal
8	Slow	No	Yes	Animal
9	Slow	No	No	Fish
10	Slow	No	Yes	Fish
11	No	Long	No	Bird
12	Fast	No	No	Bird

$$P(C_1|X) = P(x_1|C_1)\,P(x_2|C_1)\,P(x_3|C_1)\,P(C_1)$$
$$= \frac{2}{3} \times \frac{0}{3} \times \frac{2}{3} \times \frac{3}{12} = 0$$

$$P(C_2|X) = P(x_1|C_2)\,P(x_2|C_2)\,P(x_3|C_2)\,P(C_2)$$
$$= \frac{2}{5} \times \frac{1}{5} \times \frac{3}{5} \times \frac{5}{12}$$
$$= \frac{1}{50}$$

$$P(C_3|X) = P(x_1|C_3)\,P(x_2|C_3)\,P(x_3|C_3)\,P(C_3)$$
$$= \frac{0}{4} \times \frac{0}{4} \times \frac{4}{4} \times \frac{4}{12} = 0$$

$P(C_1) = \frac{3}{12}$, $P(C_2) = \frac{5}{12}$, $P(C_3) = \frac{4}{12}$

The test instance is (Slow, Rarely, No) (x_1, x_2, x_3) ⟶ C_2 → Animal

C_1 = Yes , C_2 = No $P(C_1) = \frac{5}{10}$, $P(C_2) = \frac{5}{10}$

Use naive Bayes algorithm to determine whether a red domestic SUV car is a stolen car or not using the following data:

Example no.	Colour	Type	Origin	Whether stolen
1	red	sports	domestic	yes
2	red	sports	domestic	no
3	red	sports	domestic	yes
4	yellow	sports	domestic	no
5	yellow	sports	imported	yes
6	yellow	SUV	imported	no
7	yellow	SUV	imported	yes
8	yellow	SUV	domestic	no
9	red	SUV	imported	no
10	red	sports	imported	yes

$$P(C_1|X) = P(x_1|C_1)\, P(x_2|C_1)\, P(x_3|C_1)\, P(C_1)$$
$$= \frac{3}{5} \times \frac{1}{5} \times \frac{2}{5} \times \frac{1}{2}$$
$$= \frac{3}{125}$$

$$P(C_2|X) = P(x_1|C_2)\, P(x_2|C_2)\, P(x_3|C_2) \cdot P(C_2)$$
$$= \frac{2}{5} \times \frac{3}{5} \times \frac{3}{5} \times \frac{1}{2}$$
$$= \frac{9}{125}$$

X (Red x_1, SUV x_2, domestic x_3) $\Rightarrow$ C_2 → Not

Consider the following dataset showing the result whether a person has passed or failed the exam based on various factors. Suppose the factors are independent to each other. We want to classify a new instance with Confident=Yes, Studied=Yes, and Sick=No.

Confident	Studied	Sick	Result
Yes	No	No	Fail
Yes	No	Yes	Pass
No	Yes	Yes	Fail
No	Yes	No	Pass
Yes	Yes	Yes	Pass

Yes Yes No ⇒

(True/False) Naïve Bayes algorithm can be used when some input features are in continuous range?

Salary

75,000

50,000

1 Lac

Salary

0 - 50

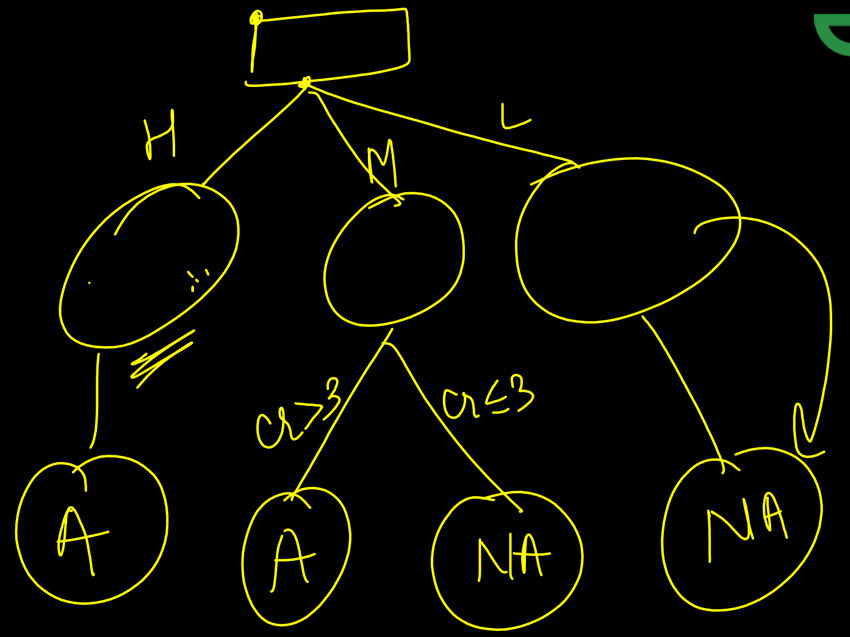
50 - 000

100 - 150

What is the naive assumption in a Naive Bayes Classifier?

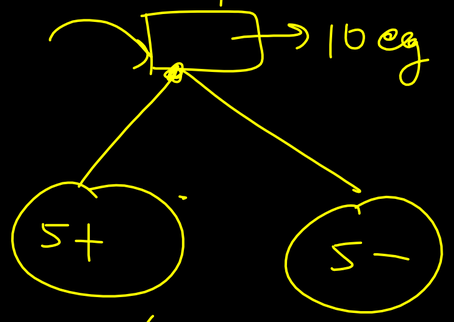
a. All the classes are independent of each other
b. All the features of a class are independent of each other
c. The most probable feature for a class is the most important feature to be considered for classification
d. All the features of a class are conditionally dependent on each other.

Decision Tree

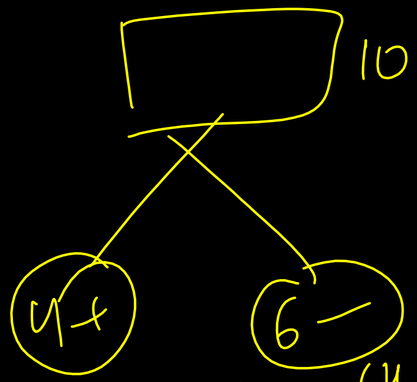


Salary	Credit score	Loan Approved/ NA
H	1	A
M	2	NA
L	3	.
	4	
Entropy	5	

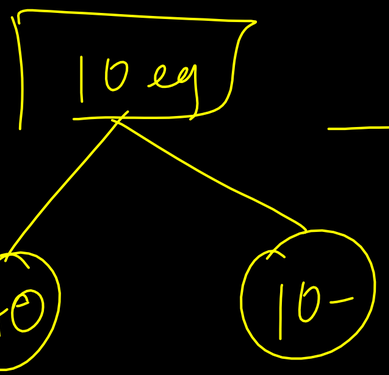
$$= -\sum P_i \log_2 P_i$$



$$E = -\left(\frac{5}{10}\log_2\frac{5}{10} + \frac{5}{10}\log_2\frac{5}{10}\right) = 1$$



$$E = -\left(\frac{4}{10}\log_2\frac{4}{10} + \frac{6}{10}\log_2\frac{6}{10}\right)$$



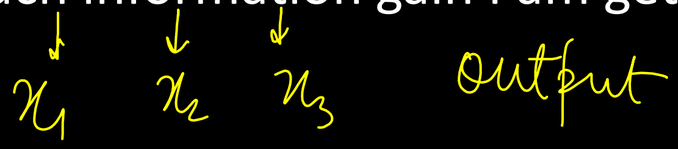
$\longrightarrow E = 0$

Entropy

Measure of “impurity” in a dataset

Information gain

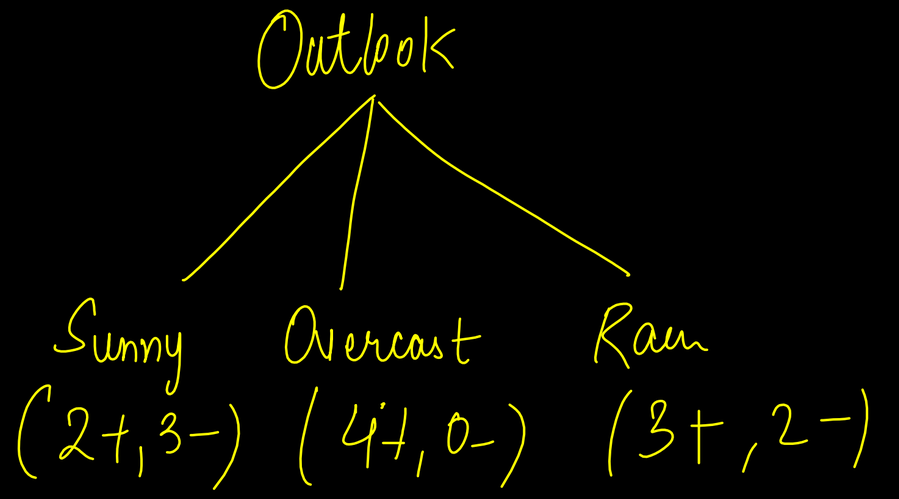
How much information gain I am getting from particular node.



Day	Outlook	Temp	Humidity	Wind	Tennis?
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Day	Outlook	Temp	Humidity	Wind	Tennis?
D1	Sunny	Hot	High	Weak	*No*
D2	Sunny	Hot	High	Strong	*No*
D3	Overcast	Hot	High	Weak	*Yes*
D4	Rain	Mild	High	Weak	*Yes*
D5	Rain	Cool	Normal	Weak	*Yes*
D6	Rain	Cool	Normal	Strong	*No*
D7	Overcast	Cool	Normal	Strong	*Yes*
D8	Sunny	Mild	High	Weak	*No*
D9	Sunny	Cool	Normal	Weak	*Yes*
D10	Rain	Mild	Normal	Weak	*Yes*
D11	Sunny	Mild	Normal	Strong	*Yes*
D12	Overcast	Mild	High	Strong	*Yes*
D13	Overcast	Hot	Normal	Weak	*Yes*
D14	Rain	Mild	High	Strong	*No*

Day	Outlook	Temp	Humidity	Wind	Tennis?
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No



$$IG = Entropy(S) - \sum_{v \in category} \frac{|S_v|}{|S|} Entropy(S, v)$$

$$= 0.940 - \left[\frac{5}{14} \times 0.971 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.971\right]$$

$$= 0.247$$

$$Entropy(S) = -\left(\frac{5}{14}\log_2\frac{5}{14} + \frac{9}{14}\log_2\frac{9}{14}\right) = 0.940$$

$$Entropy(S, Sunny) = -\left(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right) = 0.971$$

$$Entropy(S, Overcast) = 0$$

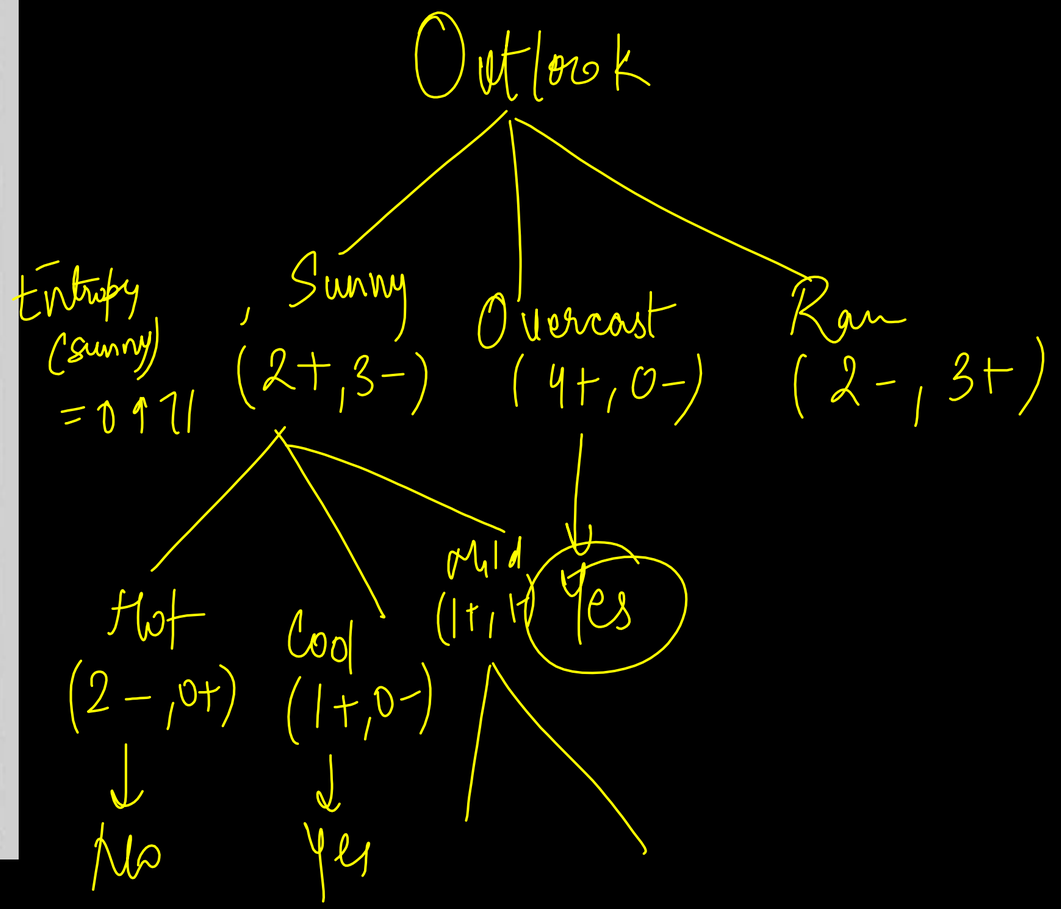
$$Entropy(S, Rain) = 0.971$$

$IG(Wind) = 0.048$

$IG(Temp) = 0.029$

$IG(Hum) = 0.151$

Day	Outlook	Temp	Humidity	Wind	Tennis?
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No



Day	Outlook	Temp	Humidity	Wind	Tennis?
D1	Sunny	Hot	High	Weak	*No*
D2	Sunny	Hot	High	Strong	*No*
D3	Overcast	Hot	High	Weak	*Yes*
D4	Rain	Mild	High	Weak	*Yes*
D5	Rain	Cool	Normal	Weak	*Yes*
D6	Rain	Cool	Normal	Strong	*No*
D7	Overcast	Cool	Normal	Strong	*Yes*
D8	Sunny	Mild	High	Weak	*No*
D9	Sunny	Cool	Normal	Weak	*Yes*
D10	Rain	Mild	Normal	Weak	*Yes*
D11	Sunny	Mild	Normal	Strong	*Yes*
D12	Overcast	Mild	High	Strong	*Yes*
D13	Overcast	Hot	Normal	Weak	*Yes*
D14	Rain	Mild	High	Strong	*No*

Day	Outlook	Temp	Humidity	Wind	Tennis?
D1	Sunny	Hot	High	Weak	*No*
D2	Sunny	Hot	High	Strong	*No*
D3	Overcast	Hot	High	Weak	*Yes*
D4	Rain	Mild	High	Weak	*Yes*
D5	Rain	Cool	Normal	Weak	*Yes*
D6	Rain	Cool	Normal	Strong	*No*
D7	Overcast	Cool	Normal	Strong	*Yes*
D8	Sunny	Mild	High	Weak	*No*
D9	Sunny	Cool	Normal	Weak	*Yes*
D10	Rain	Mild	Normal	Weak	*Yes*
D11	Sunny	Mild	Normal	Strong	*Yes*
D12	Overcast	Mild	High	Strong	*Yes*
D13	Overcast	Hot	Normal	Weak	*Yes*
D14	Rain	Mild	High	Strong	*No*

Day	Outlook	Temp	Humidity	Wind	Tennis?
D1	Sunny	Hot	High	Weak	*No*
D2	Sunny	Hot	High	Strong	*No*
D3	Overcast	Hot	High	Weak	*Yes*
D4	Rain	Mild	High	Weak	*Yes*
D5	Rain	Cool	Normal	Weak	*Yes*
D6	Rain	Cool	Normal	Strong	*No*
D7	Overcast	Cool	Normal	Strong	*Yes*
D8	Sunny	Mild	High	Weak	*No*
D9	Sunny	Cool	Normal	Weak	*Yes*
D10	Rain	Mild	Normal	Weak	*Yes*
D11	Sunny	Mild	Normal	Strong	*Yes*
D12	Overcast	Mild	High	Strong	*Yes*
D13	Overcast	Hot	Normal	Weak	*Yes*
D14	Rain	Mild	High	Strong	*No*

Consider the dataset, S given below:

Elevation	Road Type	Speed Limit	Speed
steep	Uneven	Yes	Slow
steep	Smooth	Yes	Slow
flat	Uneven	No	Fast
steep	Smooth	No	Fast

Elevation, Road Type and speed Limit are the features and Speed is the target label that we want to predict.

Elevation

steep (2S, 1F) flat (1F, 0S)

2S, 2F

$Entropy = -\left(\frac{2}{4}\log_2\frac{2}{4} + \frac{2}{4}\log_2\frac{2}{4}\right) = 1$

$0.918 = Entropy(S, steep) = -\left[\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}\right]$

$Entropy(S, flat) = -\left[\frac{1}{1}\log_2 1 + 0\right] = 0$

$IG = 1 - \left[\frac{3}{4} \times 0.918 + \frac{1}{4} \times 0\right] = 0.3$

Find the entropy of the dataset, S as given above:

a. 0.5
b. 0
c. 1
d. 0.7

Find the information Gain if the dataset is split at the feature "Elevation":

a. 1
b. 0
c. 0.675
d. 0.325

Find the feature on which the parent node must be chosen to split the dataset, S based on information gain:

a. Speed Limit
b. Road Type
c. Elevation

GATE DA 2024

Details of ten international cricket games between two teams “Green” and “Blue” are given in Table C. This table consists of matches played on different pitches, across formats along with their winners. The attribute Pitch can take one of two values: spin-friendly (represented as S) or pace-friendly (represented as F). The attribute Format can take one of two values: one-day match (represented as O) or test match (represented as T).

A cricket organization would like to use the information given in Table C to develop a decision-tree model to predict outcomes of future games between these two teams.

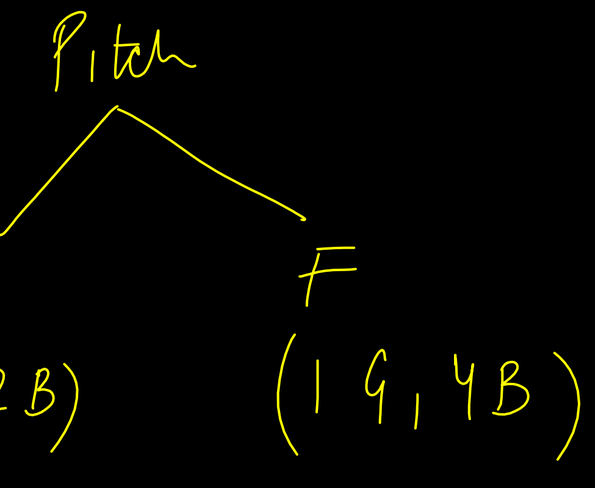
To develop such a model, the computed InformationGain(C, Pitch) with respect to the Target is _____ (rounded off to two decimal places).

