



# Machine Learning

30

Marks	Placed
24	Placed not placed

Classification  
Regression

$x$

Height

168 cm

166 5

⋮

$f(x)$

Weight

82 kg

81 kg

Size  
of house

1200 sq ft

⋮

⋮

⋮

⋮

$f(x)$

↑

Price

50 L

price

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

→

$x_1 \quad x_2 \quad \dots \quad x_n \quad \underbrace{f(x)}$

↓  
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

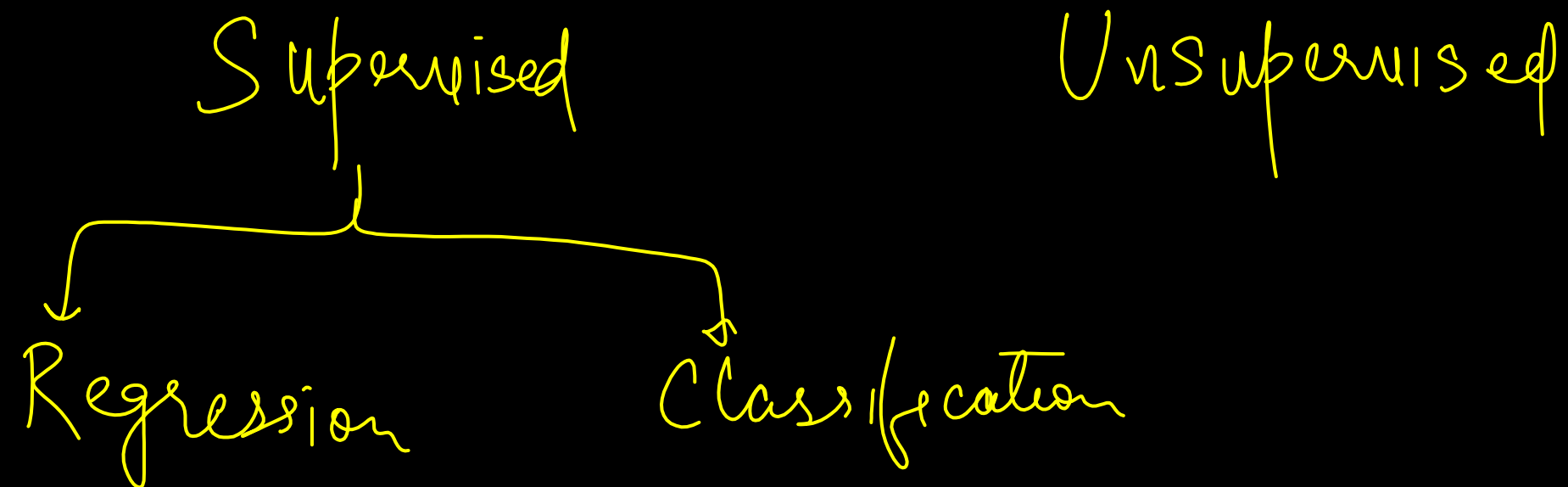
Which of the following are supervised learning problems (Multiple Correct)?

- 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

↓  
Reinforcement

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

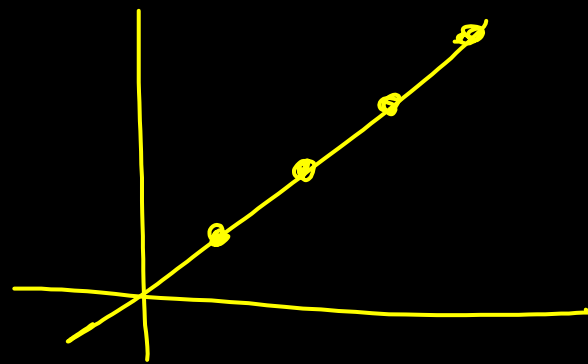


$$x \text{ and } y \left\{ \begin{array}{l} y = ax + b \end{array} \right.$$

# Linear Regression

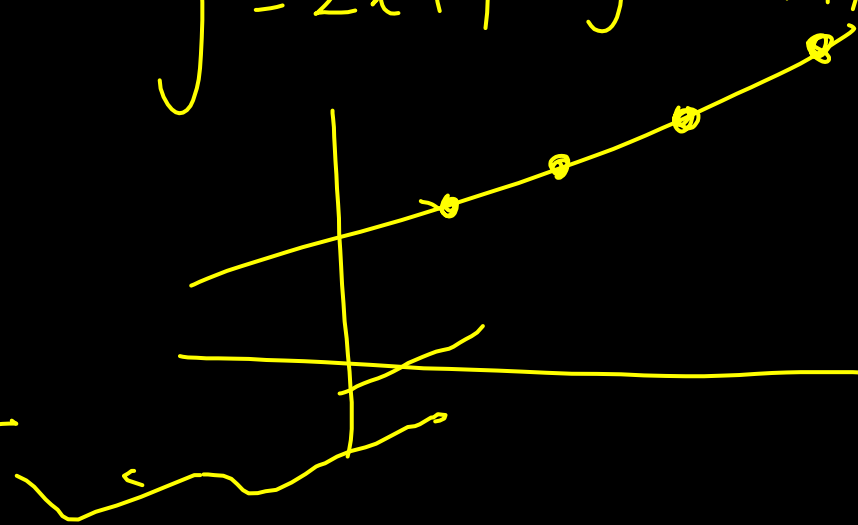
$x$	$y$
2	4
3	6
1	2
4	8

$$y = 2x$$



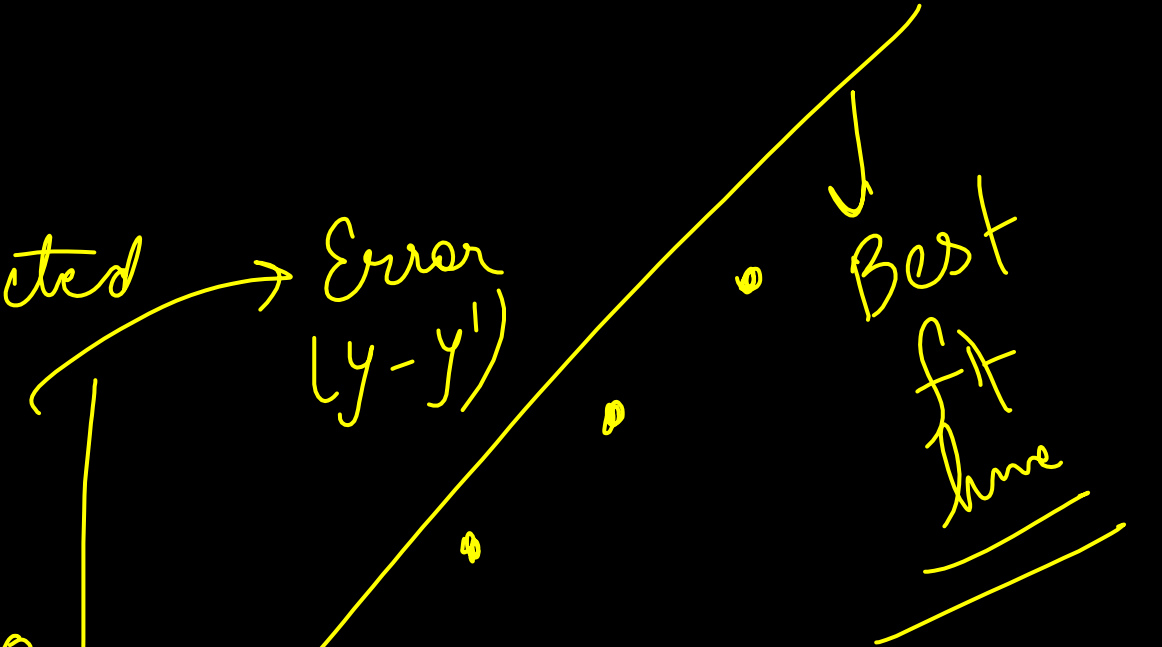
$x$	$y$	$y'$
2	5	5
3	7	7
1	3	3
4	9	9

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



	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
		<u>1</u>	

$$y' = 2x \rightarrow \text{predicted value}$$



$x$	$y$	$y'$	Error $(y - y')$	Error <sup>2</sup>	Error
1	3	2	1	0	0
2	4	4	0	0	0
3	6	6	0	0	1
4	8	8	-1	1	
5	9	10			
			<u>0</u>	<u>2</u>	<u>2</u>

$$y = 2x$$

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

$$= \frac{1}{2n} \left[ (y_1 - y'_1)^2 + (y_2 - y'_2)^2 + \dots + (y_n - y'_n)^2 \right]$$

$$\begin{aligned} \text{Mean Absolute Error} \\ &= \sum_{i=1}^n \frac{1}{2n} |y_i - y'_i| \end{aligned}$$

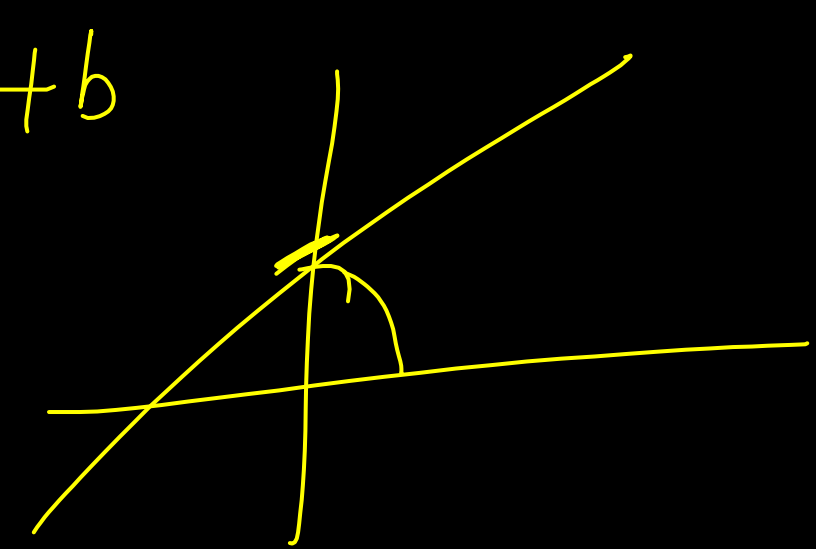




What is Best Fit Line?