C

4G,6B

$$Entropy(C) = -\left[\frac{4}{10}\log_2\frac{4}{10} + \frac{6}{10}\log_2\frac{6}{10}\right]$$
$$= 0.970$$

Match Number	Pitch	Format	Winner (Target)
1	S	T	Green
2	S	T	Blue
3	F	O	Blue
4	S	O	Blue
5	F	T	Green
6	F	O	Blue
7	S	O	Green
8	F	T	Blue
9	F	O	Blue
10	S	O	Green



$$Entropy(Pitch, S) = -\left[\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right]$$
$$= 0.970$$

$$Entropy(Pitch, F) = -\left[\frac{1}{5}\log_2\frac{1}{5} + \frac{4}{5}\log_2\frac{4}{5}\right]$$
$$= 0.721$$

Information Gain (G, Pitch)

$$= 0.970 - \left[\frac{5}{10} \times 0.970 + \frac{5}{10} \times 0.721\right]$$
$$= 0.125$$

True/False. Can we use a decision tree for both Regressions as well as Classification?

a) True

b) False

Find the decision tree.

x_1	1	3	4	6	10	15	2	7	16	0
x_2	12	23	21	10	27	23	35	12	27	17
y	10.1	15.3	11.5	13.9	17.8	23.1	12.7	43.0	17.6	14.9

T

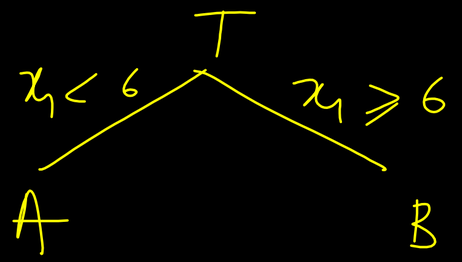
x_1	1	3	4	6	10	15	2	7	16	0
x_2	12	23	21	10	27	23	35	12	27	17
y	10.1	15.3	11.5	13.9	17.8	23.1	12.7	43.0	17.6	14.9

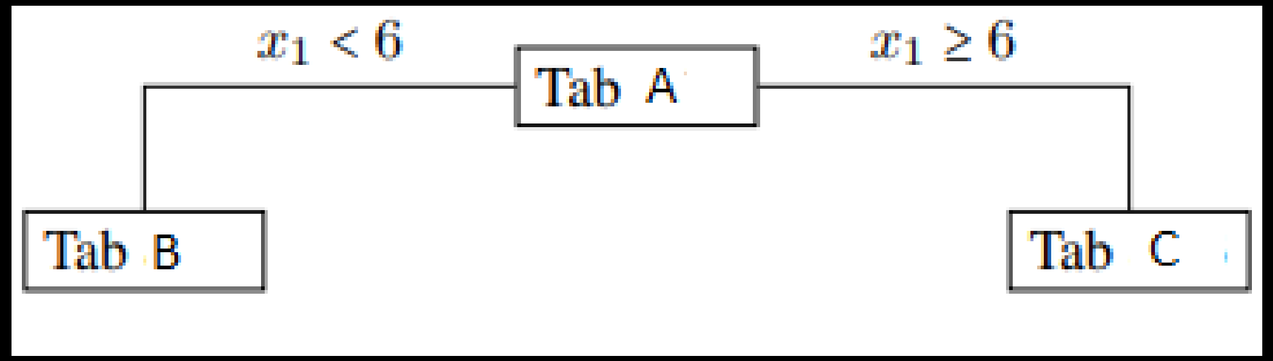
A

x_1	1	3	4	2	0
x_2	12	23	21	35	17
y	10.1	15.3	11.5	12.7	14.9

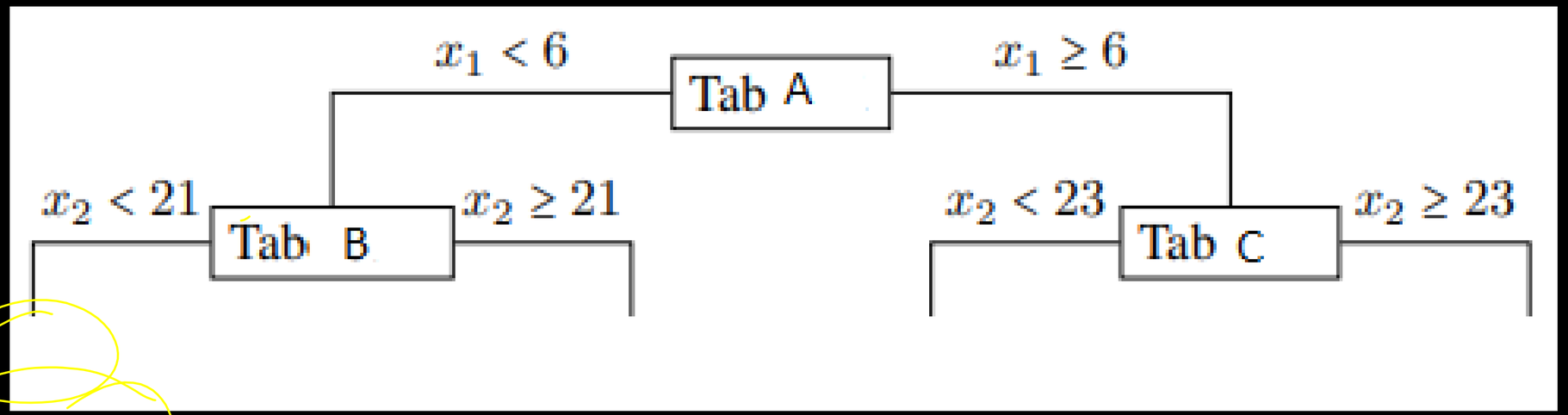
B

x_1	6	10	15	7	16
x_2	10	27	23	12	27
y	13.9	17.8	23.1	43.0	17.6



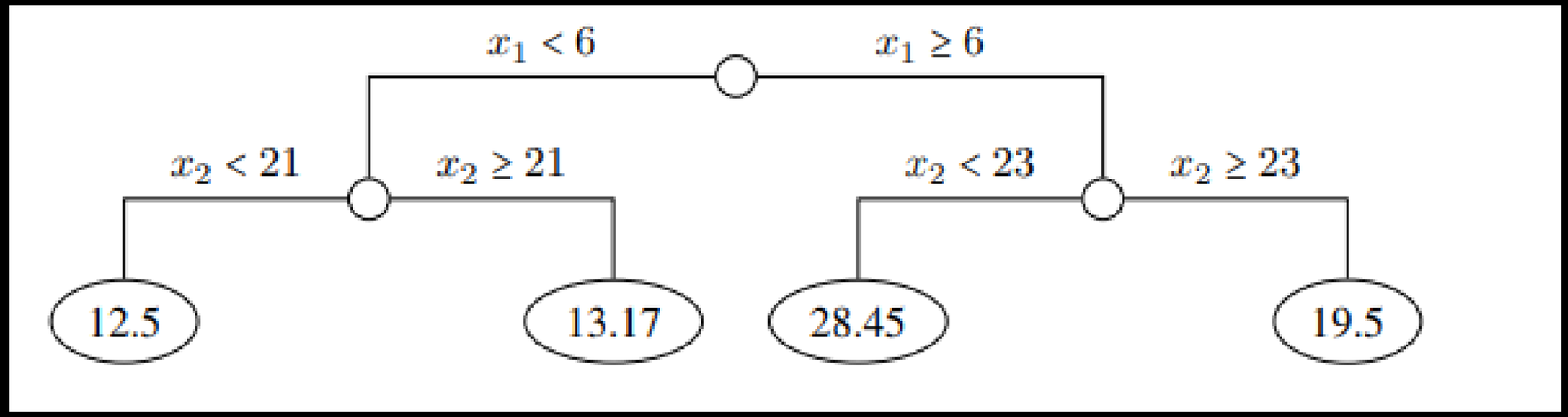


$x_1 = 5, x_2 = 10$



$\frac{10+14+9}{2}$

1


$x_1 < 6$
$x_1 \geq 6$
$x_2 < 21$
$x_2 \geq 21$
$x_2 < 23$
$x_2 \geq 23$
12.5
13.17
28.45
19.5

Decision tree is a _____.

a) Non-linear ML technique.

b) Non-Parametric technique.

c) Supervised Learning technique.

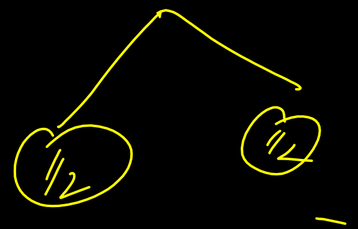
d) All of the above.

Entropy value for the data sample that has 50-50 split belonging to two categories is

a. 1

b. 0

c. None



Binary class

$0 < E < 1$

Which of the following is true for a decision tree?

a. A decision tree is an example of a linear classifier.
b. The entropy of a node typically decreases as we go down a decision tree.
c. Entropy is a measure of purity.
d. An attribute with lower mutual information should be preferred to other attributes.

Day	Outlook	Temp	Humidity	Wind	Tennis?
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

5M, 9Y

N: $\frac{5}{14}$ Y: $\frac{9}{14}$

Find the Gini Index of tennis.

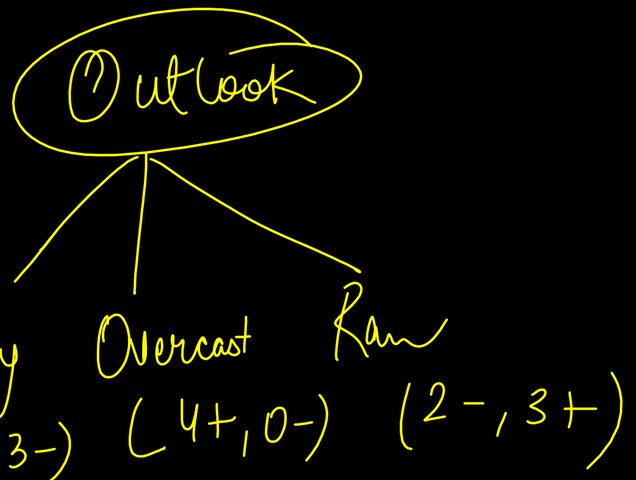
$$GI = 1 - \sum P_i^2$$

$$= 1 - \left[\left(\frac{5}{14}\right)^2 + \left(\frac{9}{14}\right)^2\right]$$

→ Gini split index = $\sum \frac{|S_v|}{|S|} GI(S, v)$

Choose feature which has minimum Gini split index.

Find the Gini Split Index of Outlook.



$$GI(S, \text{Sunny}) = 1 - \left[\left(\frac{2}{5}\right)^2 + \left(\frac{3}{5}\right)^2\right] = 0.48$$

$$GI(S, \text{Overcast}) = 1 - \left[\left(\frac{4}{4}\right)^2\right]$$

$$= 0$$

$$GI(S, \text{Rain}) = 0.48$$

$$\text{Gini Split Index (Outlook)} = \sum \frac{|S_v|}{|S|} \times GI(S, v)$$

$$= \frac{5}{14} \times 0.48 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.48$$

$$= 0.342$$

Overfitting in decision trees

Causes:

1. The tree has too many levels (i.e., it's too deep).
2. The tree splits down to very specific details of the training set

To deal with an overfitted Decision Tree, you can use one or more of the following methods:

Early stopping
Feature selection

Pruning: Pruning is a technique used to simplify the decision tree by removing branches that do not improve the tree's performance on new data.

Pruning involves building the full decision tree and then removing branches that do not improve the tree's performance on a validation set. The idea is to remove branches that are unlikely to be useful for predicting new data, thereby simplifying the tree and reducing the risk of overfitting.

Pruning is a technique that reduces the size of decision trees by removing sections of the tree that provide little power to classify instances. This is done in order to avoid:

a. overfitting
b. underfitting

Which of the following statements is not true about Pruning?

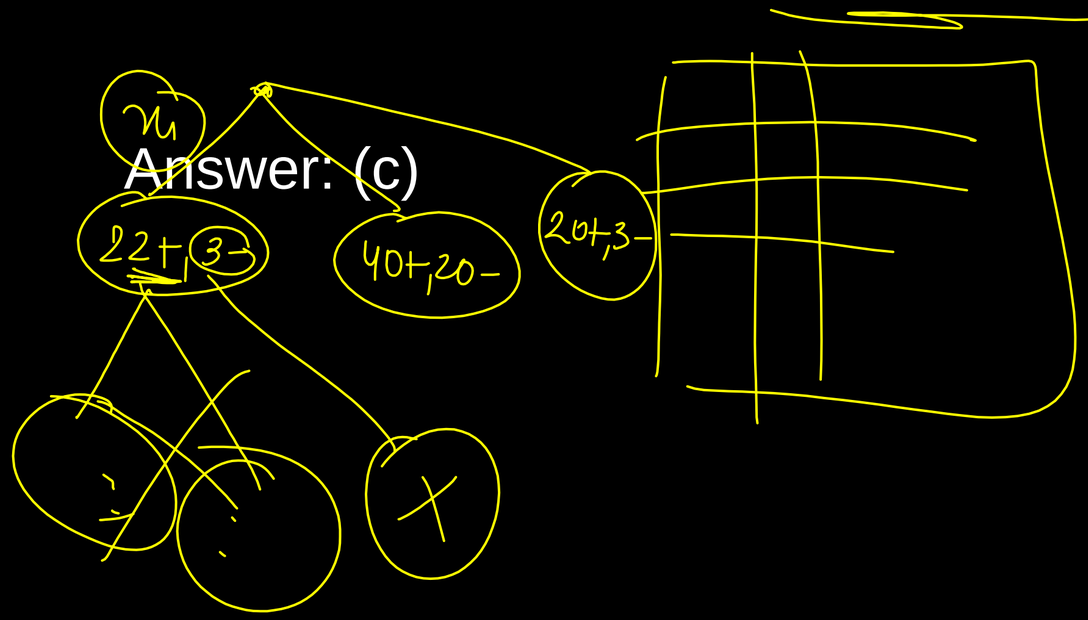
a) It removes the sections of the tree that provide little power to classify instances

b) It is a technique in machine learning and search algorithms to reduce the size of the decision trees

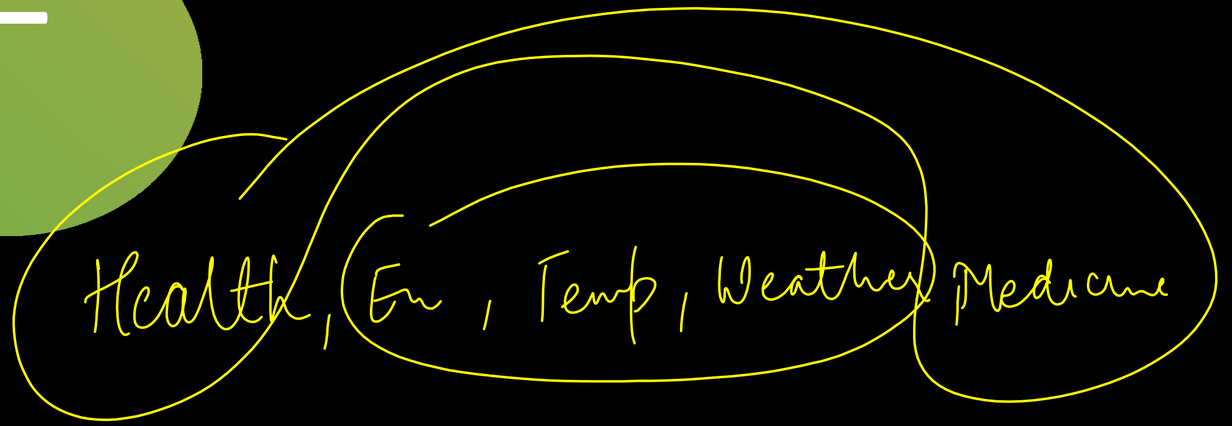
c) It increases the complexity of the final classifier

d) It improves the predictive accuracy by the reduction of overfitting

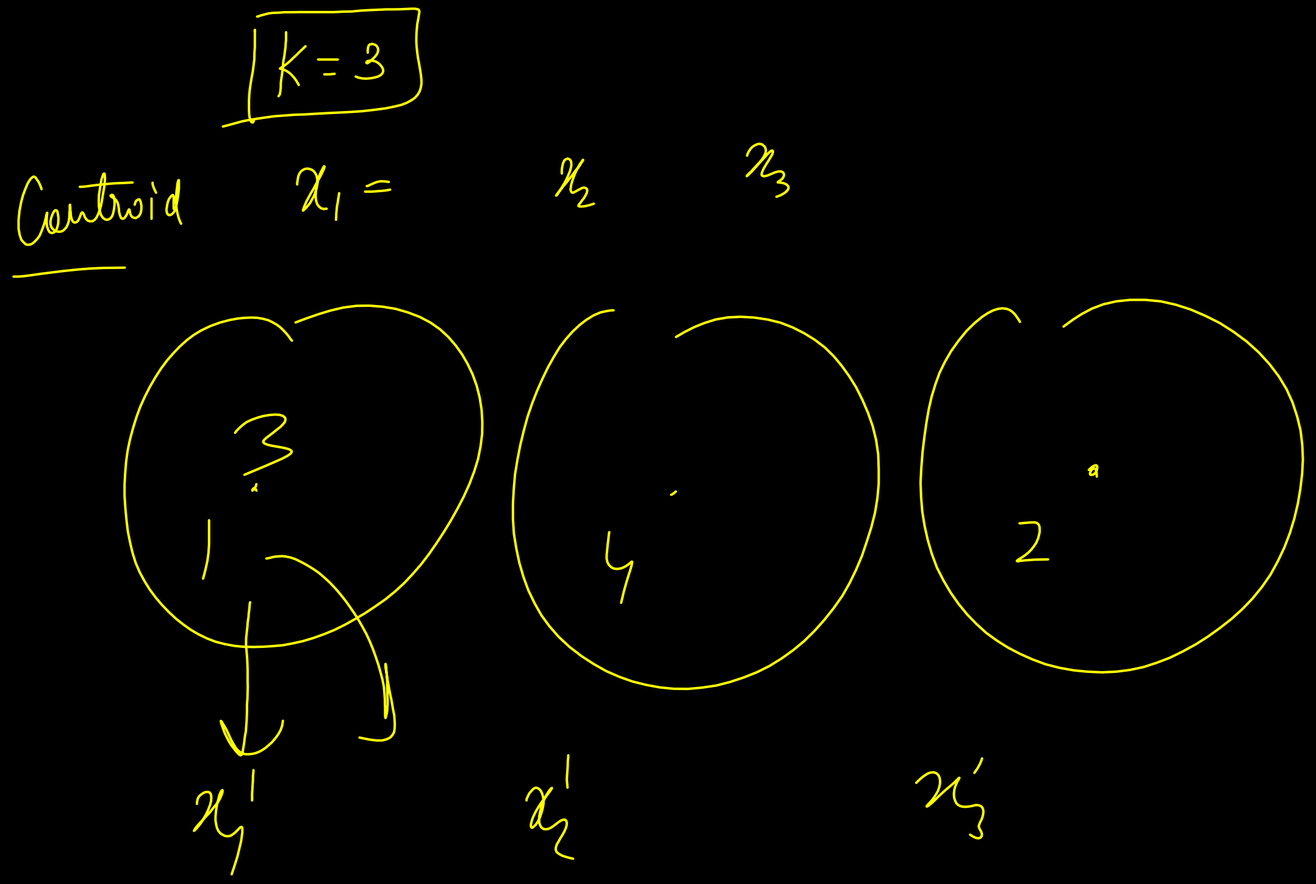
Answer: (c)



Clustering Algorithm – K Means

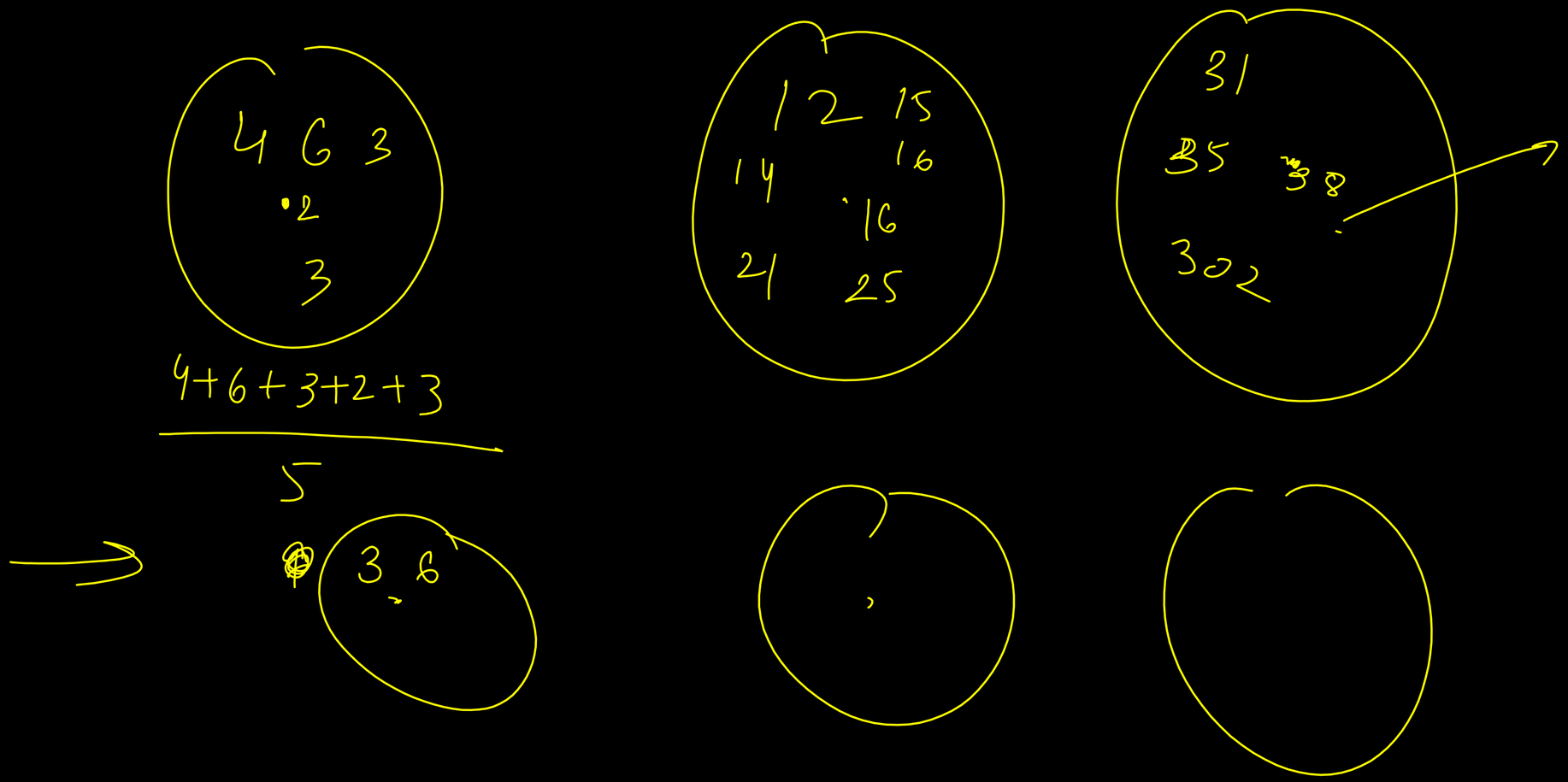


K-Means Clustering

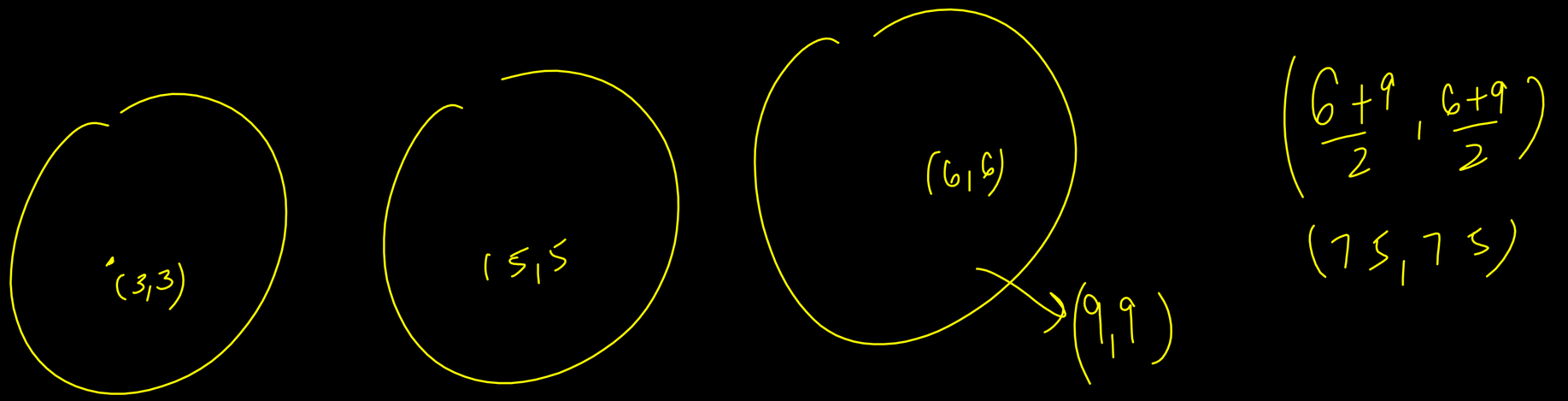


Problem Apply k-means algorithm for given data with k = 3. Use $C_1(2)$, $C_2(16)$ and $C_3(38)$ as initial centers.