$$\boxed{\begin{aligned} y_i' &= wx_i + b \\ y_i' &= wx_i + b \end{aligned}}$$

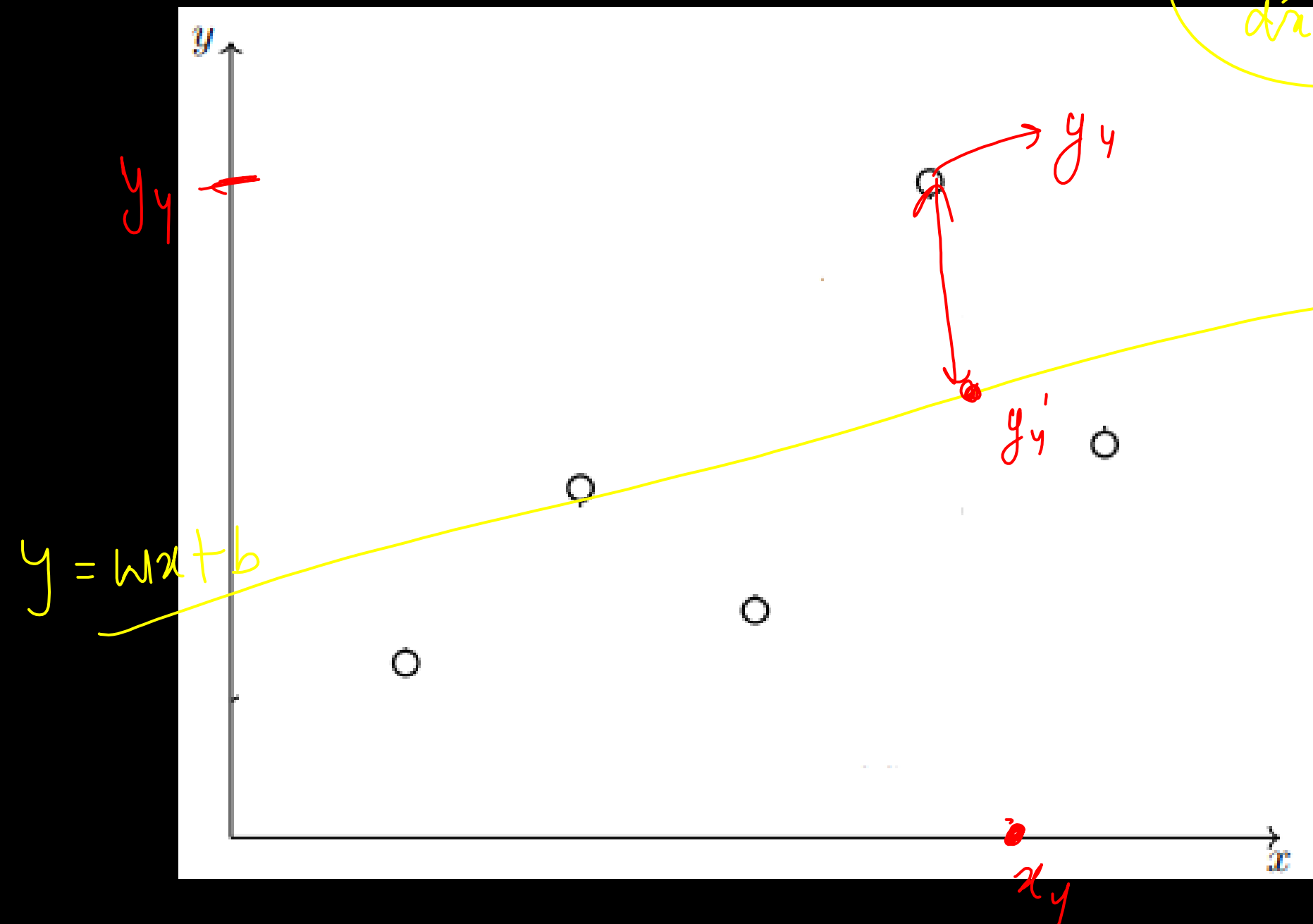
$$y = ax + b$$



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

$$\begin{aligned} y &= e^x \\ \frac{dy}{dx} &= e^x = 0 \end{aligned}$$

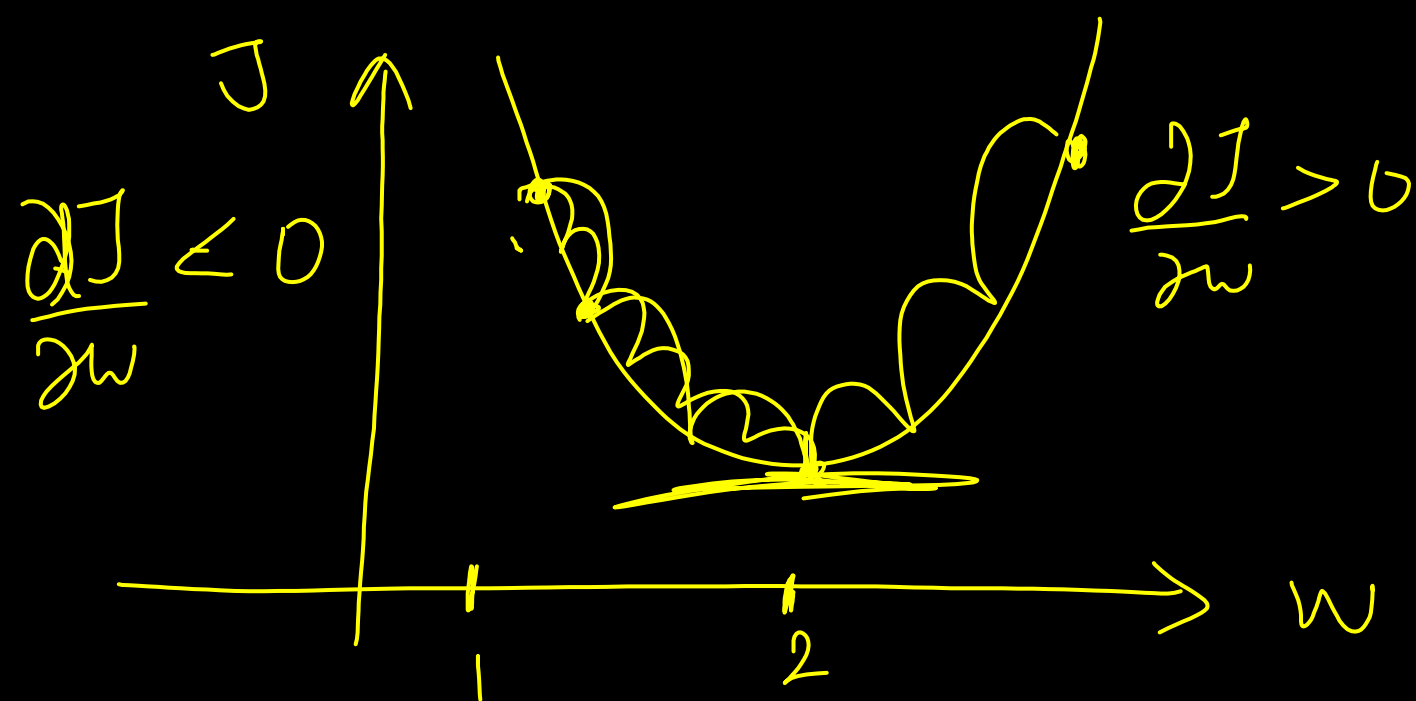
$$y = wx + b$$



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

$$\begin{aligned} J(w, b) \\ \text{Cost function} &= \frac{1}{2n} \sum (y_i - (wx_i + b))^2 \\ \text{Goal} &\rightarrow w, b \end{aligned}$$