Data: 2, 4, 6, 3, 31, 12, 15, 16, 38, 35, 14, 21, 3, 25, 30 2.



Given the dataset: (1, 1), (3, 3), (4, 4), (5, 5), (6, 6), (9, 9), (0, 3), (3, 0) and assuming the initial centroids for (K = 3 – means clustering) to be C_1 = (3, 3), C_2 = (5, 5) and C_3 = (6, 6). One iteration for K-means clustering, will update C_3 to (___, ___)



During a research work, you found seven observations as described with the data points below. You want to create three clusters from these observations using K-means algorithm.
After first iteration, the clusters C1, C2, C3 has following observations:
C1: {(2,2), (4,4), (6,6)}
C2: {(0,4), (4,0)}
C3: {(5,5), (9,9)}

If you want to run a second iteration then what will be the cluster centroids?

K-mean clustering algorithm has clustered the given 8 observations into 3 clusters after 1st iteration as follows:

C_1 : {(3,3), (5,5), (7,7)}

C_2 : {(0,6), (6,0), (3,0)}

C_3 : {(8,8),(4,4)}

What will be the Manhattan distance for observation (4,4) from cluster centroid C_1 in second iteration?

Answer: (2)

GATE DA 2024

Euclidean distance based k-means clustering algorithm was run on a dataset of 100 points with $k = 3$. If the points $\begin{bmatrix}1\\1\end{bmatrix}$ and $\begin{bmatrix}-1\\1\end{bmatrix}$ are both part of cluster 3, then which **ONE** of the following points is necessarily also part of cluster 3?

(A)	$\begin{bmatrix}0\\0\end{bmatrix}$
(B)	$\begin{bmatrix}0\\2\end{bmatrix}$
(C)	$\begin{bmatrix}2\\0\end{bmatrix}$
(D)	$\begin{bmatrix}0\\1\end{bmatrix}$

Which of the following can be possible termination conditions in K-Means?

1. For a fixed number of iterations.
2. Assignment of observations to clusters does not change between iterations. Except for cases with a bad local minimum.
3. Centroids do not change between successive iterations.
4. All of the above

What is true about K-Mean Clustering?

1. K-means is extremely sensitive to cluster center initializations
2. Bad initialization can lead to Poor convergence speed
3. Bad initialization can lead to bad overall clustering

a. 1 and 2
b. 1 and 3
c. All of the above
d. 2 and 3

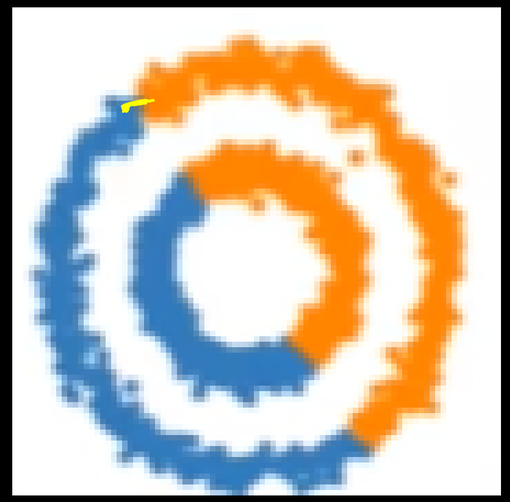
Which of the following clustering algorithms is the most sensitive to outliers?

a. K-means clustering algorithm
b. K-medians clustering algorithm
c. K-modes clustering algorithm
d. None of the above

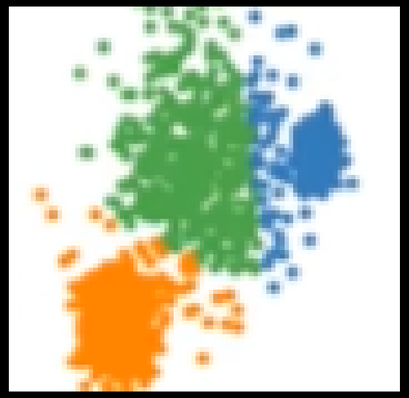
Limitations

1.Assumes Spherical Clusters:

K-Means performs well when clusters are spherical, equally sized, and have similar variances.



2. **Sensitive to Outliers:**
   Outliers can significantly impact cluster assignments and centroid positions.



3. Requires Specifying *K*:
   Determining the appropriate number of clusters (*K*) can be challenging.

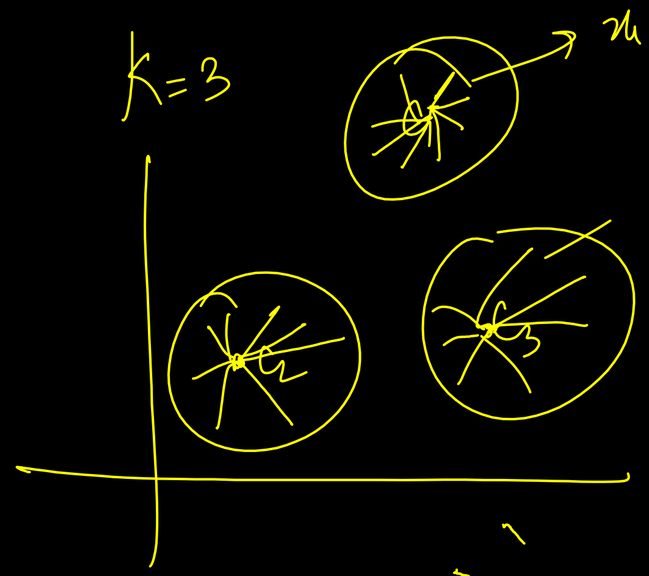
In which of the following cases will K-Means clustering fail to give good results? (Choose all the correct answers)

a. Data points with outliers
b. Data points with round shapes
c. Data points with non-convex shapes
d. Data points with different densities

SSE (or WCSS) in K Means

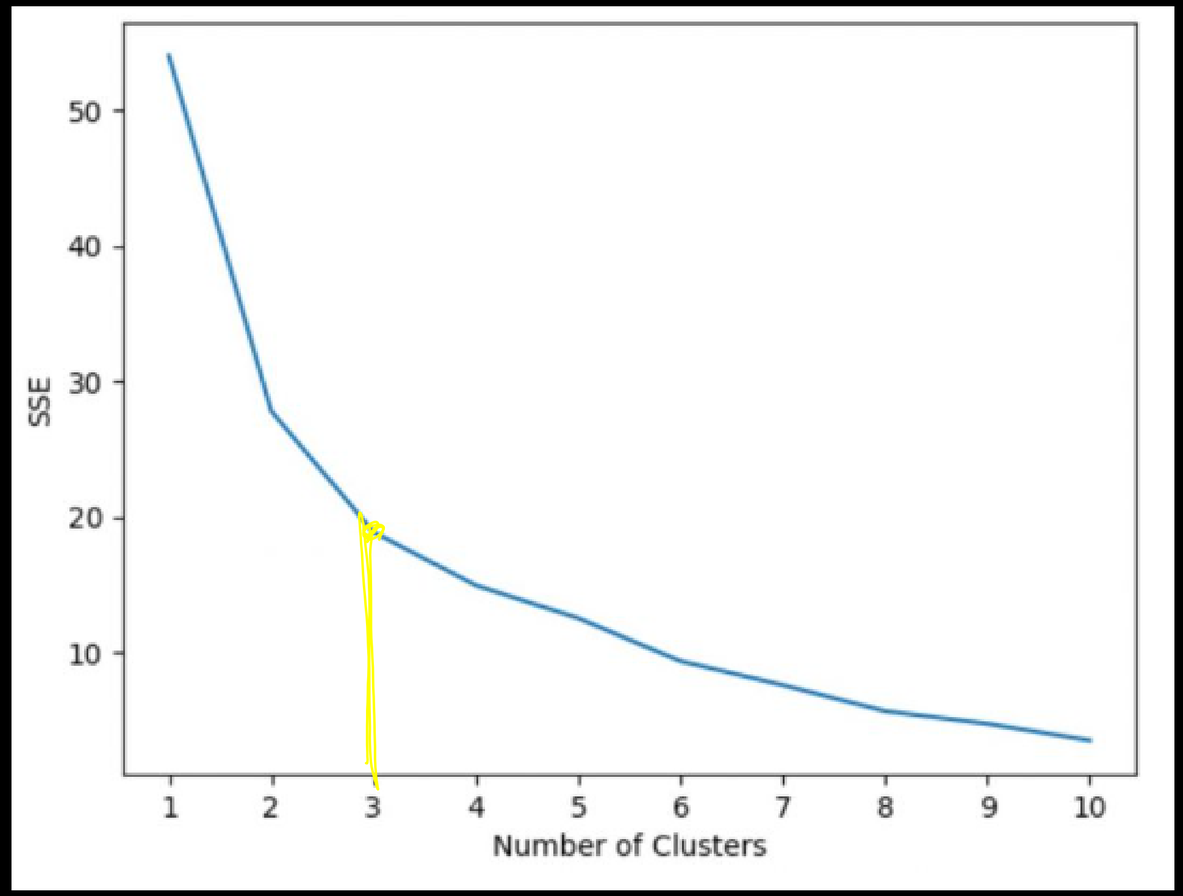
$$\text{WCSS} = \sum_{k=1}^{K} \sum_{x \in C_k} \|x - \mu_k\|^2$$

$$SSE = \sum_{k \in \text{Clusters}} \sum_{i \in C} |x_{ik} - c_i|^2$$



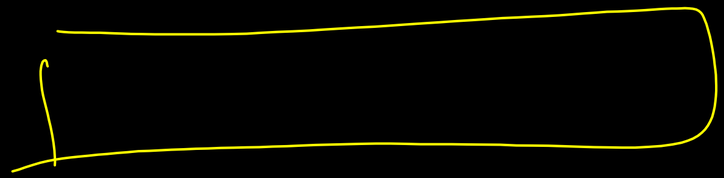
Elbow method

To determine the optimal number of clusters, we'll create a plot that displays the number of clusters along with the SSE (sum of squared errors) of the model.

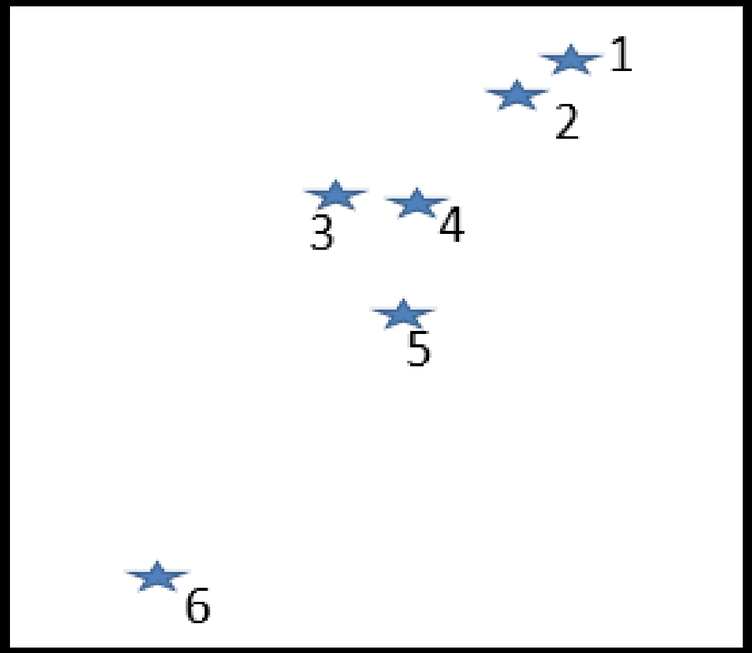


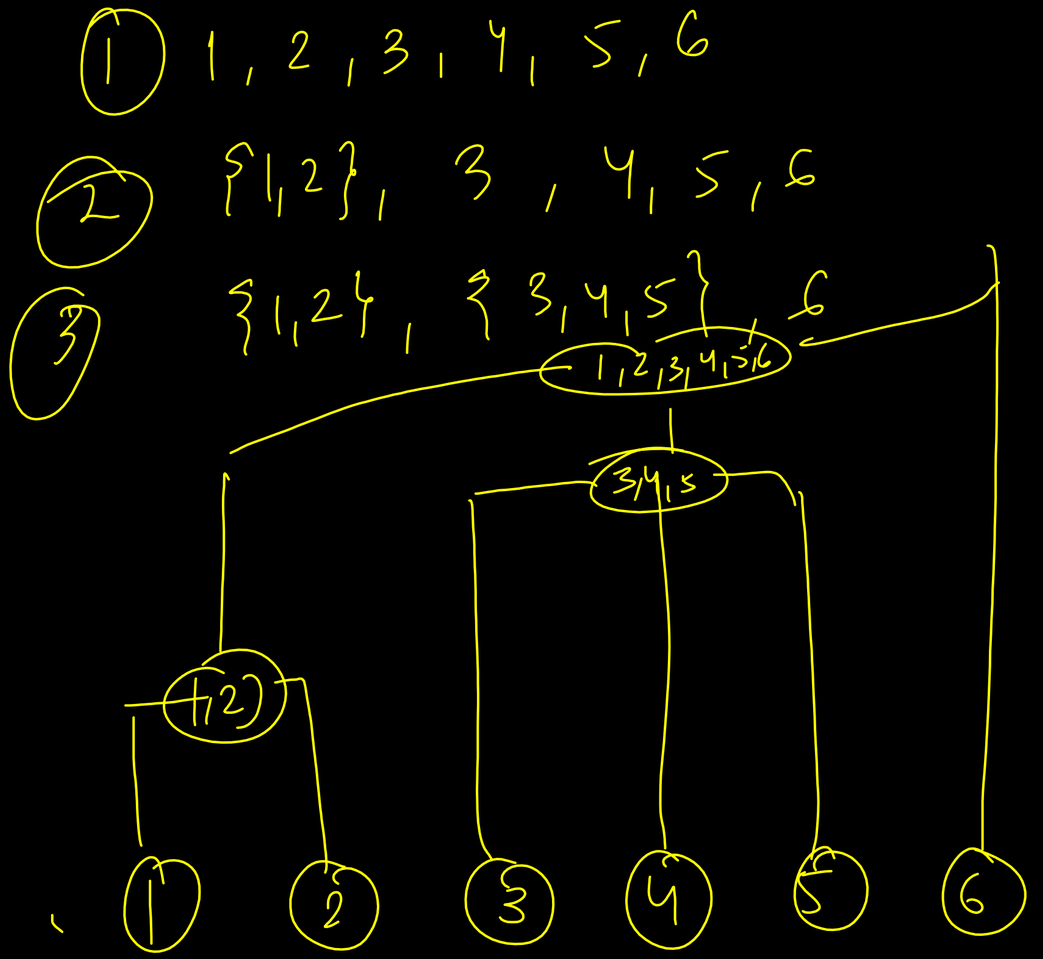
k = 3 clusters

Hierarchical Clustering

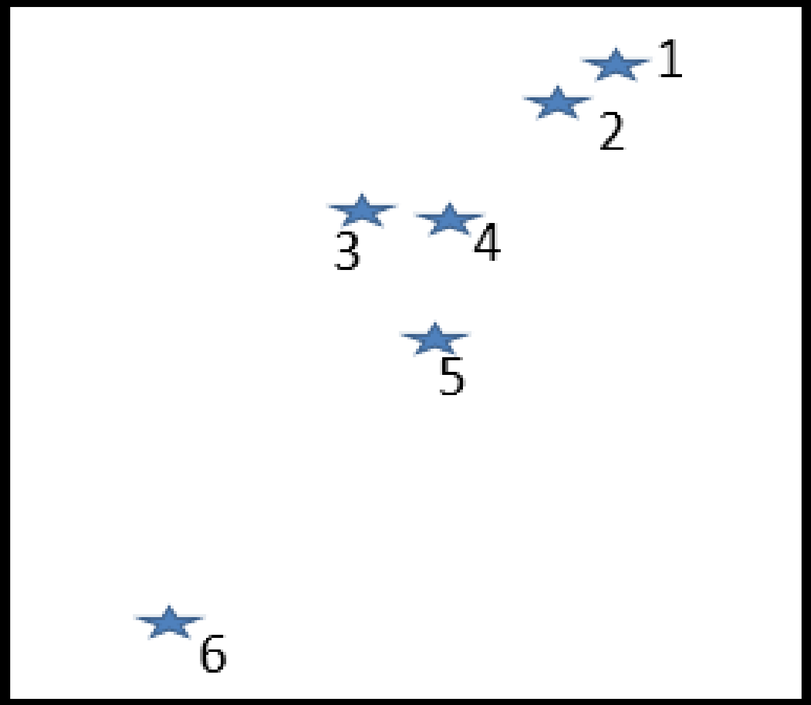


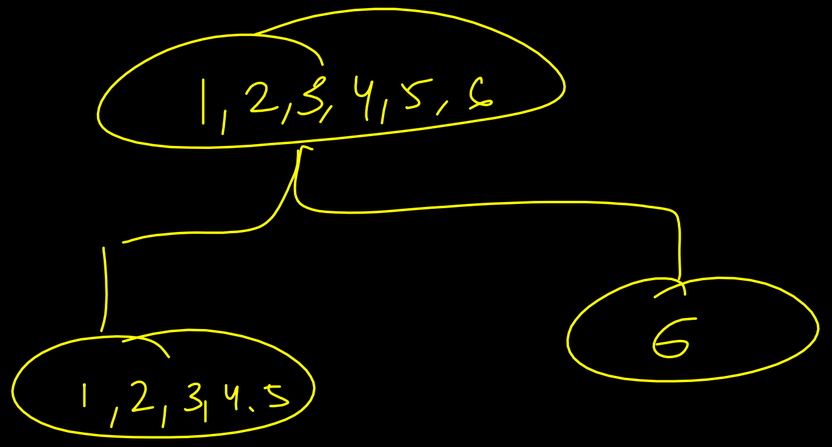
Agglomerative clustering (bottom up)





Divisive clustering (top down)





Measures of dissimilarity

The decision regarding whether two clusters are to be merged or not is taken based on the measure of dissimilarity between the clusters.

Measures of distance between data points

Numeric data

Name	Formula
Euclidean distance	$\|\vec{x} - \vec{y}\|_2 = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}$
Squared Euclidean distance	$\|\vec{x} - \vec{y}\|_2^2 = (x_1 - y_1)^2 + \cdots + (x_n - y_n)^2$
Manhattan distance	$\|\vec{x} - \vec{y}\|_1 = \|x_1 - y_1\| + \cdots + \|x_n - y_n\|$
Maximum distance	$\|\vec{x} - \vec{y}\|_\infty = \max\{\|x_1 - y_1\|, \ldots, \|x_n - y_n\|\}$

Non-numeric data

the Levenshtein distance between “kitten” and “sitting” is 3

Measures of dissimilarity

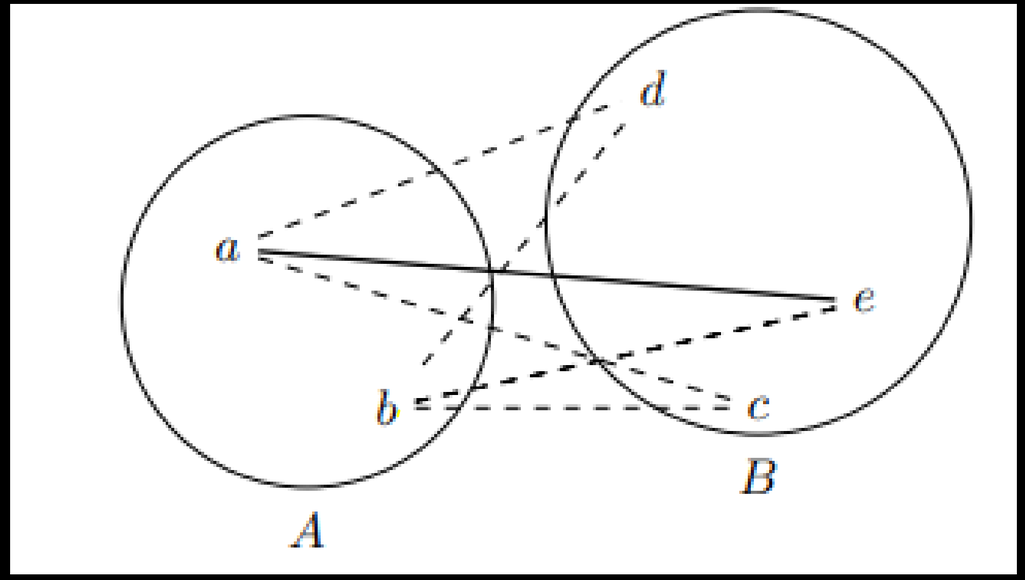
Measures of distance between groups of data points

Let d(x, y) = distance between x and y (say Euclidean distance formula)
d(A, B) = distance between the groups A and B.

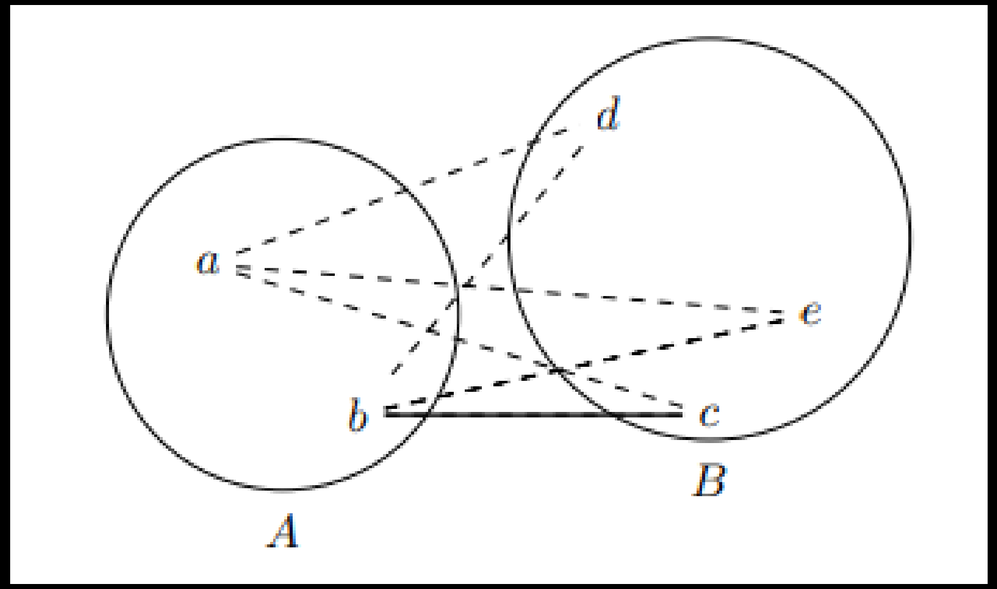
The following are some of the different methods in which d(A, B) is defined.