# Gradient Descent



$$W = W - \alpha \left( \frac{\partial J}{\partial w} \right)$$

$$W = W - \alpha \left( \frac{\partial J}{\partial w} \right)$$

$$= 1 - \alpha (-2)$$

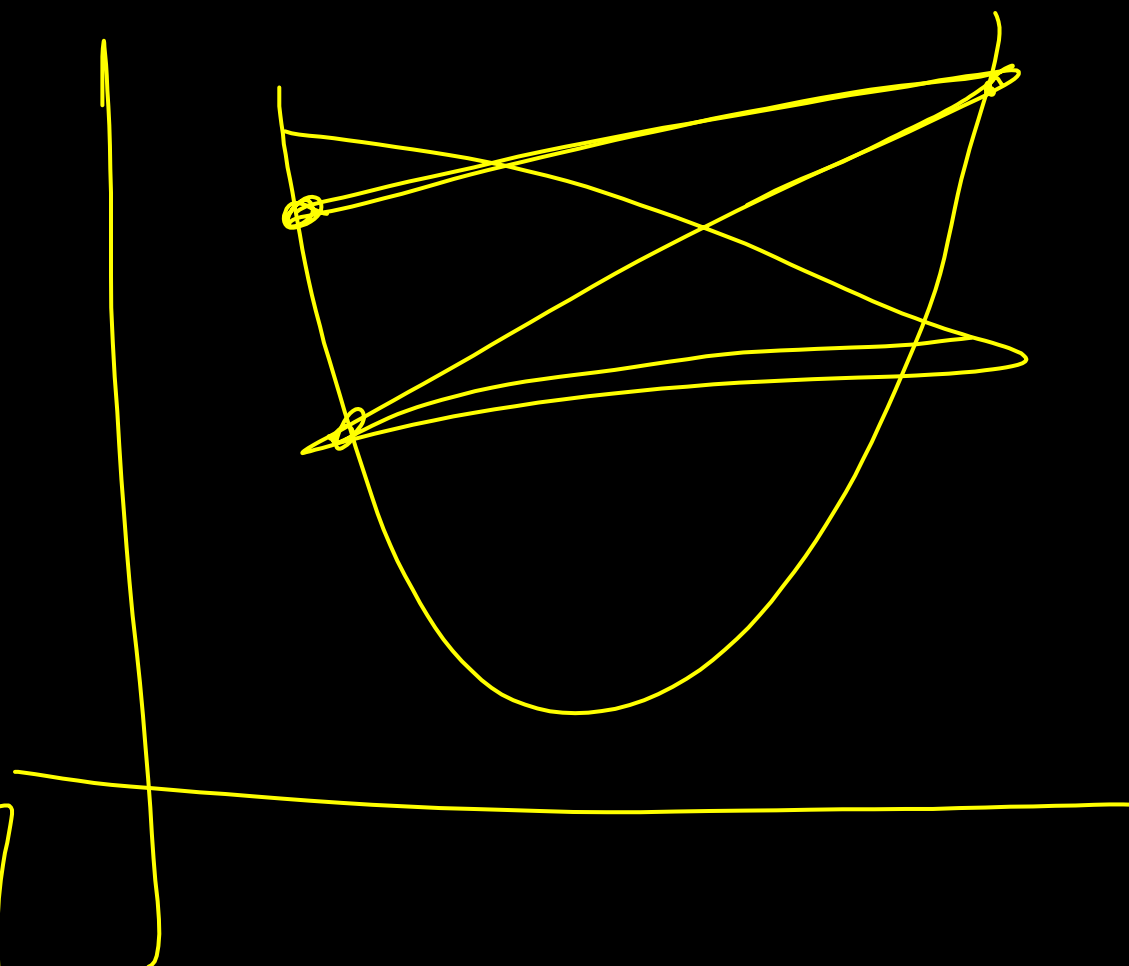
$$= 1 + 0.1 \times 2$$

$$= 1.2$$

$$W = 1.2 - 0.1 f(1.2)$$

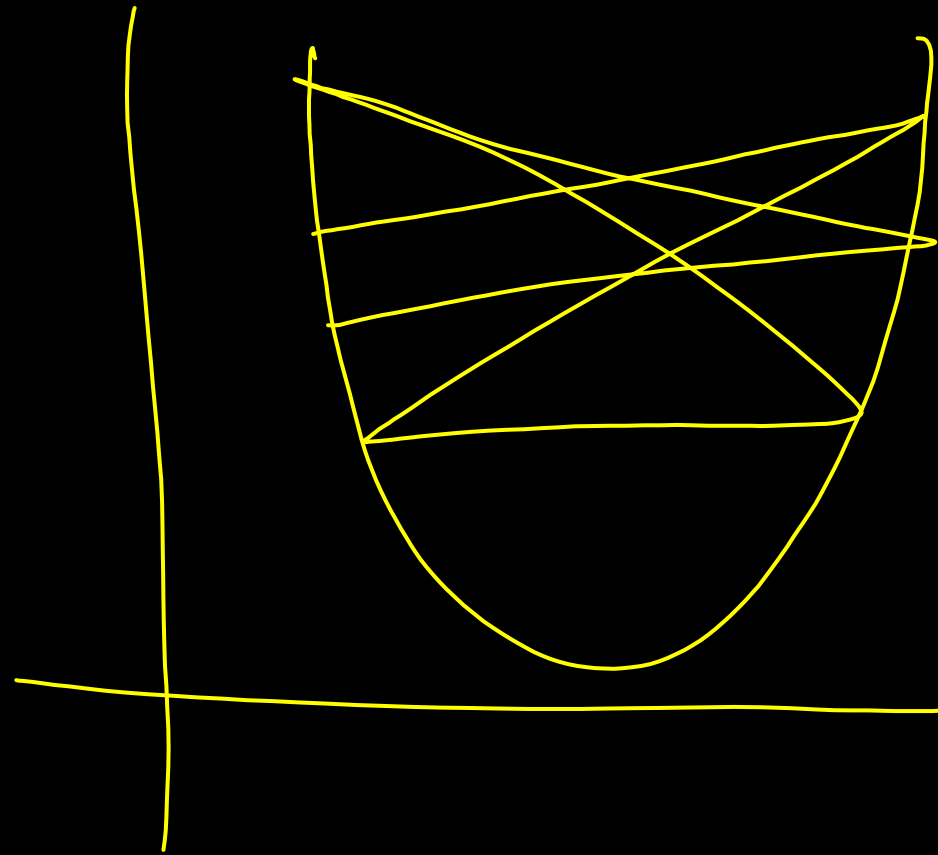
$$= 1.2$$

$$b = b - \alpha \left( \frac{\partial J}{\partial b} \right)$$

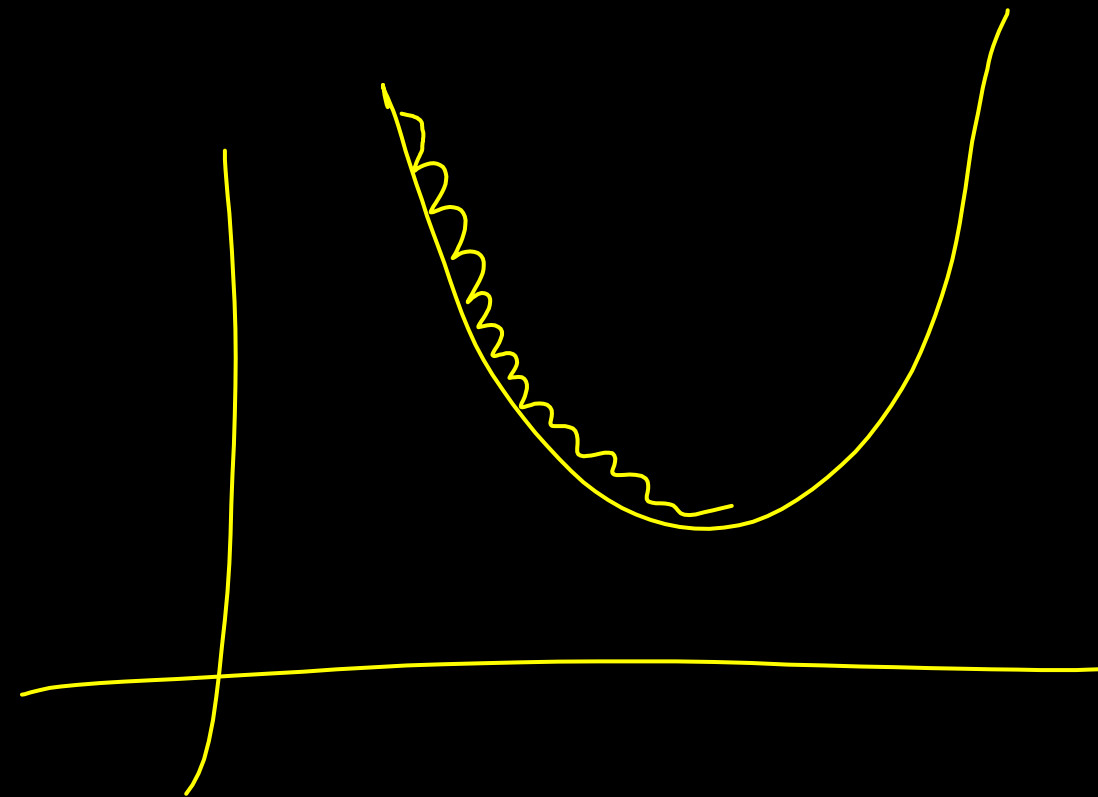


# Learning Rate

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



$\alpha \rightarrow$  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

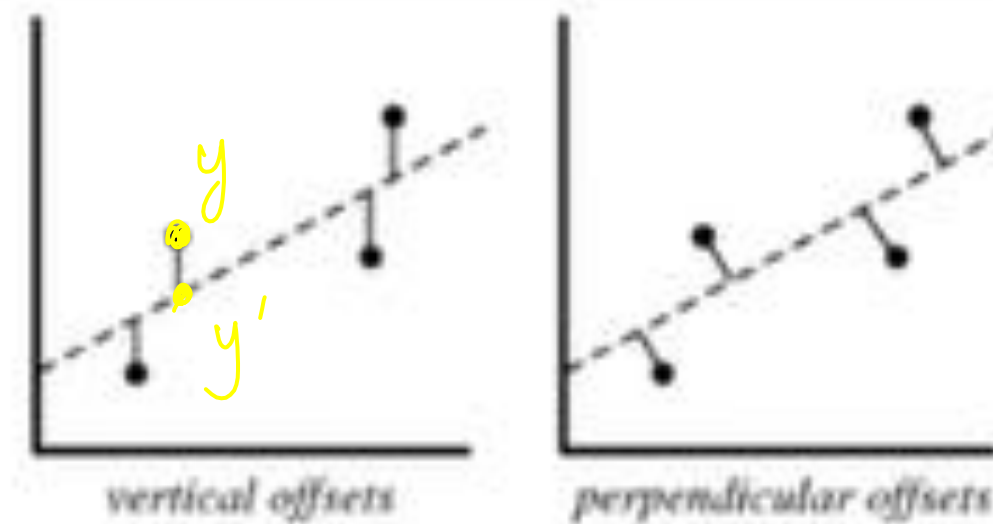
$$y = \frac{1}{2}x + 0$$

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.



→

PCA

- ✓ 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?