1. $d(A, B) = \max\{d(x, y) : x \in A, y \in B\}$.

Agglomerative hierarchical clustering using this measure of dissimilarity is known as complete linkage clustering. The method is also known as farthest neighbour clustering.



2. d(A, B) = min{d(x, y) : x ∈ A, y ∈ B}.
Agglomerative hierarchical clustering using this measure of dissimilarity is known as single linkage clustering.
The method is also known as nearest neighbour clustering.



3. $d(A, B) = \frac{1}{|A||B|} \sum_{x \epsilon A, y \epsilon B} d(x, y)$ where $|A|$, $|B|$ are respectively the number of elements in A and B.

Problem Given the dataset {a, b, c, d, e} and the following distance matrix, construct a dendrogram by complete linkage hierarchical clustering using the agglomerative method.

	a	b	c	d	e
a	0	9	3	6	11
b	9	0	7	5	10
c	3	7	0	9	2
d	6	5	9	0	8
e	11	10	2	8	0

① a, b, c, d, e

② a, b, d, {c, e}

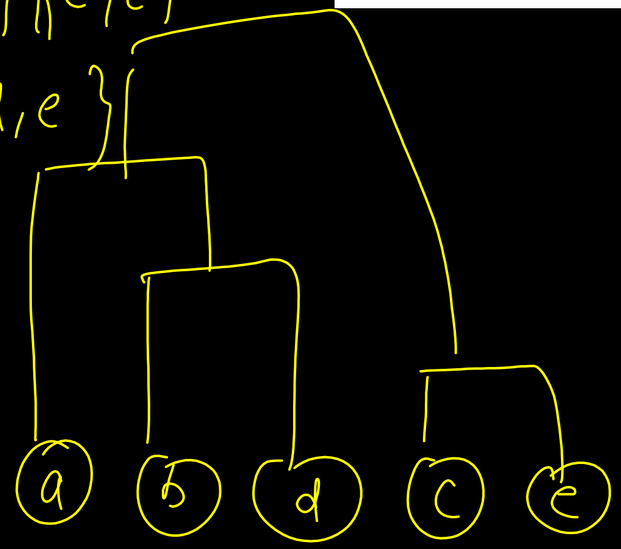
③ a, {b, d}, {c, e}

④ {a, b, d}, {c, e}

⑤ {a, b, c, d, e}

	a	b	d	{c, e}
a	0	9	6	11
b	9	0	5	10
d	6	5	0	9
{c, e}	11	10	9	0

	a	{b, d}	{c, e}
a	0	9	11
{b, d}	9	0	10
{c, e}	11	10	0



Problem Given the dataset {a, b, c, d, e} and the following distance matrix, construct a dendrogram by single linkage hierarchical clustering using the agglomerative method.

① a, b, c, d, e

② a, b, d, {c, e}

③ a, b, {d, c, e}

④

	a	b	c	d	e
a	0	9	3	6	11
b	9	0	7	5	10
c	3	7	0	9	2
d	6	5	9	0	8
e	11	10	2	8	0

	a	b	d	{c, e}
a	0	9	6	3
b	9	0	5	7
d	6	5	0	8
{c, e}	3	7	8	0

Given the following distance matrix, construct the dendogram using agglomerative clustering with average linkage.

	A	B	C	D	E
A	0	1	2	2	3
B	1	0	2	4	3
C	2	2	0	1	5
D	2	4	1	0	3
E	3	3	5	3	0

$\frac{2+2+2+4}{2 \times 2} =$

$\frac{3+3}{2 \times 1}$

$\frac{5+3}{2 \times 1}$

=

① A, B, C, D, E

② {A, B}, {C, D}, {E}

③ {A, B, C, D}, {E}

④ {A, B, C, D, E}

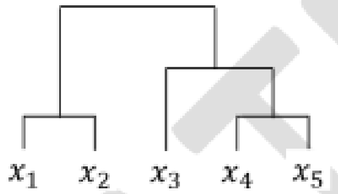
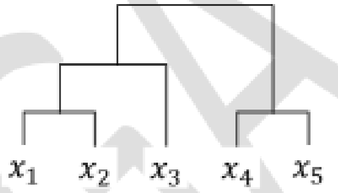
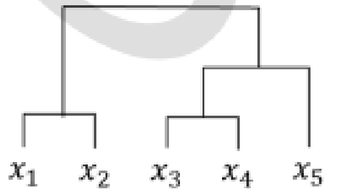
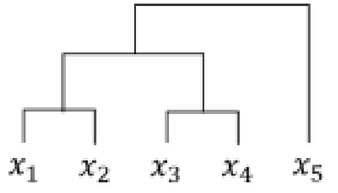
	{A,B}	{C,D}	E
{A,B}	0	2 5	3
{C,D}	2 5	0	4
E	3	4	0

GATE DA 2024

Consider the table below, where the $(i,j)^{th}$ element of the table is the distance between points x_i and x_j. Single linkage clustering is performed on data points, x_1, x_2, x_3, x_4, x_5.

	x_1	x_2	x_3	x_4	x_5
x_1	0	1	4	3	6
x_2	1	0	3	5	3
x_3	4	3	0	2	5
x_4	3	5	2	0	1
x_5	6	3	5	1	0

Which **ONE** of the following is the correct representation of the clusters produced?

(A)	
(B)	
(C)	
(D)	

Dimensionality reduction

Two ways of Dimensionality Reduction

1. Feature Selection
2. Feature Extraction

Feature Selection – 3 Methods

1.Filter Method
- Correlation
- Chi-Square Test, etc.

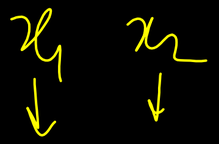
2.Wrapper Method
- Forward Selection
- Backward Selection
- Bi-directional Elimination

3.Embedded Method
- LASSO
- Elastic Net
- Ridge Regression, etc.

Feature Extraction

Feature extraction is the process of transforming the space containing many dimensions into space with fewer dimensions.

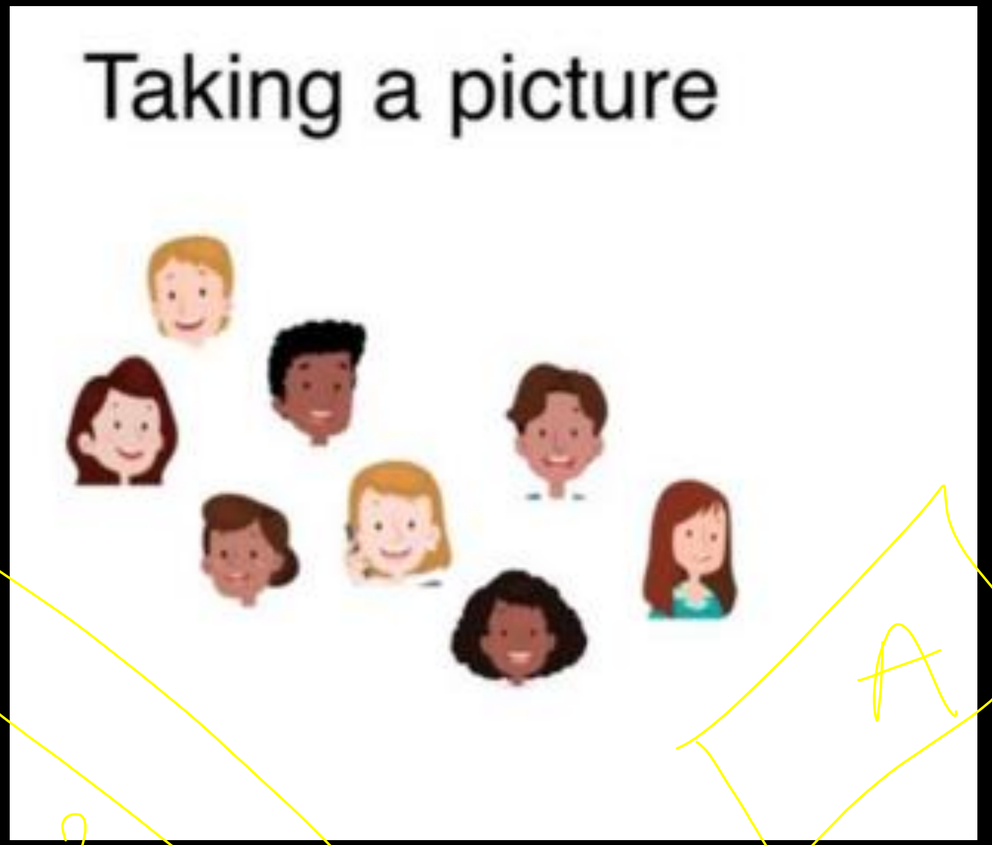
This approach is useful when we want to keep the whole information but use fewer resources while processing the information



Some common feature extraction techniques are:

1. Principal Component Analysis (PCA)
2. Linear Discriminant Analysis (LDA)
3. Kernel PCA
4. Quadratic Discriminant Analysis (QDA)etc