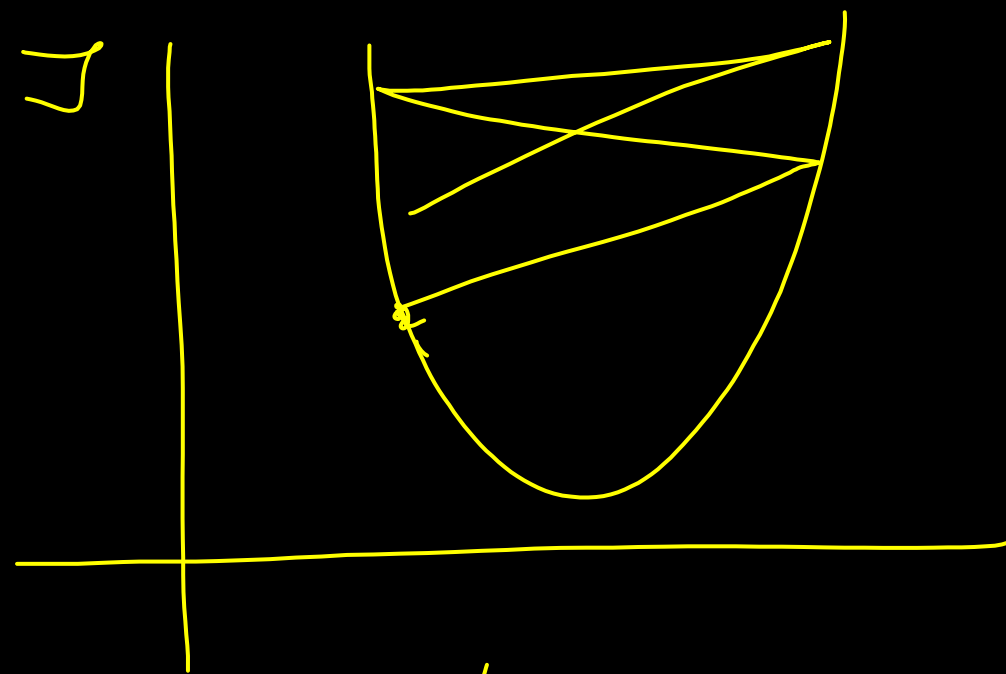
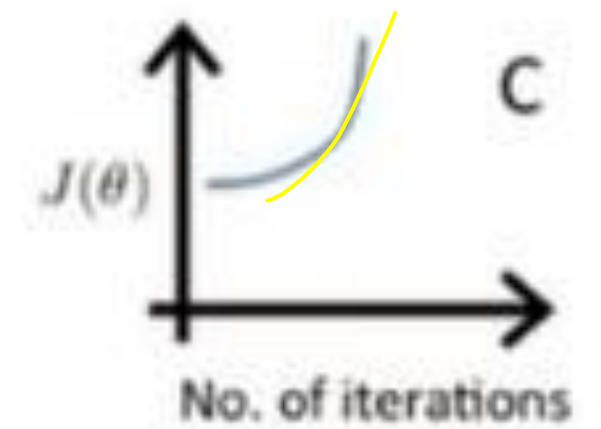
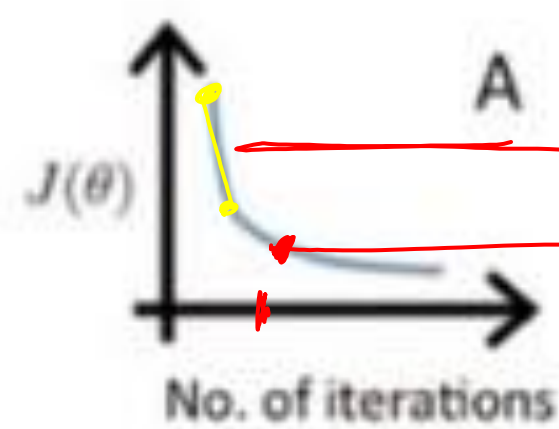
- ☒ A) Linear regression is sensitive to outliers
- ☐ B) Linear regression is not sensitive to outliers
- ☐ C) Can't say
- ☐ D) None of these