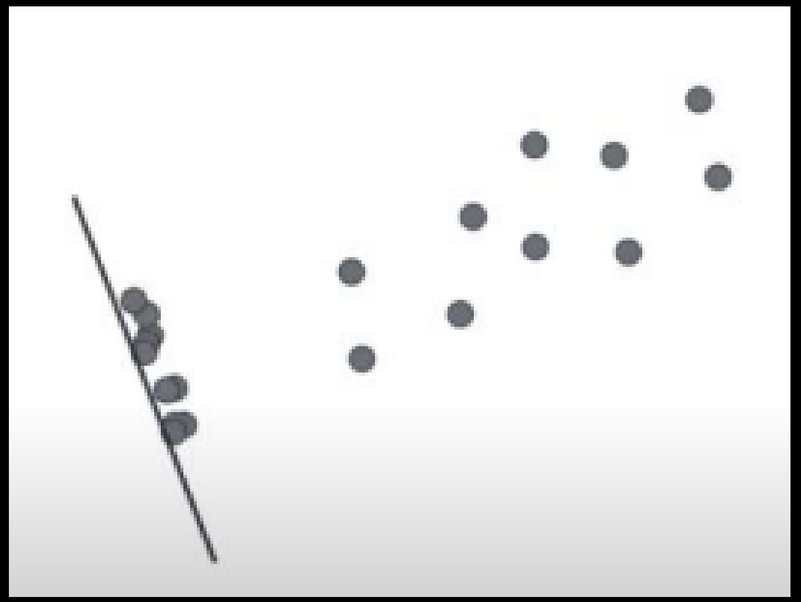
PCA

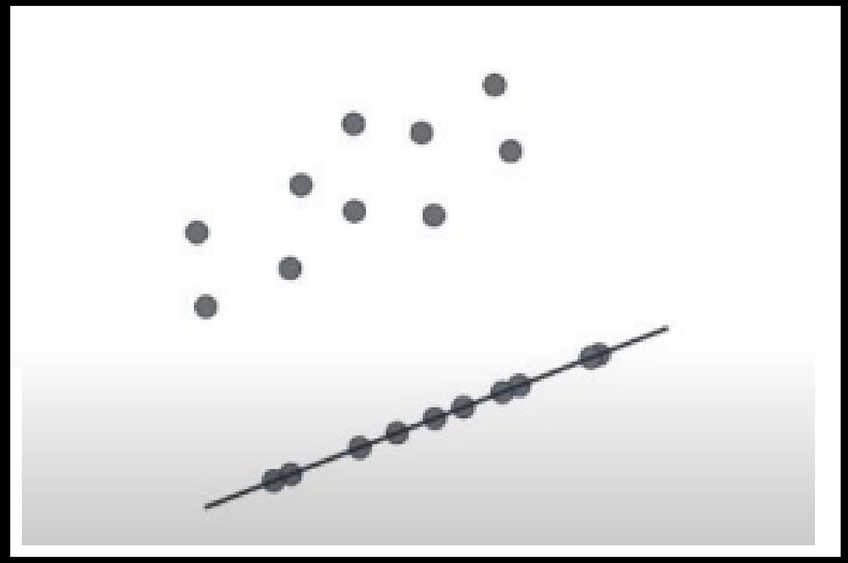


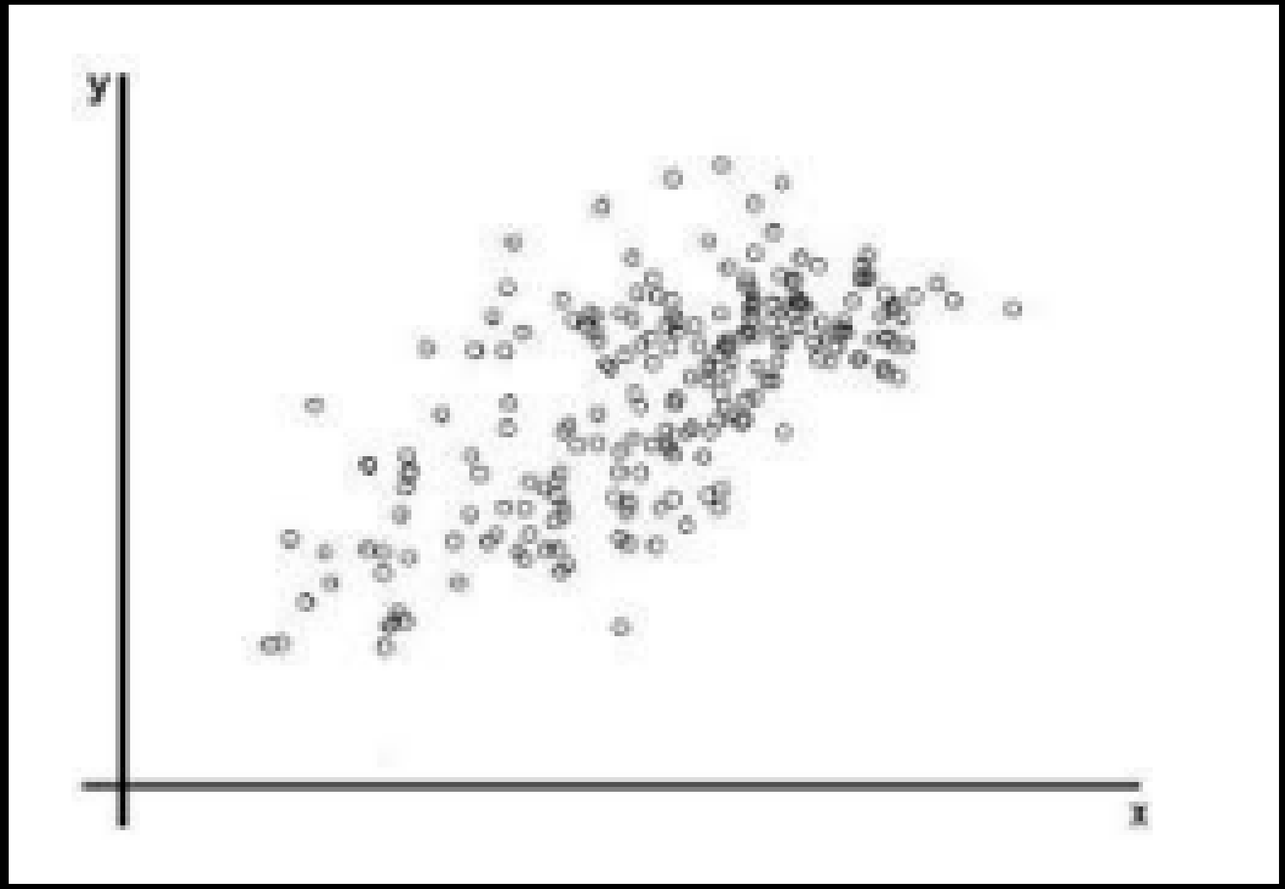

Taking a picture

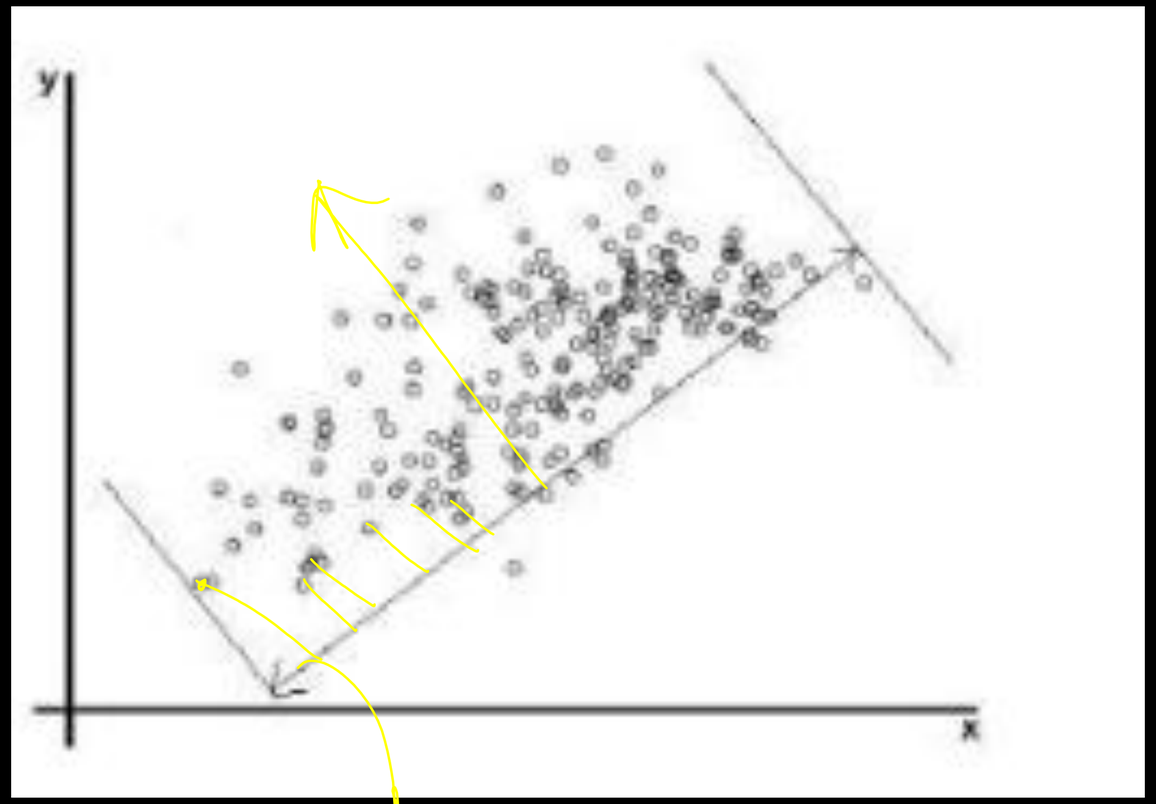
Taking a picture

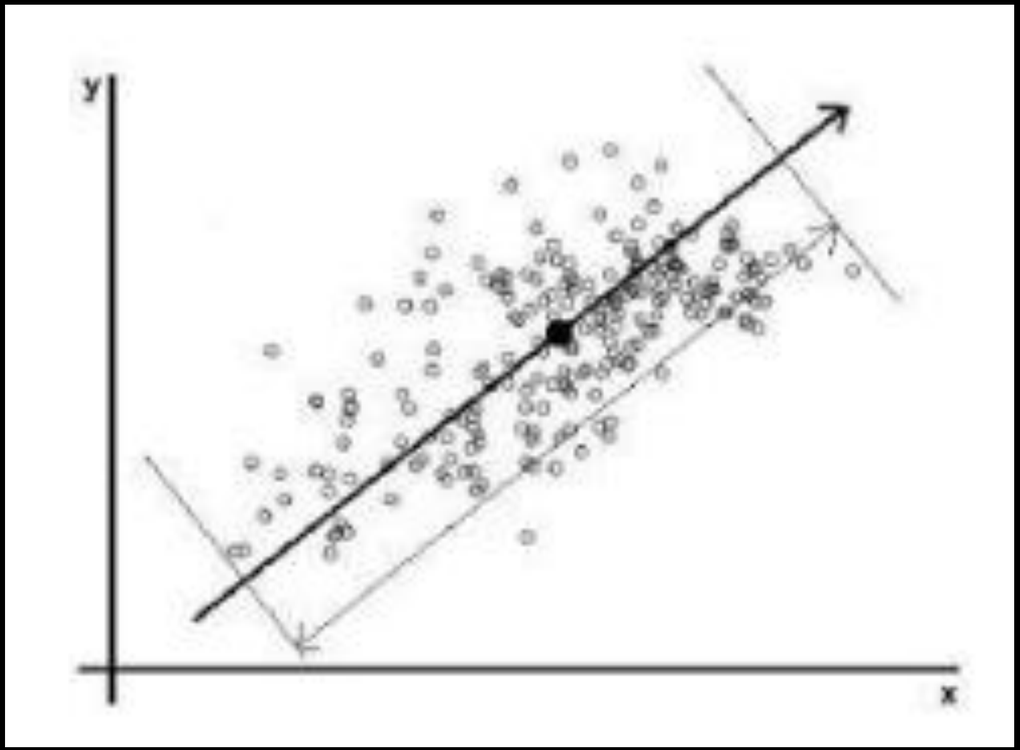

Taking a picture

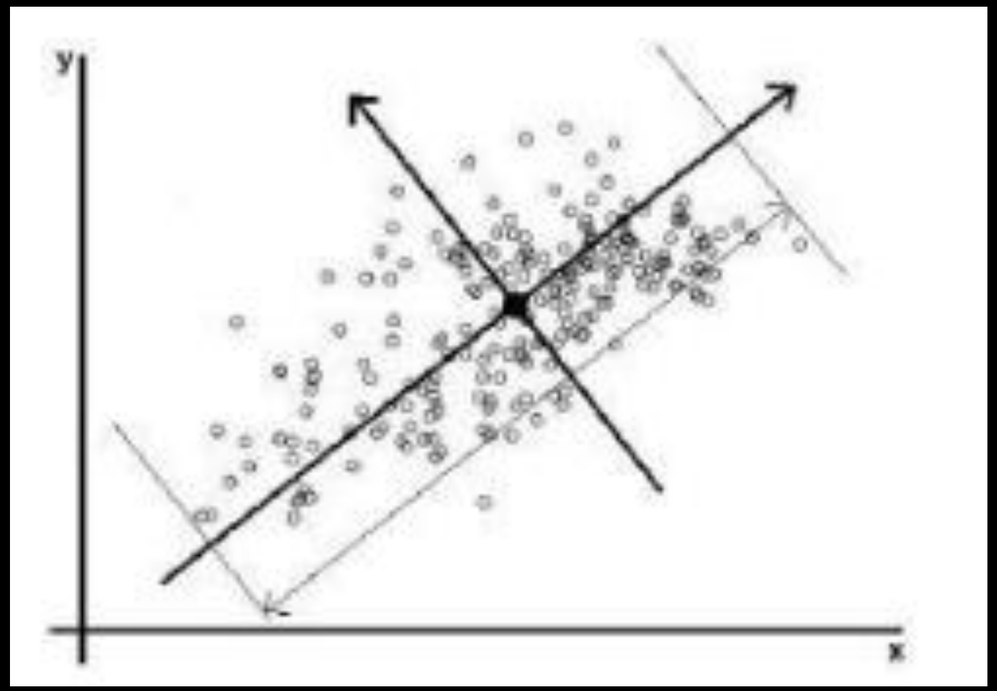


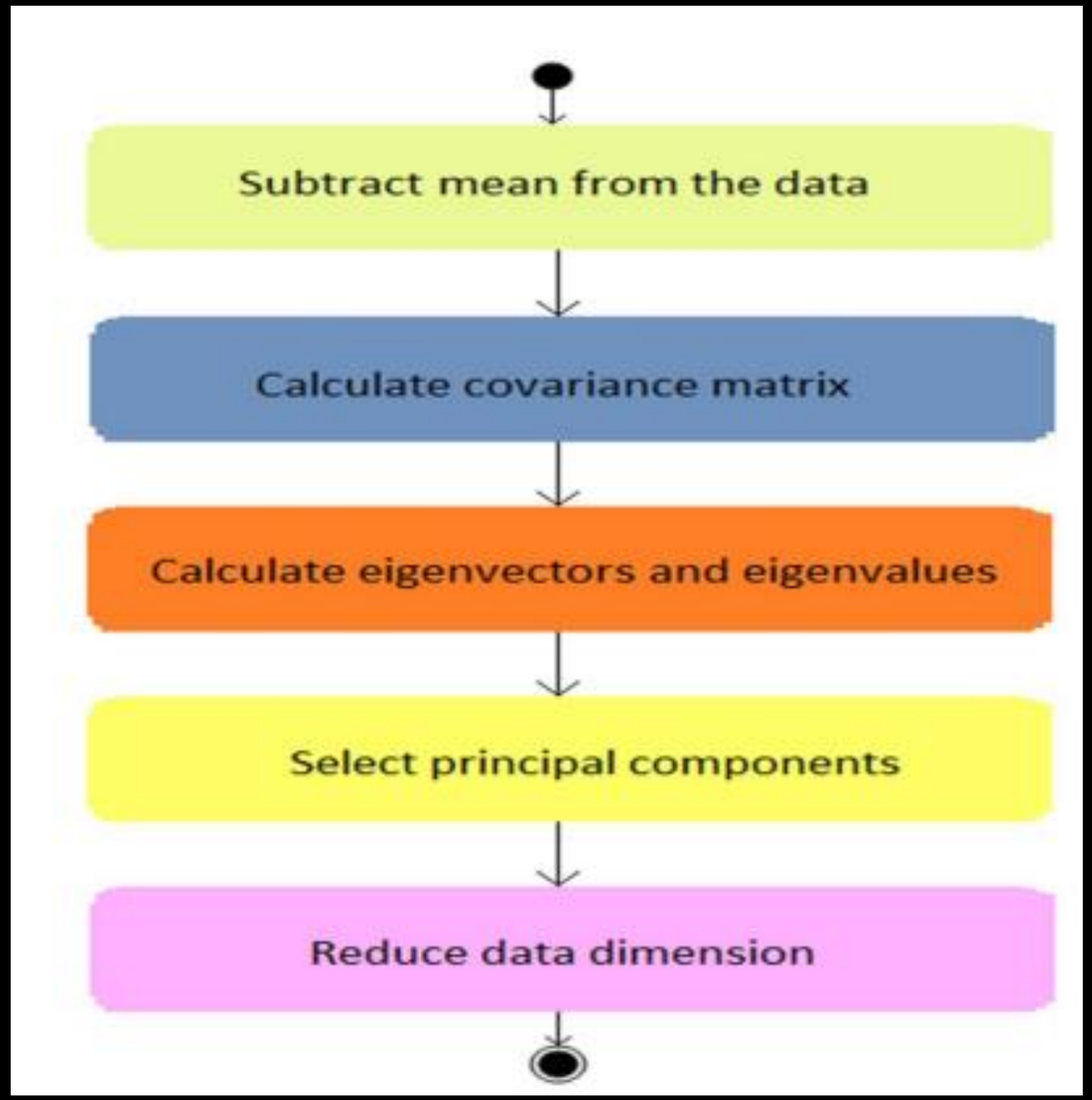



y
x


y
x
IPC8


y
x


y
x


Subtract mean from the data
Calculate covariance matrix
Calculate eigenvectors and eigenvalues
Select principal components
Reduce data dimension

Feature	Example 1	Example 2	Example 3	Example 4
X_1	x_{11} 4	x_{12} 8	x_{13} 13	x_{14} 7
X_2	x_{21} 11	x_{22} 4	x_{23} 5	x_{24} 14

$\bar{X}_1 = 8$

$\bar{X}_2 = 8.5$

$cov(X_1, X_1) = var\, X_1 = \frac{\sum (x_i - \mu)^2}{n} = \frac{(4-8)^2 + (8-8)^2 + (13-8)^2 + (7-8)^2}{4} = 14$

$Cov = \begin{bmatrix} cov(X_1,X_1) & cov(X_1,X_2) \\ cov(X_2,X_1) & cov(X_2,X_2) \end{bmatrix}$

$cov(X_2, X_2) = 23$

$cov(X_1, X_2) = \frac{1}{n-1} \sum (X_{1k} - \bar{X}_1)(X_{2k} - \bar{X}_2) = -11$

$Cov = \begin{bmatrix} 14 & -11 \\ -11 & 23 \end{bmatrix}$ → Eigen vector → unit vector → e_1

$e_1^T \left(\begin{bmatrix} 4-8 & 8-8 & 13-8 & 7-8 \\ 11-8.5 & 4-8.5 & 5-8.5 & 14-8.5 \end{bmatrix} \right)$

The most popularly used dimensionality reduction algorithm is Principal Component Analysis (PCA). Which of the following is/are true about PCA?

1.PCA is an unsupervised method of optimization

2.It searches for the directions that data have the largest variance

3.Maximum number of principal components <= number of features

4.All principal components are orthogonal to each other

Answer: 1,2,3&4

What will happen when eigenvalues are roughly equal?

A. PCA will perform outstandingly

B. PCA will perform badly

C. Can’t say

D. None of above

Solution: (B)

What happens when you get features in lower dimensions using PCA?

1.The features will still have interpretability

2.The features will lose interpretability

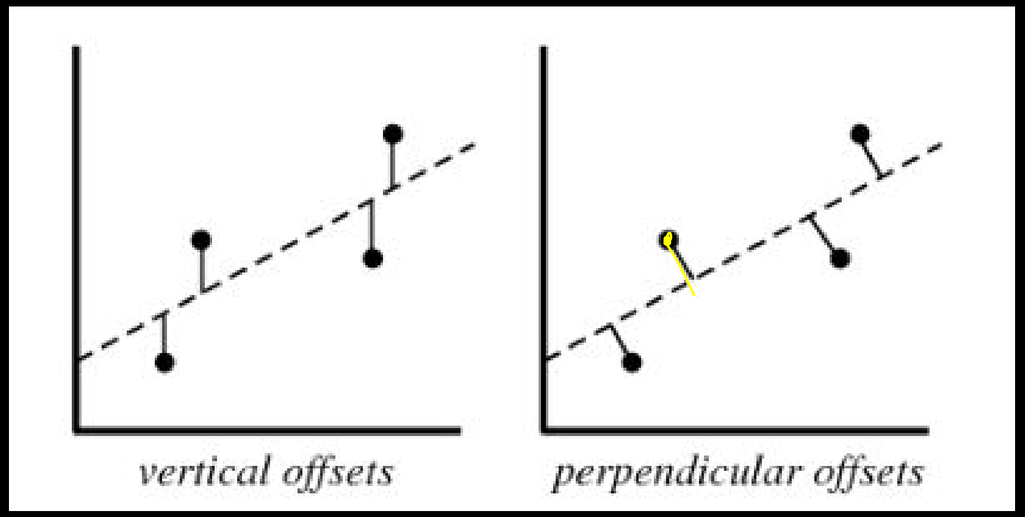
3.The features must carry all information present in the data

4.The features may not carry all information present in the data

Answer: 2 & 4

Which of the following offset do we consider in PCA?

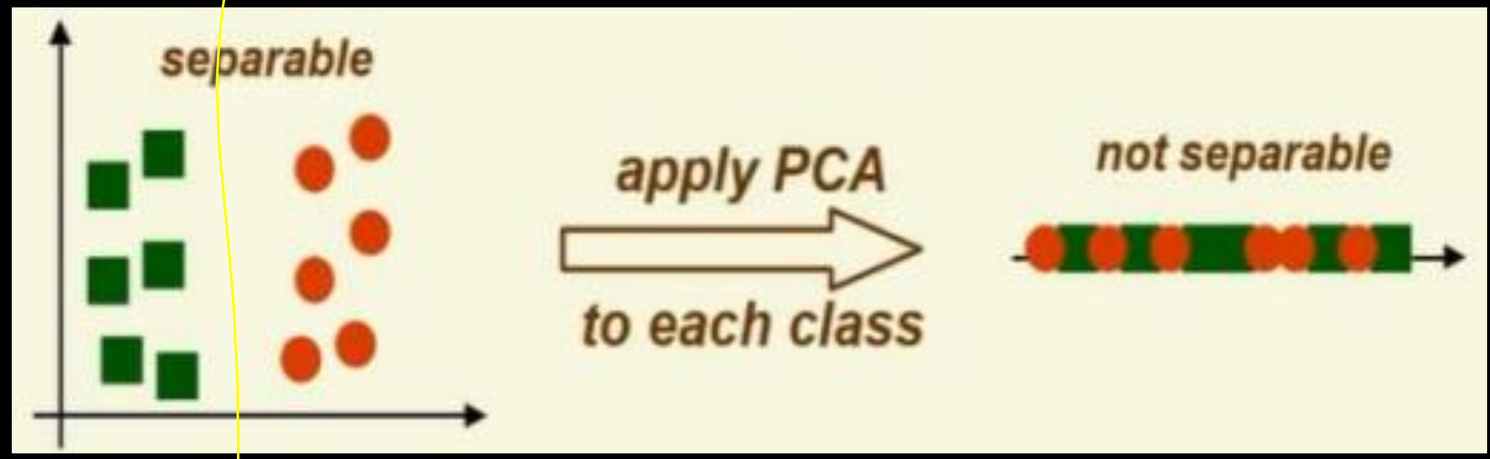
A. Vertical offset
B. Perpendicular offset
C. Both
D. None of these



Solution: (B)

PCA projects the data in the directions of maximum variance.

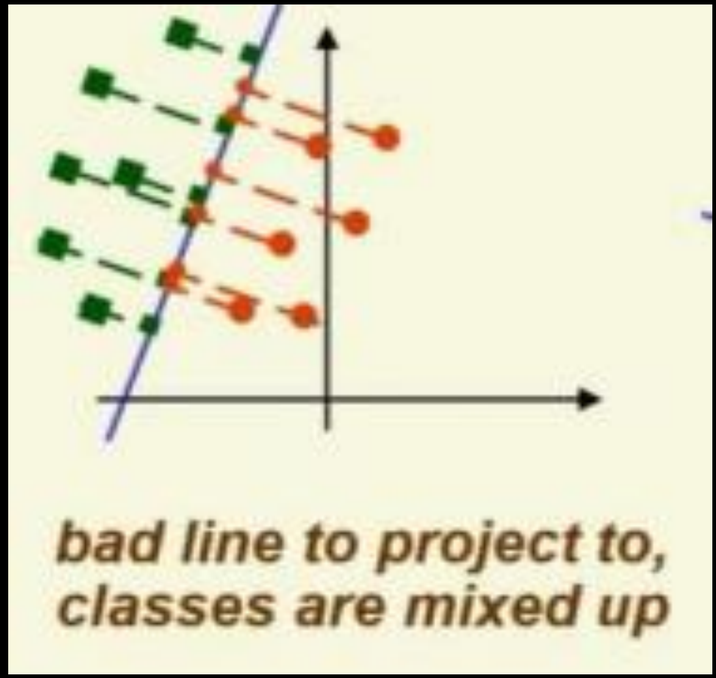
However the directions of maximum variance may be useless for classification



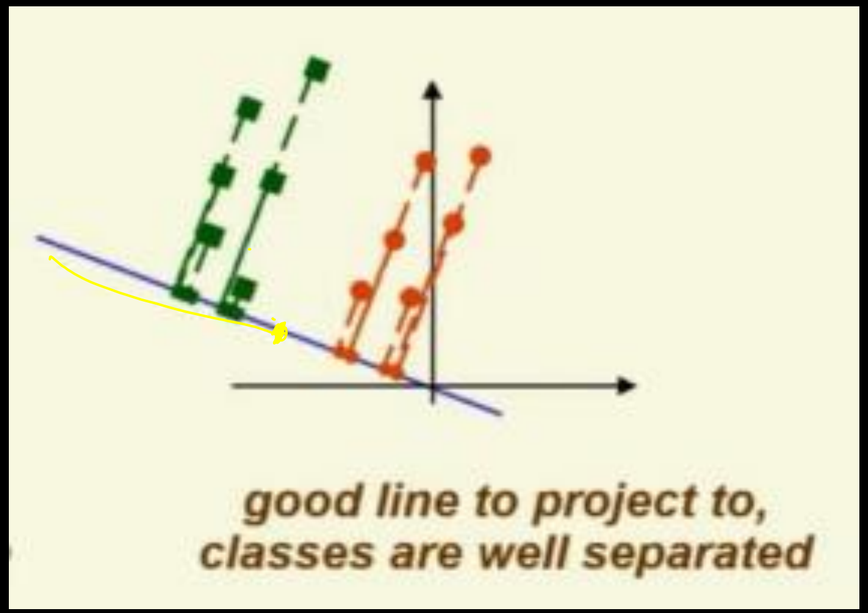
In such condition LDA which is also called as Fisher LDA works well.

Fisher Linear Discriminant

- Main idea: find projection to a line s.t. samples from different classes are well separated

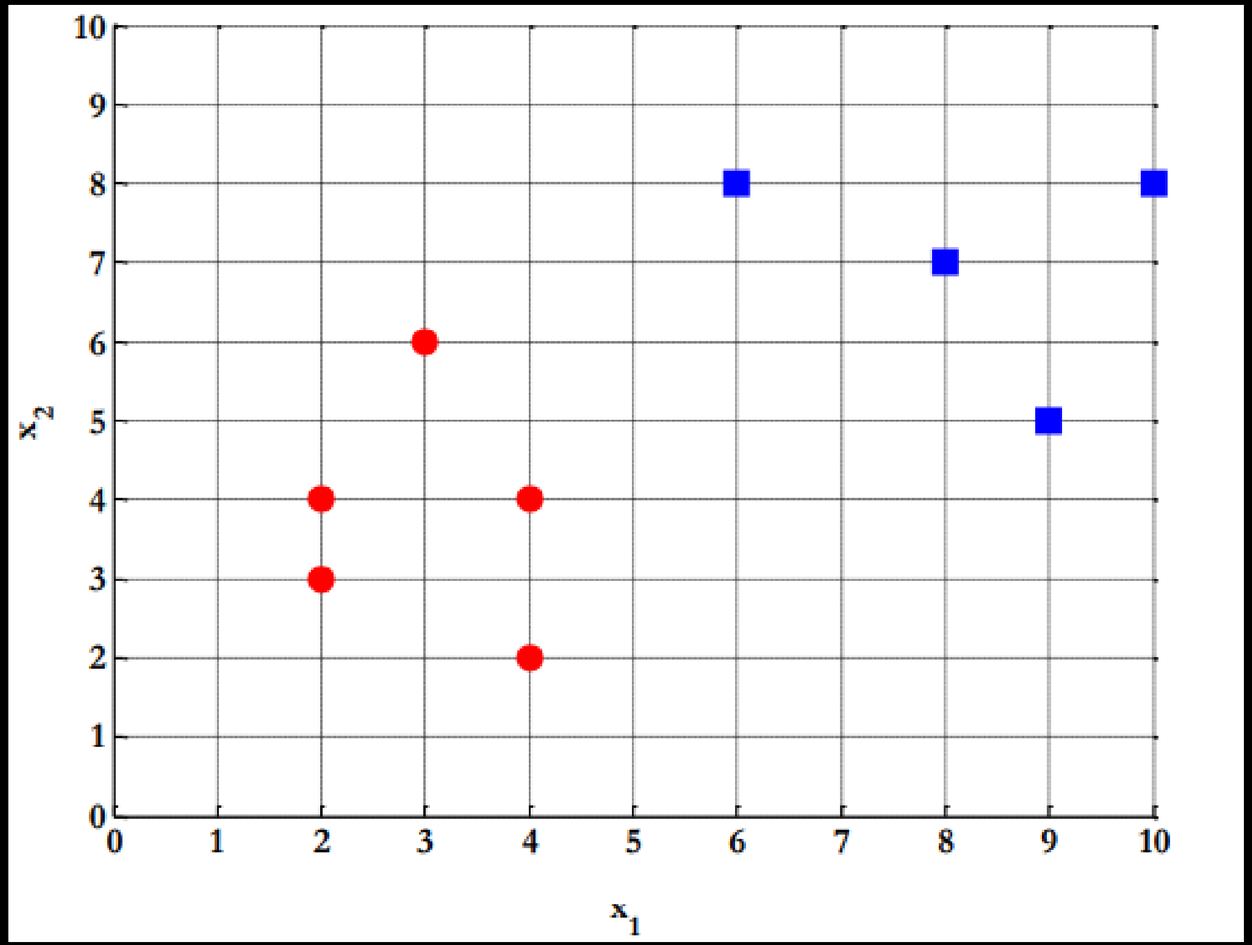


bad line to project to, classes are mixed up


good line to project to,
classes are well separated

Samples for class ω_1 : $\mathbf{X_1}=(x_1,x_2)=\{(4,2),(2,4),(2,3),(3,6),(4,4)\}$

Sample for class ω_2 : $\mathbf{X_2}=(x_1,x_2)=\{(9,10),(6,8),(9,5),(8,7),(10,8)\}$



The classes mean are :

$$\mu_1 = \frac{1}{N_1}\sum_{x\in\omega_1} x = \frac{1}{5}\left[\begin{pmatrix}4\\2\end{pmatrix}+\begin{pmatrix}2\\4\end{pmatrix}+\begin{pmatrix}2\\3\end{pmatrix}+\begin{pmatrix}3\\6\end{pmatrix}+\begin{pmatrix}4\\4\end{pmatrix}\right] = \begin{pmatrix}3\\3.8\end{pmatrix}$$

$$\mu_2 = \frac{1}{N_2}\sum_{x\in\omega_2} x = \frac{1}{5}\left[\begin{pmatrix}9\\10\end{pmatrix}+\begin{pmatrix}6\\8\end{pmatrix}+\begin{pmatrix}9\\5\end{pmatrix}+\begin{pmatrix}8\\7\end{pmatrix}+\begin{pmatrix}10\\8\end{pmatrix}\right] = \begin{pmatrix}8.4\\7.6\end{pmatrix}$$

Covariance matrix of the first class:

$$S_1 = \sum_{x \in \omega_1} (x - \mu_1)(x - \mu_1)^T = \left[\begin{pmatrix} 4 \\ 2 \end{pmatrix} - \begin{pmatrix} 3 \\ 3.8 \end{pmatrix}\right]^2 + \left[\begin{pmatrix} 2 \\ 4 \end{pmatrix} - \begin{pmatrix} 3 \\ 3.8 \end{pmatrix}\right]^2$$

$$+ \left[\begin{pmatrix} 2 \\ 3 \end{pmatrix} - \begin{pmatrix} 3 \\ 3.8 \end{pmatrix}\right]^2 + \left[\begin{pmatrix} 3 \\ 6 \end{pmatrix} - \begin{pmatrix} 3 \\ 3.8 \end{pmatrix}\right]^2 + \left[\begin{pmatrix} 4 \\ 4 \end{pmatrix} - \begin{pmatrix} 3 \\ 3.8 \end{pmatrix}\right]^2$$

$$= \begin{pmatrix} 1 & -0.25 \\ -0.25 & 2.2 \end{pmatrix}$$

Covariance matrix of the second class:

$$S_2 = \sum_{x \in \omega_2} (x - \mu_2)(x - \mu_2)^T = \left[\begin{pmatrix} 9 \\ 10 \end{pmatrix} - \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix}\right]^2 + \left[\begin{pmatrix} 6 \\ 8 \end{pmatrix} - \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix}\right]^2$$

$$+ \left[\begin{pmatrix} 9 \\ 5 \end{pmatrix} - \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix}\right]^2 + \left[\begin{pmatrix} 8 \\ 7 \end{pmatrix} - \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix}\right]^2 + \left[\begin{pmatrix} 10 \\ 8 \end{pmatrix} - \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix}\right]^2$$

$$= \begin{pmatrix} 2.3 & -0.05 \\ -0.05 & 3.3 \end{pmatrix}$$

Within-class scatter matrix:

$$S_w = S_1 + S_2 = \begin{pmatrix} 1 & -0.25 \\ -0.25 & 2.2 \end{pmatrix} + \begin{pmatrix} 2.3 & -0.05 \\ -0.05 & 3.3 \end{pmatrix}$$

$$= \begin{pmatrix} 3.3 & -0.3 \\ -0.3 & 5.5 \end{pmatrix}$$

$$w^* = S_W^{-1}(\mu_1 - \mu_2) = \begin{pmatrix} 3.3 & -0.3 \\ -0.3 & 5.5 \end{pmatrix}^{-1} \left[\begin{pmatrix} 3 \\ 3.8 \end{pmatrix} - \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix} \right]$$

$$= \begin{pmatrix} 0.3045 & 0.0166 \\ 0.0166 & 0.1827 \end{pmatrix} \begin{pmatrix} -5.4 \\ -3.8 \end{pmatrix}$$

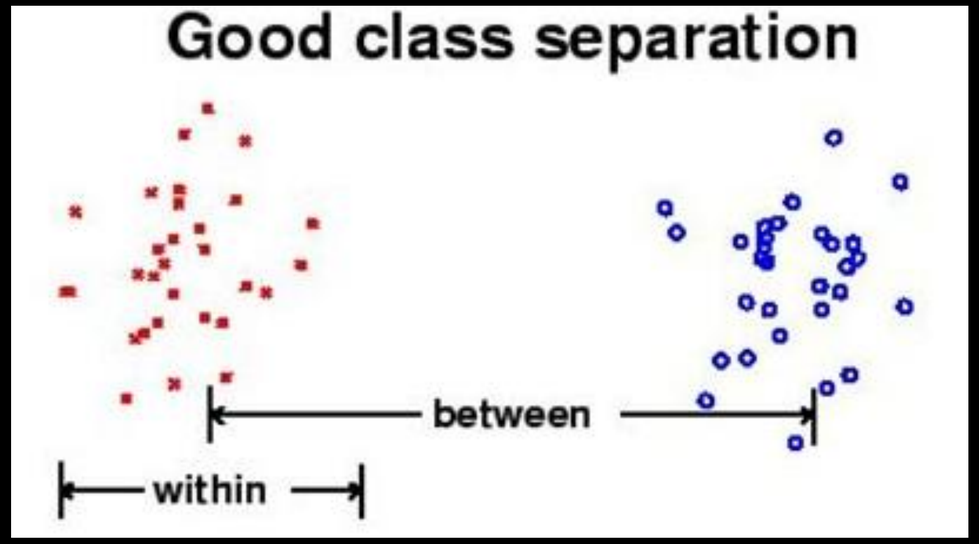
$$= \begin{pmatrix} 0.9088 \\ 0.4173 \end{pmatrix}$$

Samples for class ω_1 : $\mathbf{X_1}=(x_1,x_2)=\{(4,2),(2,4),(2,3),(3,6),(4,4)\}$

Sample for class ω_2 : $\mathbf{X_2}=(x_1,x_2)=\{(9,10),(6,8),(9,5),(8,7),(10,8)\}$

X1	4	2	2	3	4	9	6	9	8	10
X2	2	4	3	6	4	10	8	5	7	8
1st LD	4.46	3.48	3.06	5.2	5.3	12.35	8.8	10.2	10.19	12.42

Which of the following is true about LDA?



A. LDA aims to maximize the distance between classes and minimize the within-class distance.
B. LDA aims to minimize both distances between classes and the distance within the class.
C. LDA aims to minimize the distance between classes and maximize the distance within the class.
D. LDA aims to maximize both distances between classes and the distance within the class.

Solution: (A)

Which of the following comparison(s) are true about PCA and LDA?

1.Both LDA and PCA are linear transformation techniques

2.LDA uses supervised learning, whereas PCA uses unsupervised learning

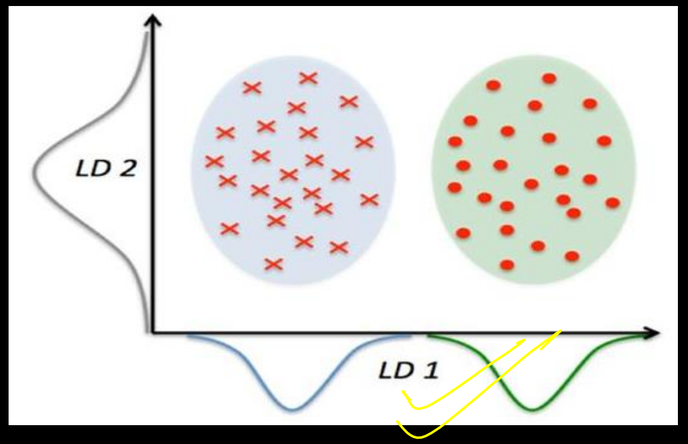
3.PCA maximizes the variance of the data, whereas LDA maximizes the separation between different classes.

4. None of these

Answer 1,2,&3

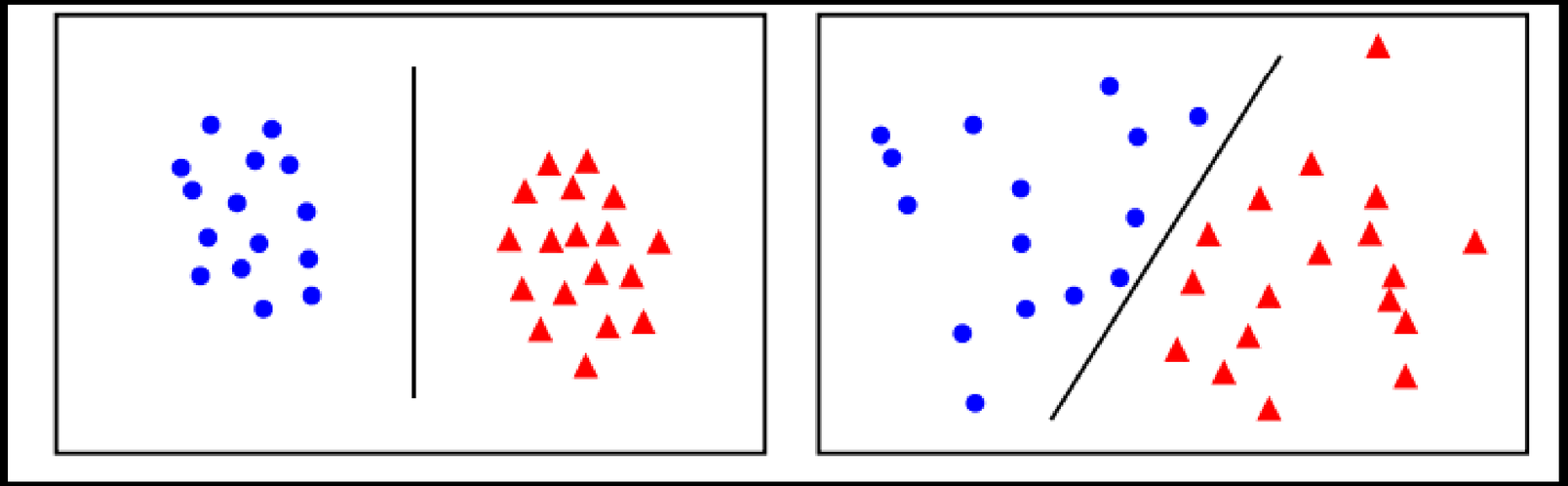
In LDA, the idea is to find the line that best separates the two classes. In the given image, which of the following is a good projection?

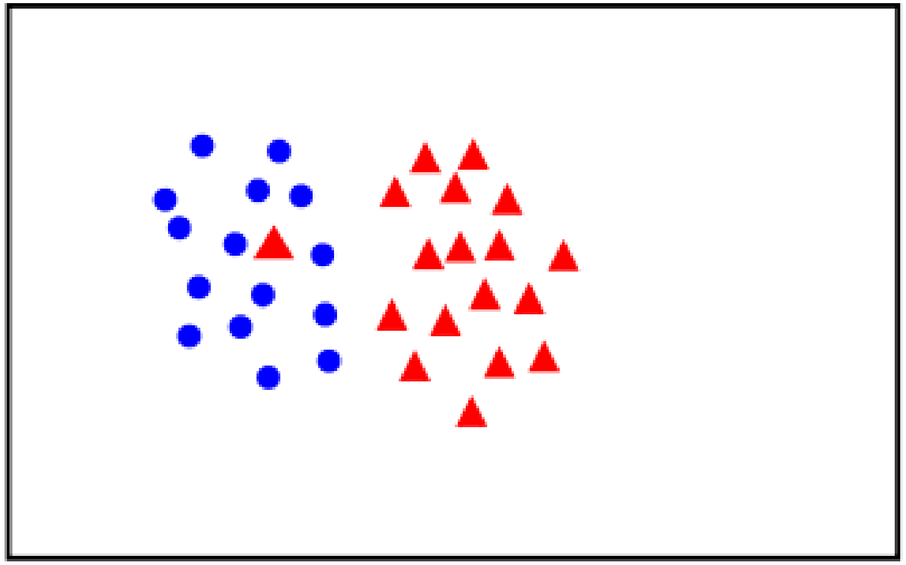
A. LD1
B. LD2
C. Both
D. None of these

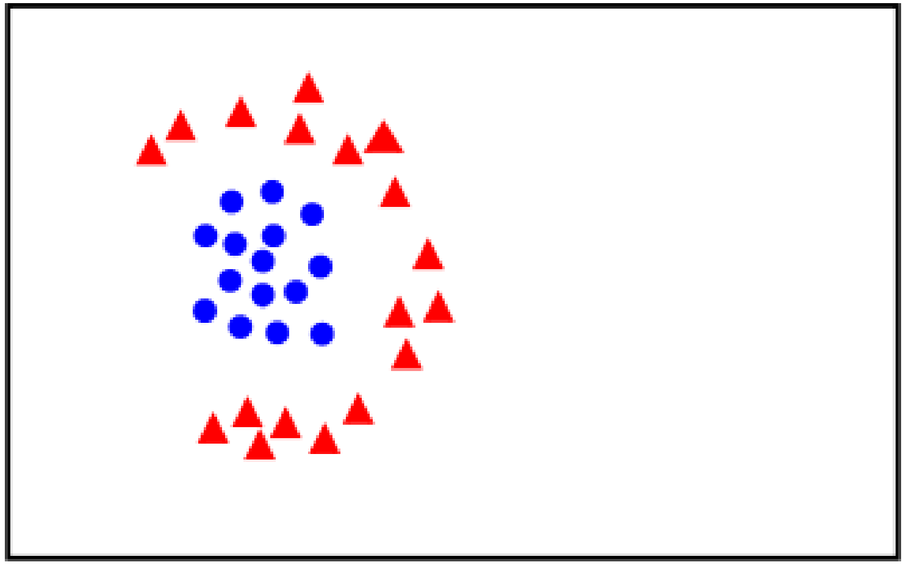


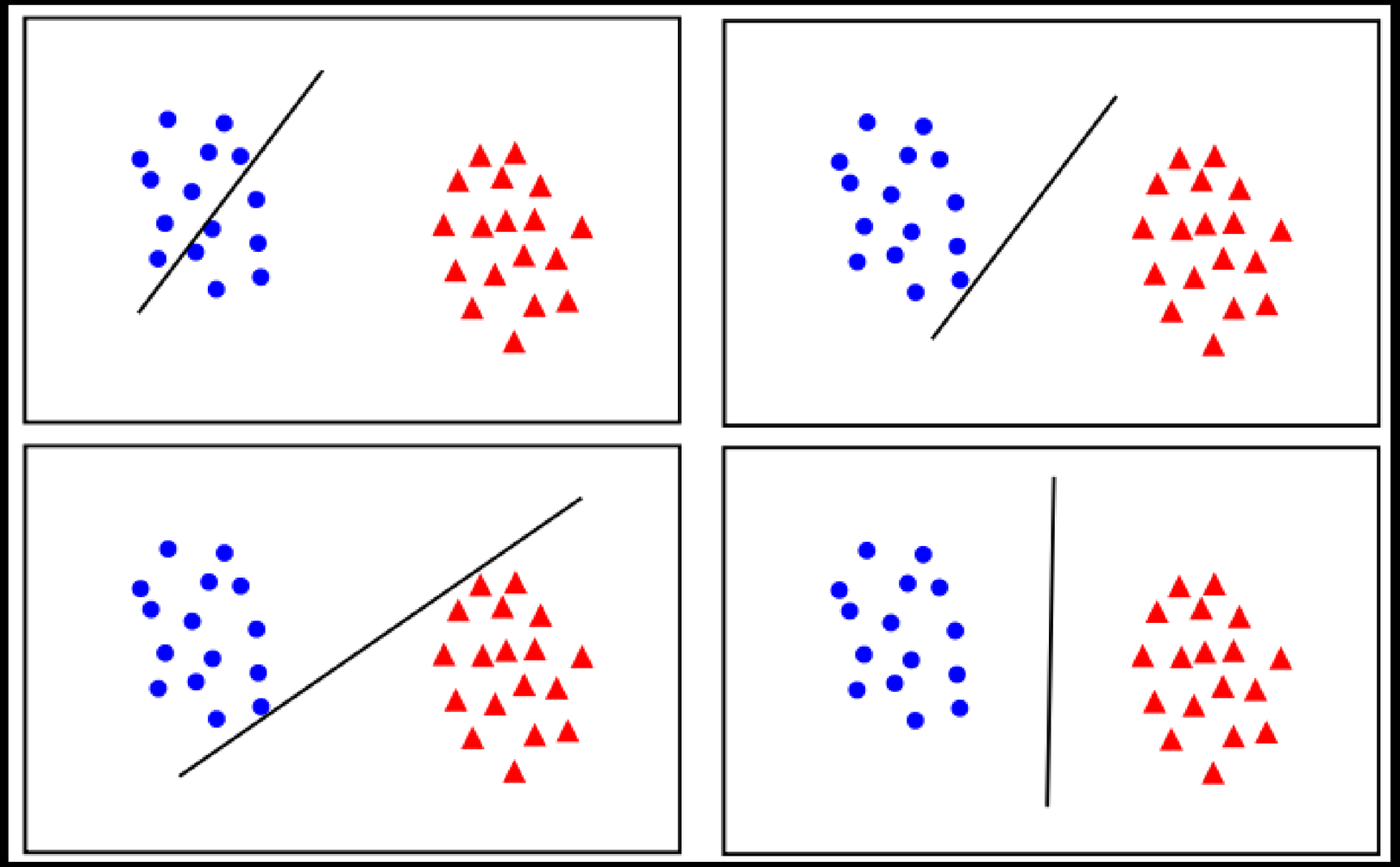
Solution: (A)

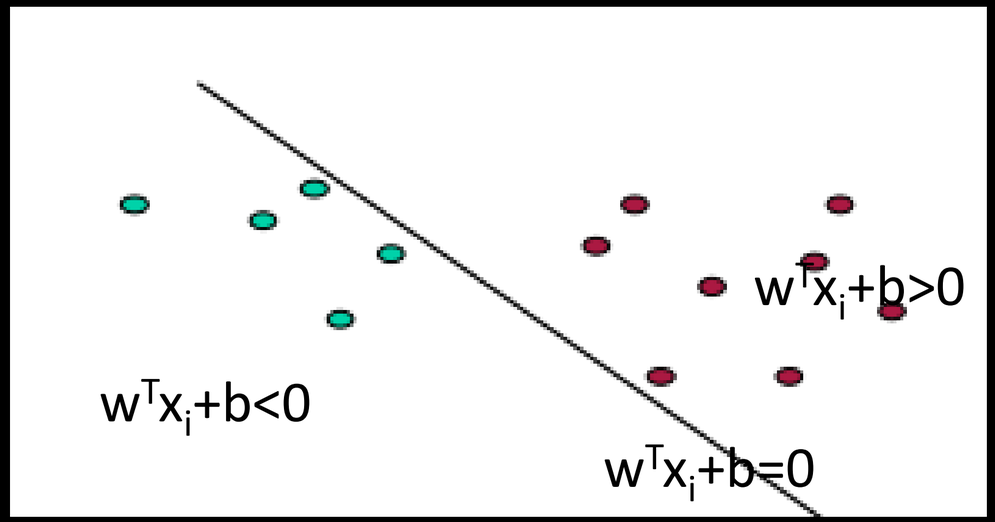
SVM

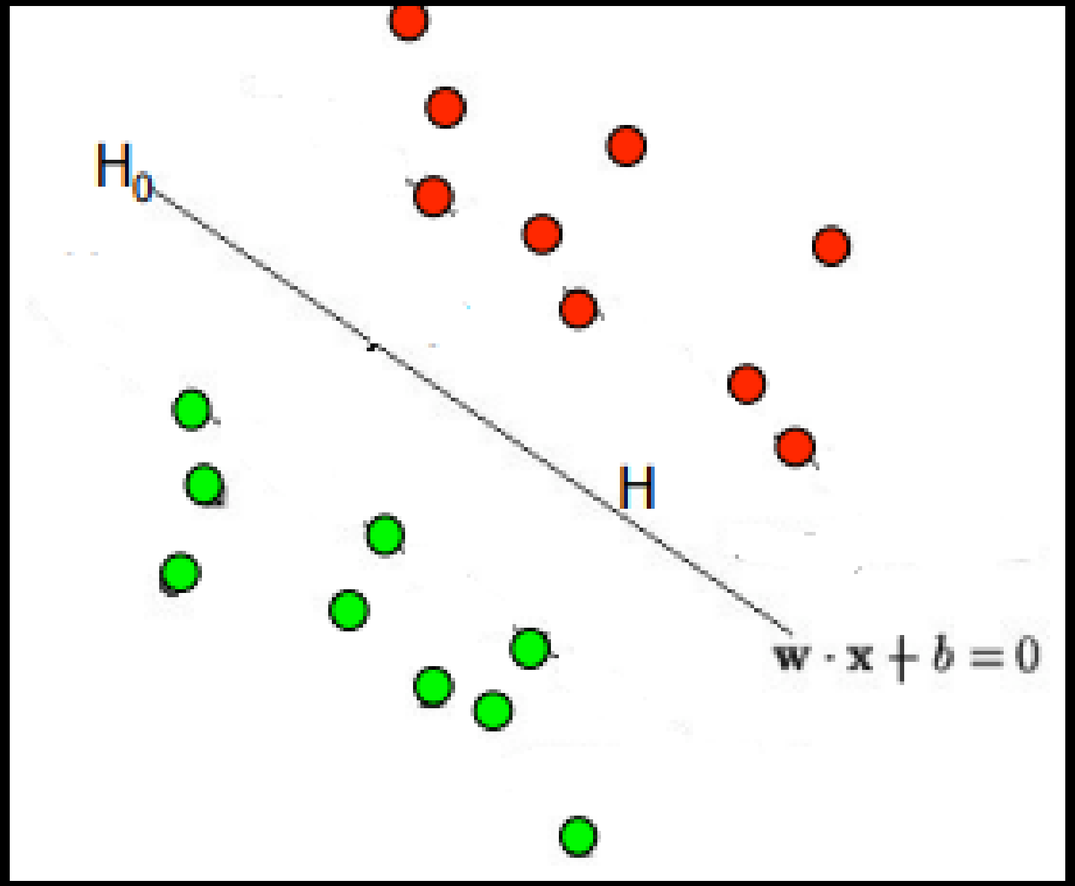


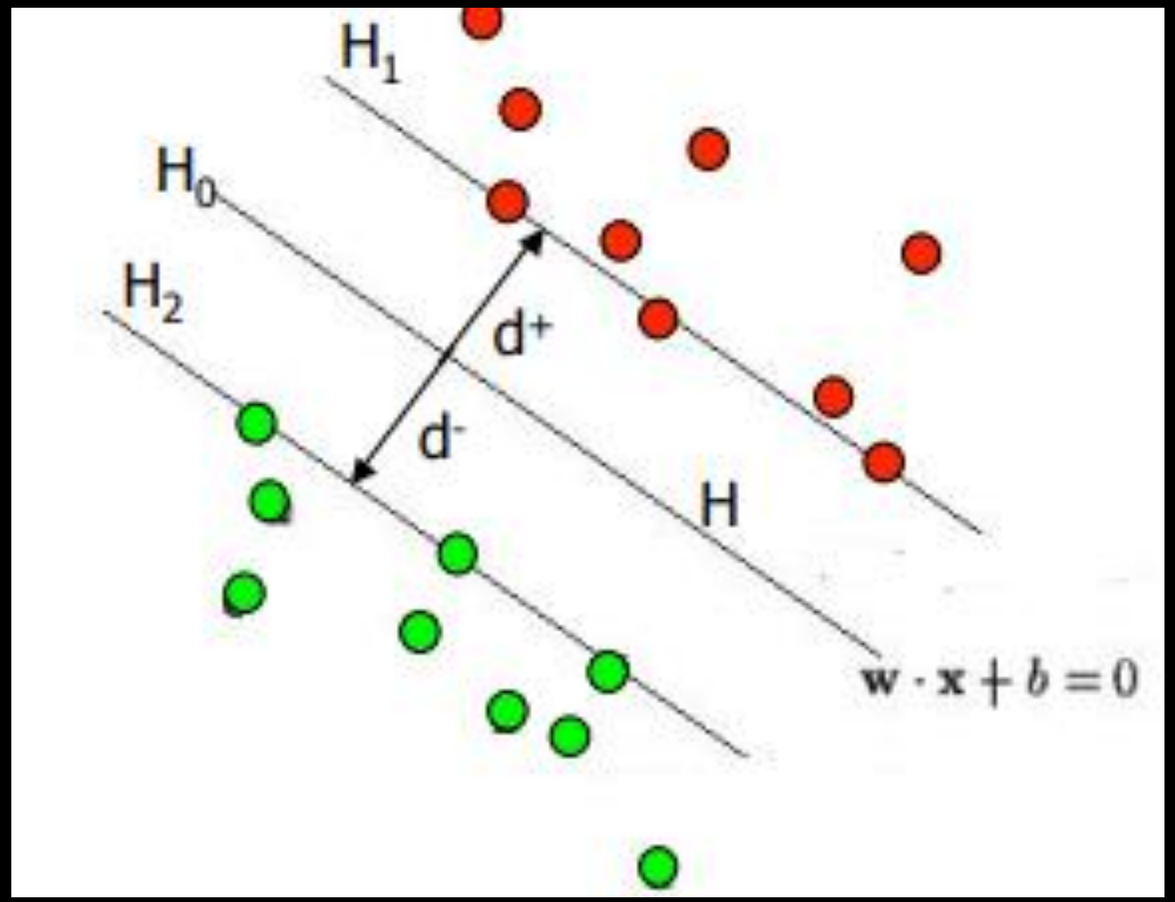


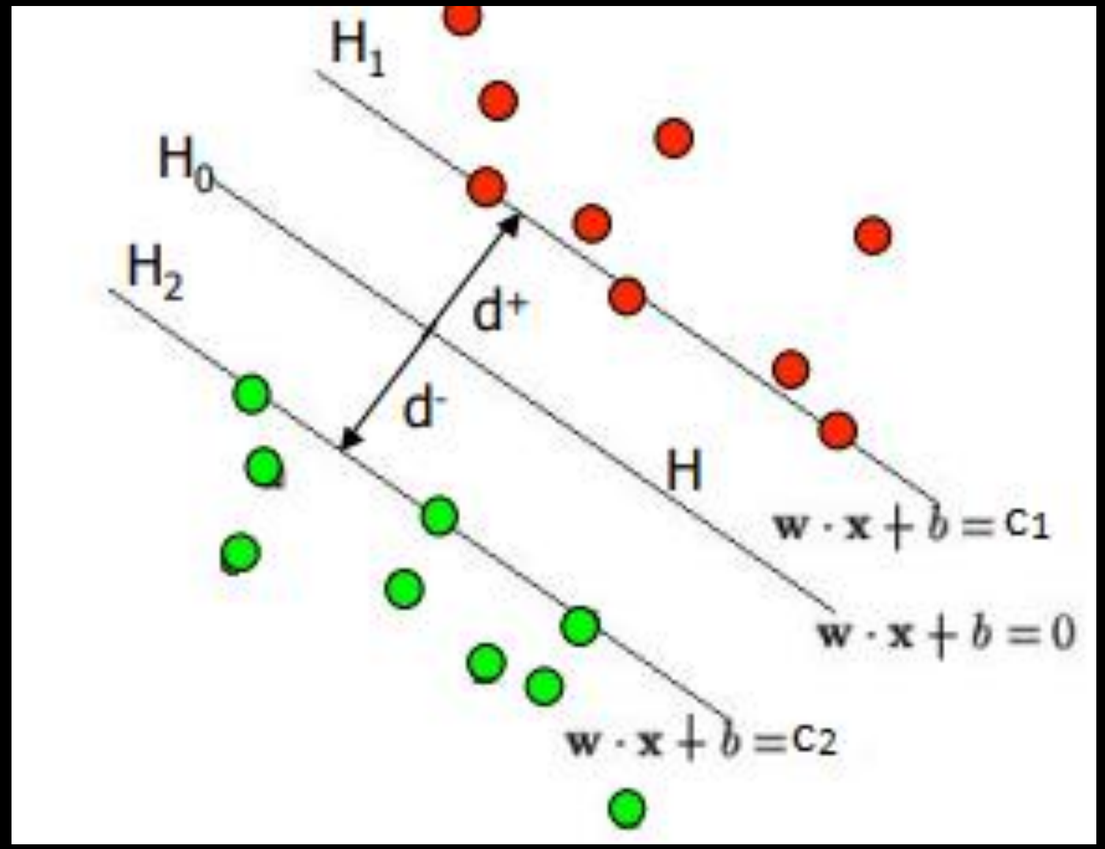


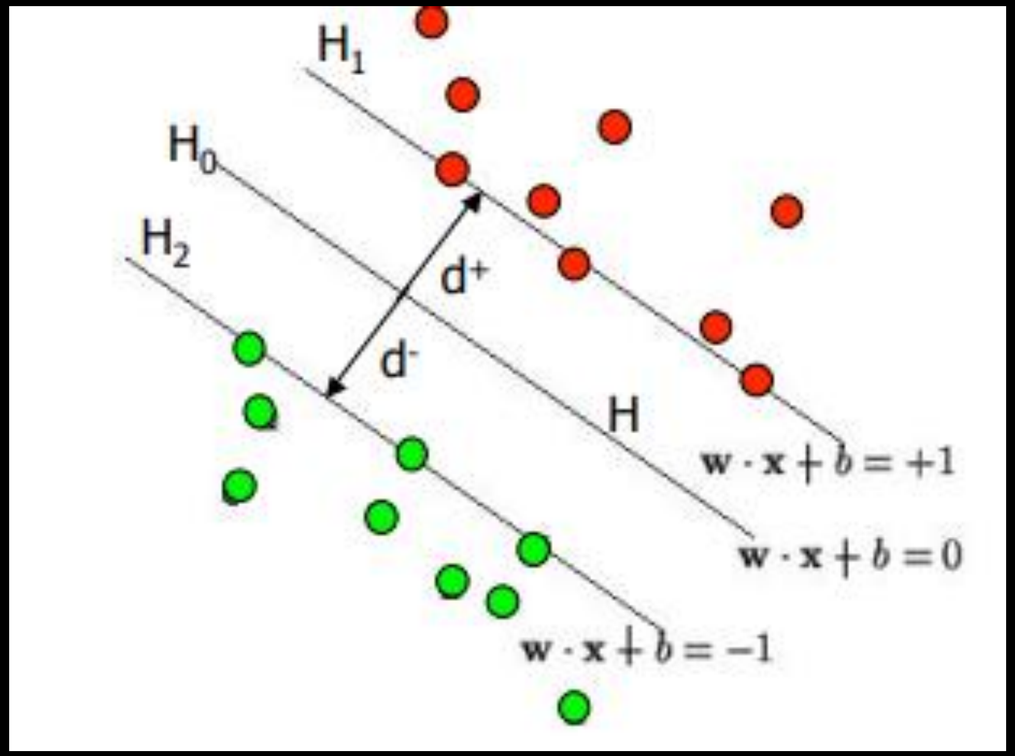



$w^T x_i + b > 0$
$w^T x_i + b < 0$
$w^T x_i + b = 0$


H0
H
w · x + b = 0

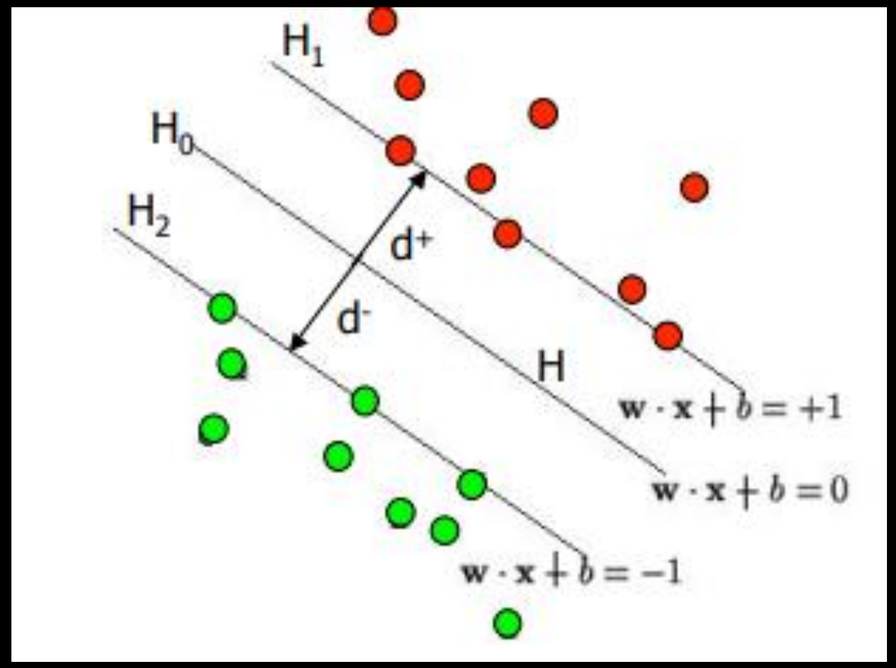

H1
H0
H2
d+
d-
H
w · x + b = 0


H1
H0
H2
d+
d-
H
w · x + b = c1
w · x + b = 0
w · x + b = c2


H1
H0
H2
d+
d-
H
w · x + b = +1
w · x + b = 0
w · x + b = −1

Recall the distance from a point(x_0,y_0) to a line: Ax+By+c = 0 is:

$$\frac{|Ax_0 + By_0 + c|}{\sqrt{A^2 + B^2}}$$



The distance between H0 and H1 is then:

$$\frac{|wx + b|}{||w||} = \frac{1}{||w||}$$

The distance between H2 and H1 is $\frac{2}{||w||}$

Optimization Problem of SVM

To summarize, the SVM optimization problem can be written as:

Objective:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2$$

Subject to the constraints:

$$y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1, \quad \forall i$$

$$\underset{\boldsymbol{w},b}{\text{minimize}} \ \frac{1}{2}\|\boldsymbol{w}\|^2$$

$$\text{subject to} \quad y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) - 1 \geq 0, \quad 1 \leq i \leq n.$$

The vector $\vec{w}$ and the scalar b are given by

$$\vec{w} = \sum_{i=1}^{N} \alpha_i y_i \vec{x}_i$$

$$b = \frac{1}{2}\left(\min_{i:y_i=+1} (\vec{w} \cdot \vec{x}_i) + \max_{i:y_i=-1} (\vec{w} \cdot \vec{x}_i) \right)$$

where $\vec{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_N)$ is a vector which maximizes

$$\sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1,j=1}^{N} \alpha_i \alpha_j y_i y_j (\vec{x}_i \cdot \vec{x}_j)$$

subject to

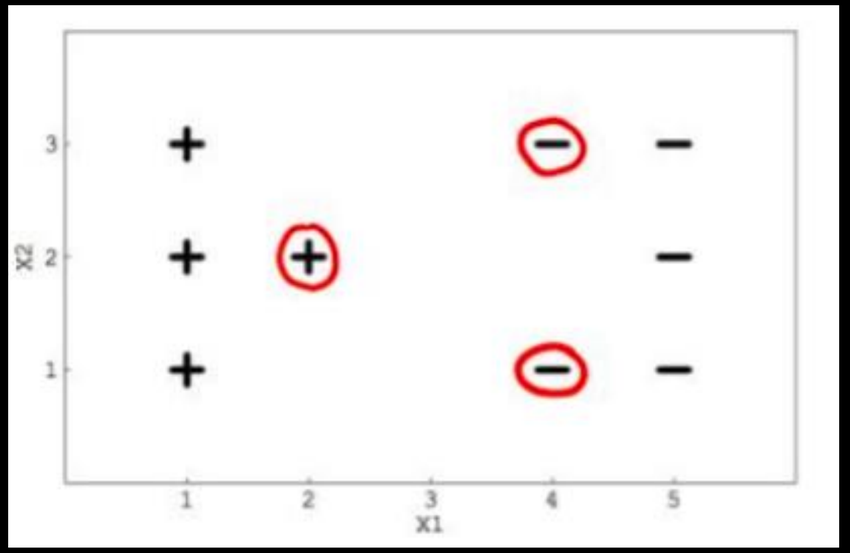
$$\sum_{i=1}^{N} \alpha_i y_i = 0$$

$$\alpha_i > 0 \text{ for } i = 1, 2, \ldots, N.$$

Using the SVM algorithm, find the SVM classifier for the follwoing data.

Example no.	x_1	x_2	Class
1	2	1	+1
2	4	3	−1

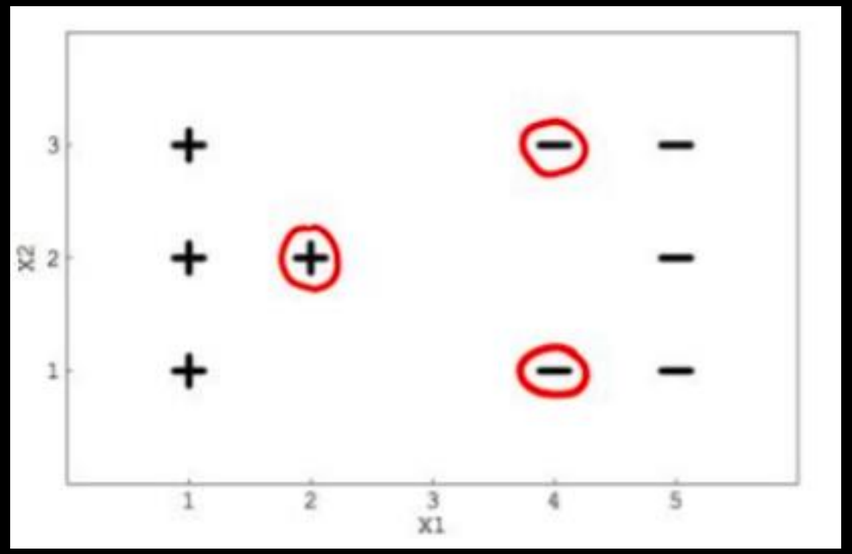
Suppose you are using a Linear SVM classifier with 2 class classification problem. Consider the following data in which the points circled red represent support vectors.



Will the decision boundary change if any of the red points are removed?

a. Yes
b. No

Suppose you are using a Linear SVM classifier with 2 class classification problem. Consider the following data in which the points circled red represent support vectors.



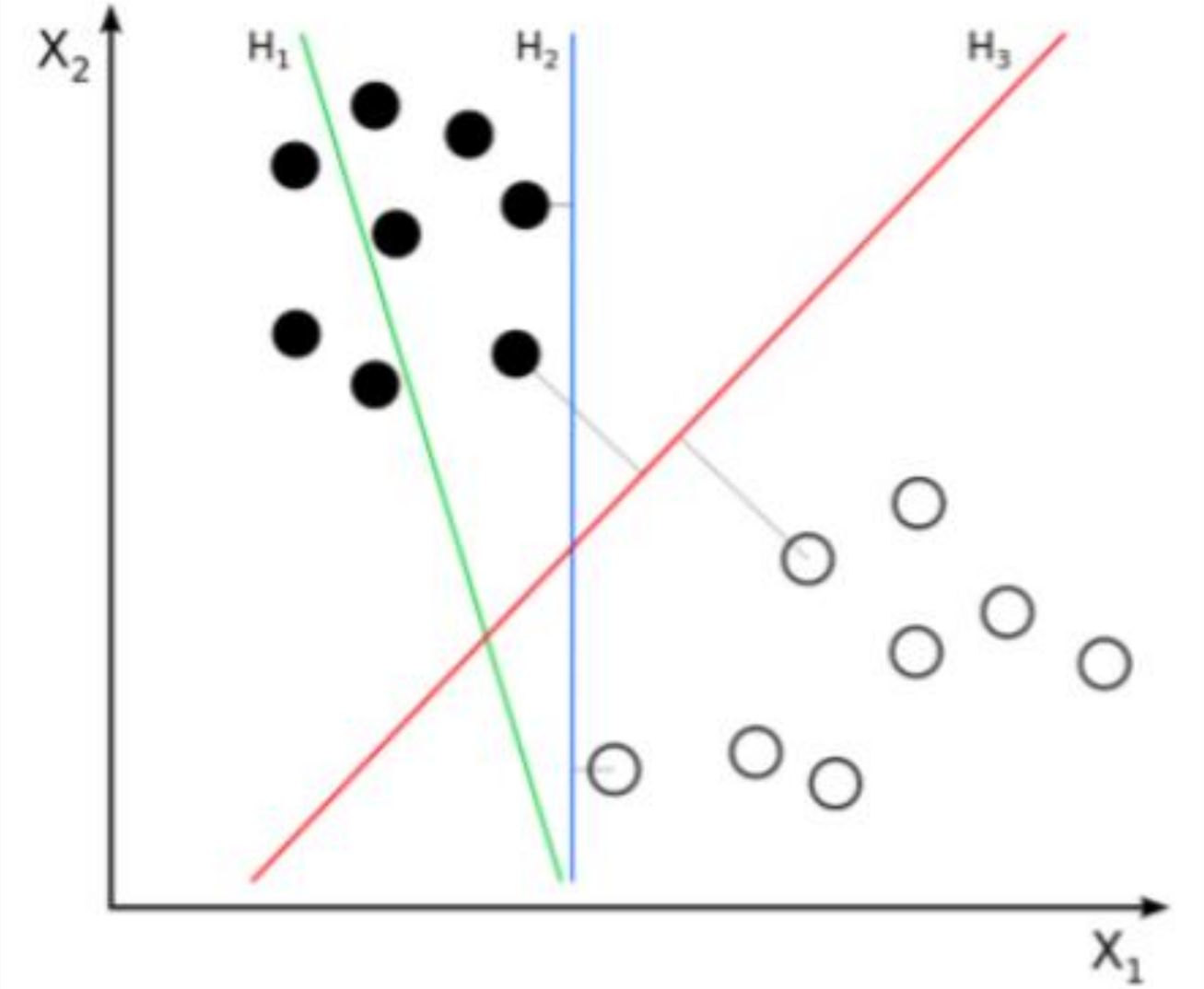
If you remove the non-red circled points from the data, the decision boundary will change?

A) True

B) False

Solution: B

Consider the data-points in the figure below.



Let us assume that the black-colored circles represent positive class whereas the white-colored circles represent negative class. Which of the following among H1, H2 and H3 is the maximum-margin hyperplane?

(a) H1

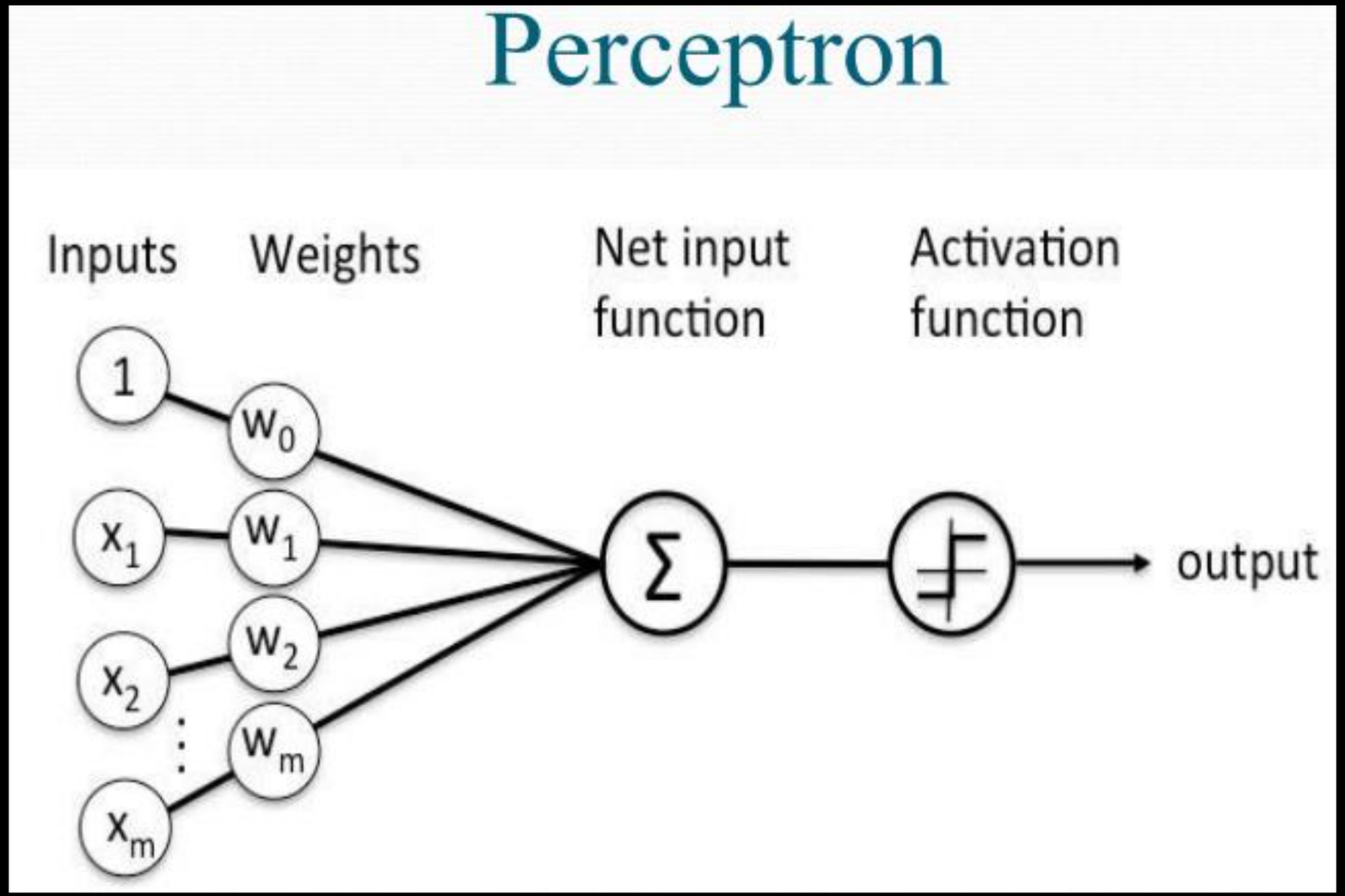
(b) H2

(c) H3

The soft margin SVM is more preferred than the hard-margin svm when:

1. The data is linearly separable
2. The data is noisy and contains overlapping point

Perceptron

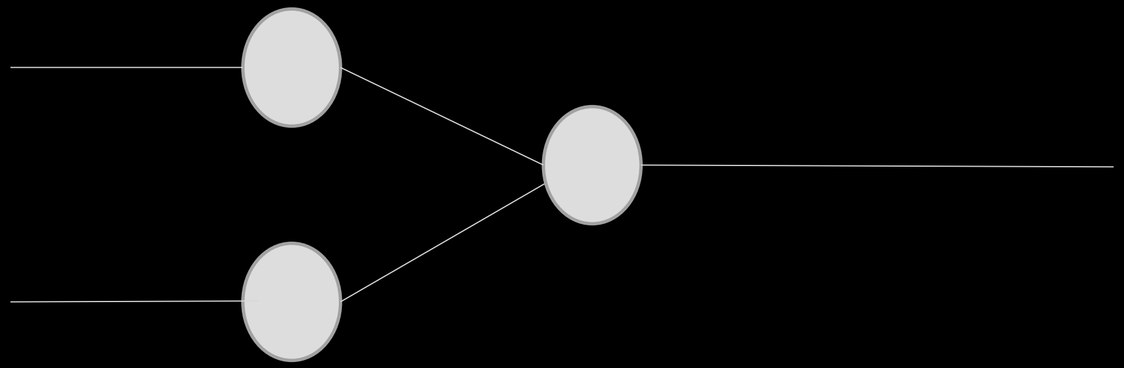

Perceptron
Inputs
Weights
Net input function
Activation function
1
x_1
x_2
x_m
w_0
w_1
w_2
w_m
Σ
output

Forward Propagation

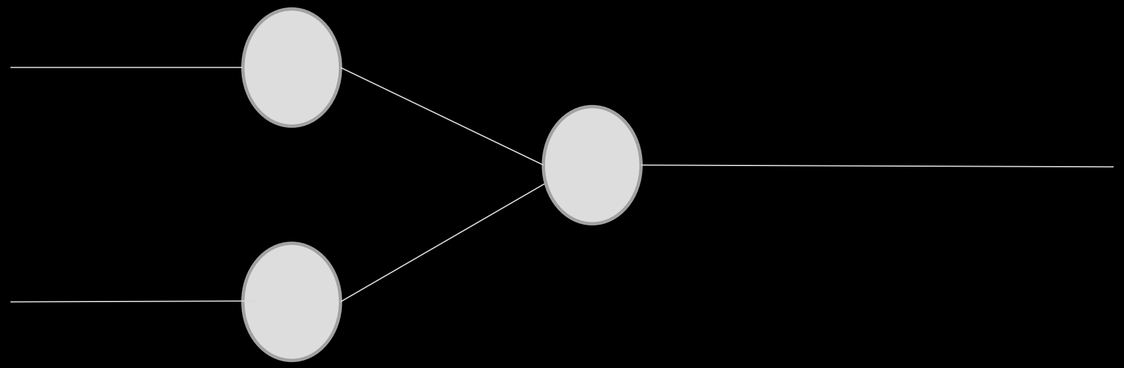
Activation Function

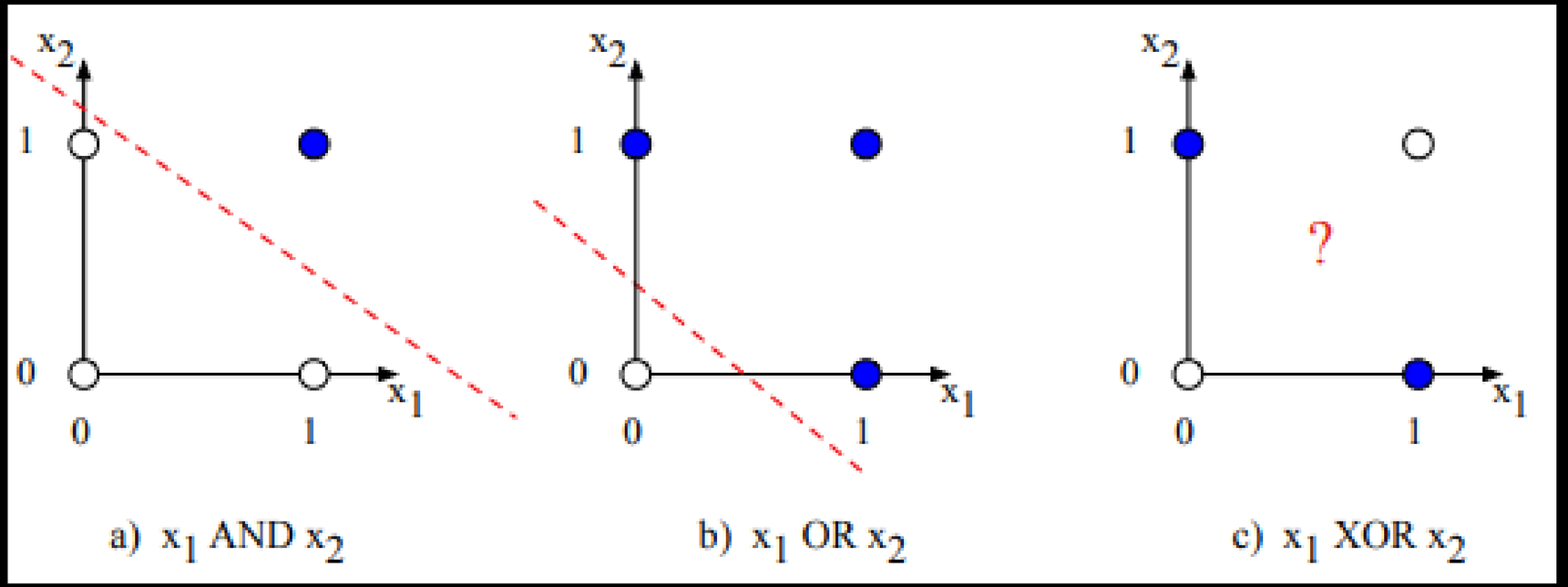
1. Sigmoid Function
2. Threshold Activation Function
3. Unit Step Function
4. Relu
5. Linear Activation Function
6. Hyperbolic tangent function

AND		
x1	x2	y
0	0	0
0	1	0
1	0	0
1	1	1

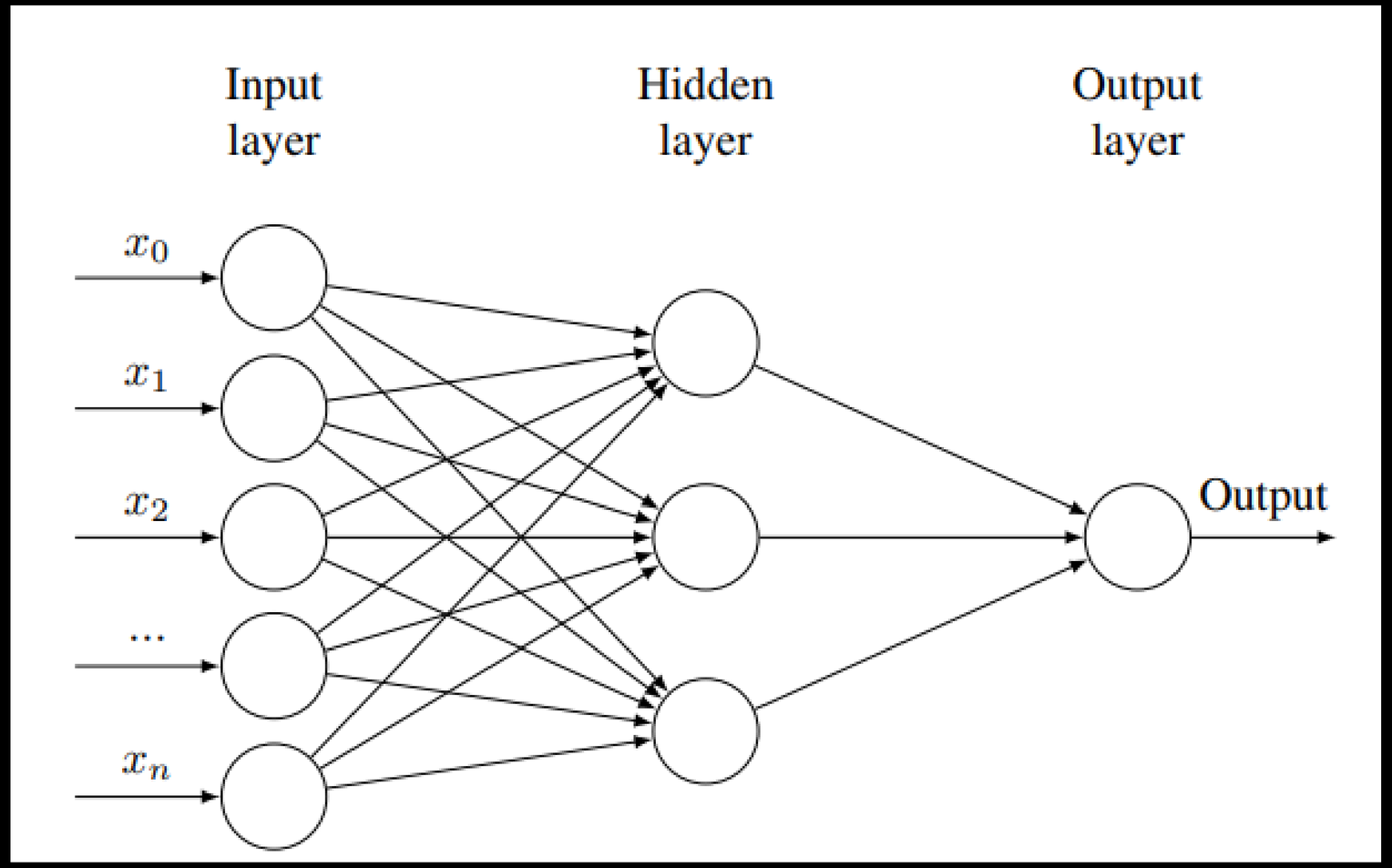


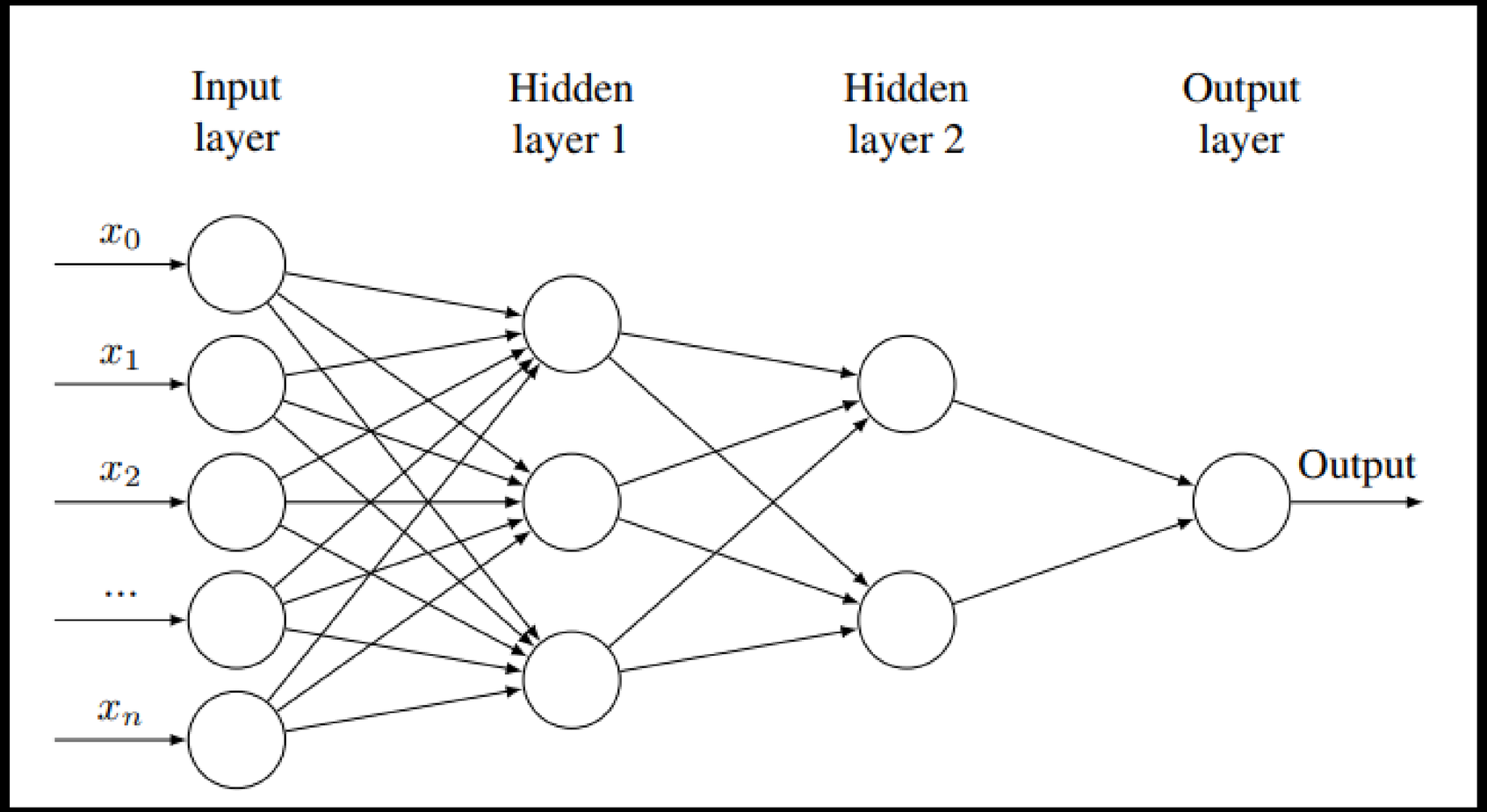
XOR		
x1	x2	y
0	0	0
0	1	1
1	0	1
1	1	0



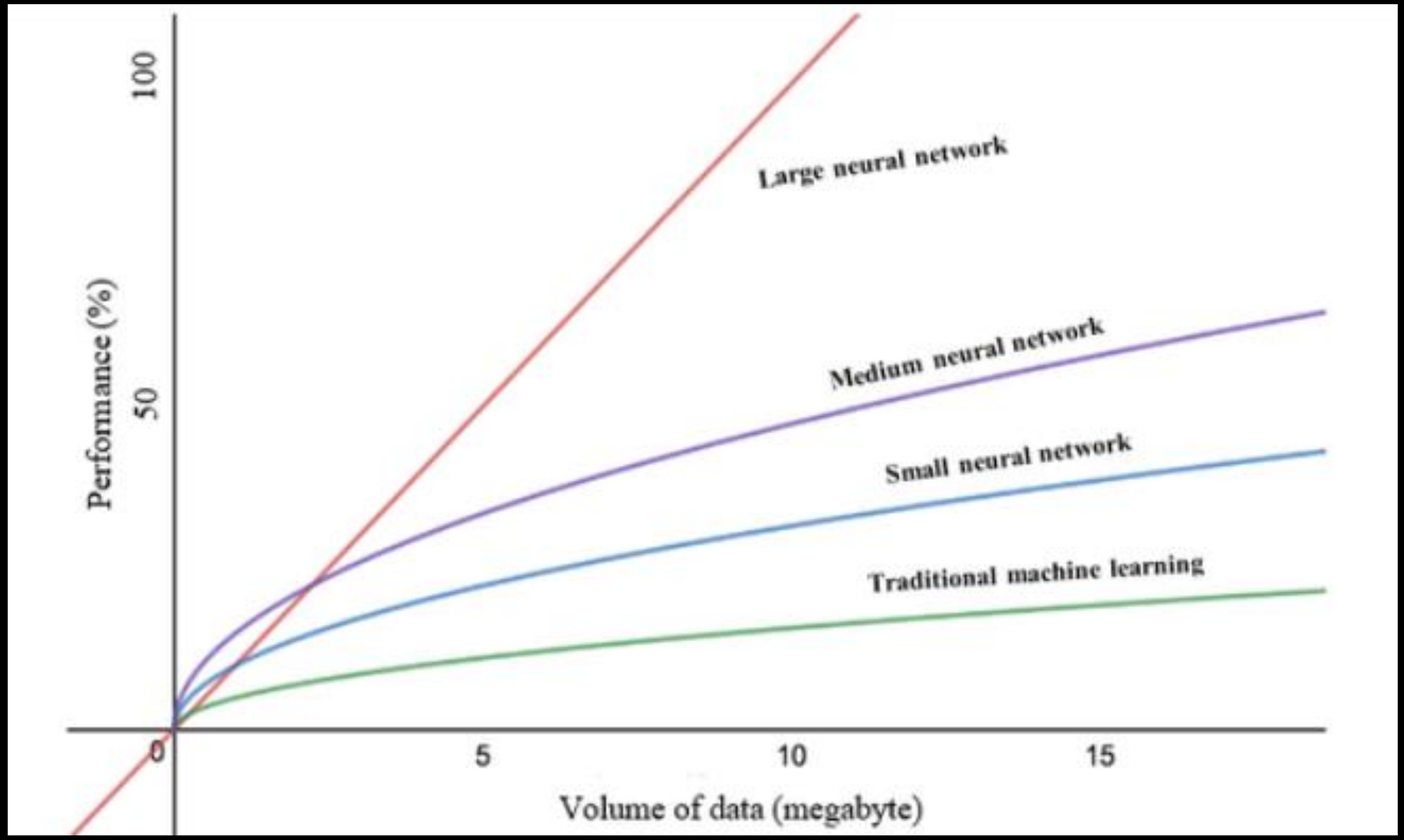

x2
1
0
0
1
x1
a) x1 AND x2
x2
1
0
0
1
x1
b) x1 OR x2
x2
1
?
0
0
1
x1
c) x1 XOR x2

Multi Layer Neural Network




Input layer
Hidden layer 1
Hidden layer 2
Output layer
x_0
x_1
x_2
...
x_n
Output

Backpropagation


Performance (%)
100
50
0
5
10
15
Volume of data (megabyte)
Large neural network
Medium neural network
Small neural network
Traditional machine learning

Q.43	Consider a Multi-Layer Perceptron (MLP) model with one hidden layer and one output layer. The hidden layer has 10 neurons, and the output layer has 3 neurons. The input to the MLP is a 5-dimensional vector. Each neuron is connected to every neuron in the previous layer, and a bias term is included for each neuron. The activation function used is the sigmoid function. The total number of trainable parameters in this MLP model is _____.	

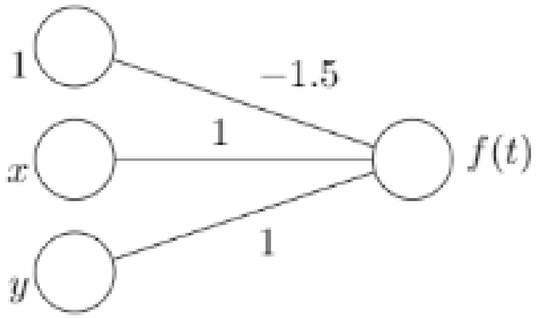
Consider the two class classification task that consists of the following points:

Class C_1 : [1 1.5]	[1 -1.5]
Class C_2 : [-2 2.5]	[-2 -2.5]

The decision boundary between the two classes using single perceptron is given by:

A. $x_1 + x_2 + 1.5 = 0$
B. $x_1 + x_2 - 1.5 = 0$
C. $x_1 + 1.5 = 0$
D. $x_1 - 1.5 = 0$

Consider a single perception with weights as given in the following figure:



and $f(t)$ is defined as

$$f(t)\begin{cases}1, t > 0\\ 0, t \leq 0\end{cases}$$

The above perception can solve

A. OR problem
B. AND problem
C. XOR problem
D. All of the above

Answer: (B)

Consider the two neural networks (NNs) shown in Figures 1 and 2, with $ReLU$ activation $(\mathrm{ReLU}(z) = \max\{0, z\}, \forall z \in \mathrm{R})$. The connections and their corresponding weights are shown in the Figures. The biases at every neuron are set to 0.

For what values of p, q, r in Figure 2 are the two NNs equivalent, when x_1, x_2, x_3 are positive?

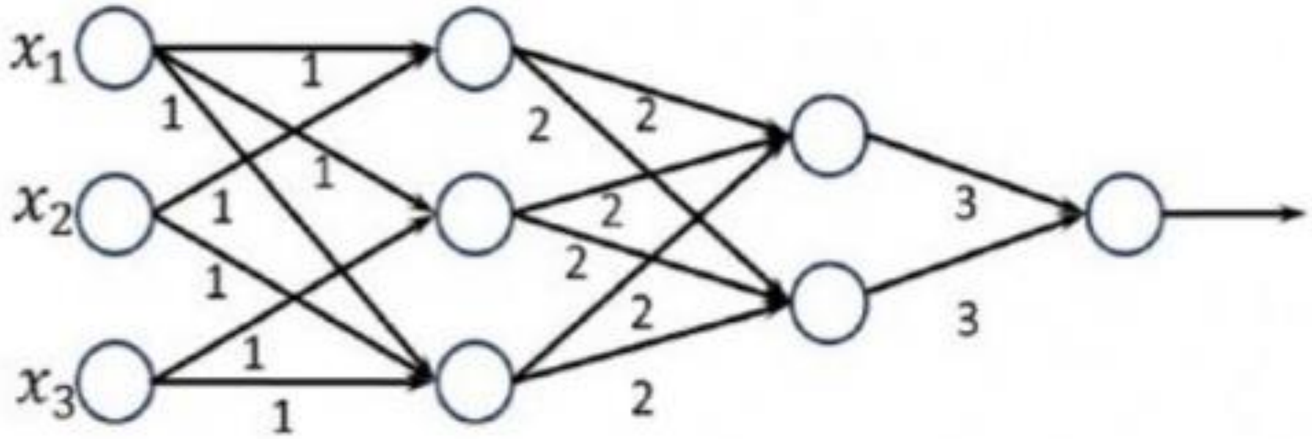


Figure 1

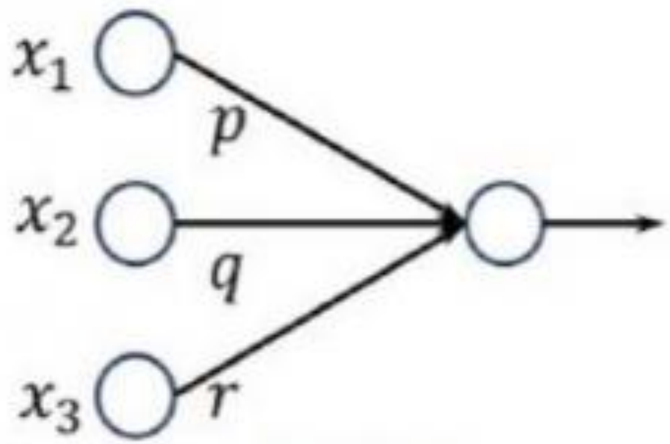


Figure 2

Note: $\mathbb{R}$ denotes the set of real numbers.

A. $p = 36, q = 24, r = 24$
B. $p = 24, q = 24, r = 36$
C. $p = 18, q = 36, r = 24$
D. $p = 36, q = 36, r = 36$

Given a feedforward neural network with a large number of hidden layers, which of the following issues is it most likely to face?

A. Overfitting and vanishing gradient

B. Underfitting and exploding gradient

C. Bias-variance trade-off and linear separability

D. None of the above

Answer: A

Thank you