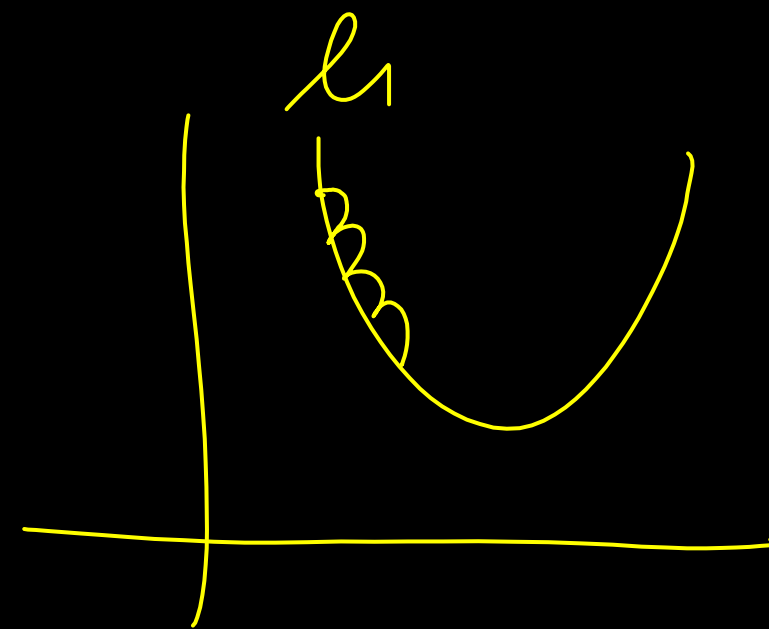
1, 2, 3, 4, 1000

Suppose  $l_1$ ,  $l_2$  and  $l_3$  are the three learning rates for A,B,C respectively. Which of the following is true about  $l_1$ ,  $l_2$  and  $l_3$ ?

- A)  $l_2 < l_1 < l_3$
- B)  $l_1 > l_2 > l_3$
- C)  $l_1 = l_2 = l_3$
- D) None of these



$\alpha \rightarrow$  too large



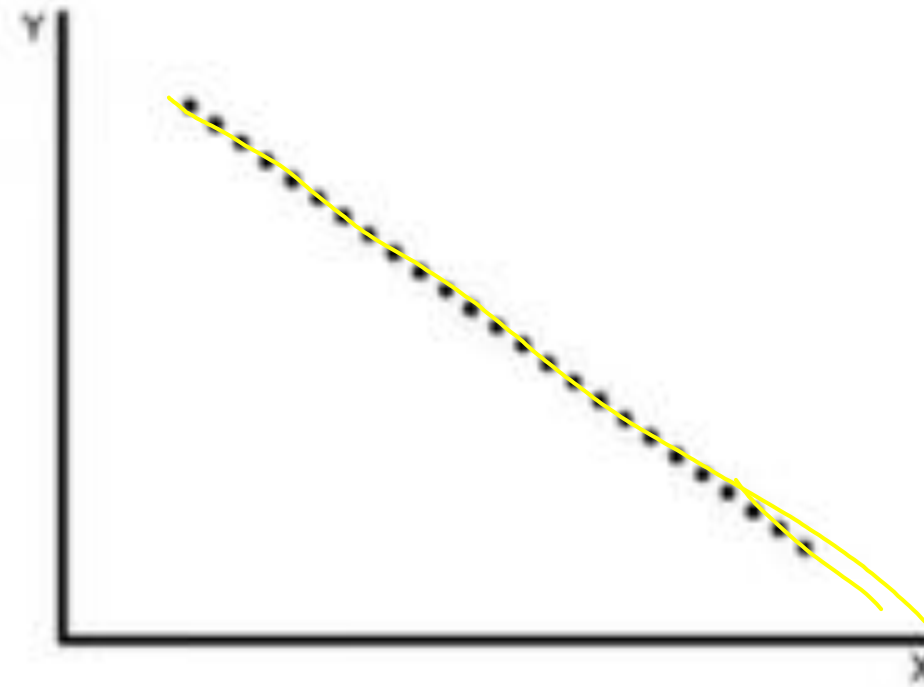
$\alpha \rightarrow$  too small

$l_2$   
Small  
 $l_2 < l_1 < l_3$

$l_3$   
 $\downarrow$   
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



$$x_1 \ x_2 \ \dots \ x_k \ y$$

$$y = w_1 x_1 + w_2 x_2 + \dots + w_k x_k + b$$

## Multiple regression

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

$$\begin{array}{cccc} x_1 & x_2 & x_3 & y \\ \text{no of room} & \text{location} & \text{size} & \text{price} \end{array}$$

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

$$\longrightarrow \text{Goal} \rightarrow w_1, w_2, w_3, b$$

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

$$y' = w^T x$$

$$y' = w_0 x + w_1 x_1 + w_2 x_2 + w_3 x_3$$

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

$$= w^T x$$

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

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



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.

$$(w_1, w_2, \dots, w_k) \rightarrow w_1=0, w_2=0, w_3 \neq 0, w_4 \neq 0, \dots$$

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

$$w_1 \neq 0, w_2 \neq 0, \dots$$



$$w_1 \approx 0$$

$$0.006/$$

## L1 Regularization (Lasso)

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

*hyperparameter*

$$y' = w_1 x_1 + w_2 x_2 + \dots + w_n x_n + b$$

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

↓  
Min

⇓  
 $w_i, b$

↓  
 $\lambda = 1000$

$1000 \times$

$= 1000$

$$w_i' = 0$$

↓  
Sparse

## L2 Regularization (Ridge):

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

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

$\downarrow$   
 $w_i \neq 0$

$$w_1 = 0 \ 0000$$

$$y = \underbrace{100}_{w_1} x_1 + \underbrace{200}_{w_2} x_2 + \underbrace{300}_{w_3} x_3 + b$$

## 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,  $\lambda_1$  and  $\lambda_2$  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)

$\lambda = 0$

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

→ 0

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)

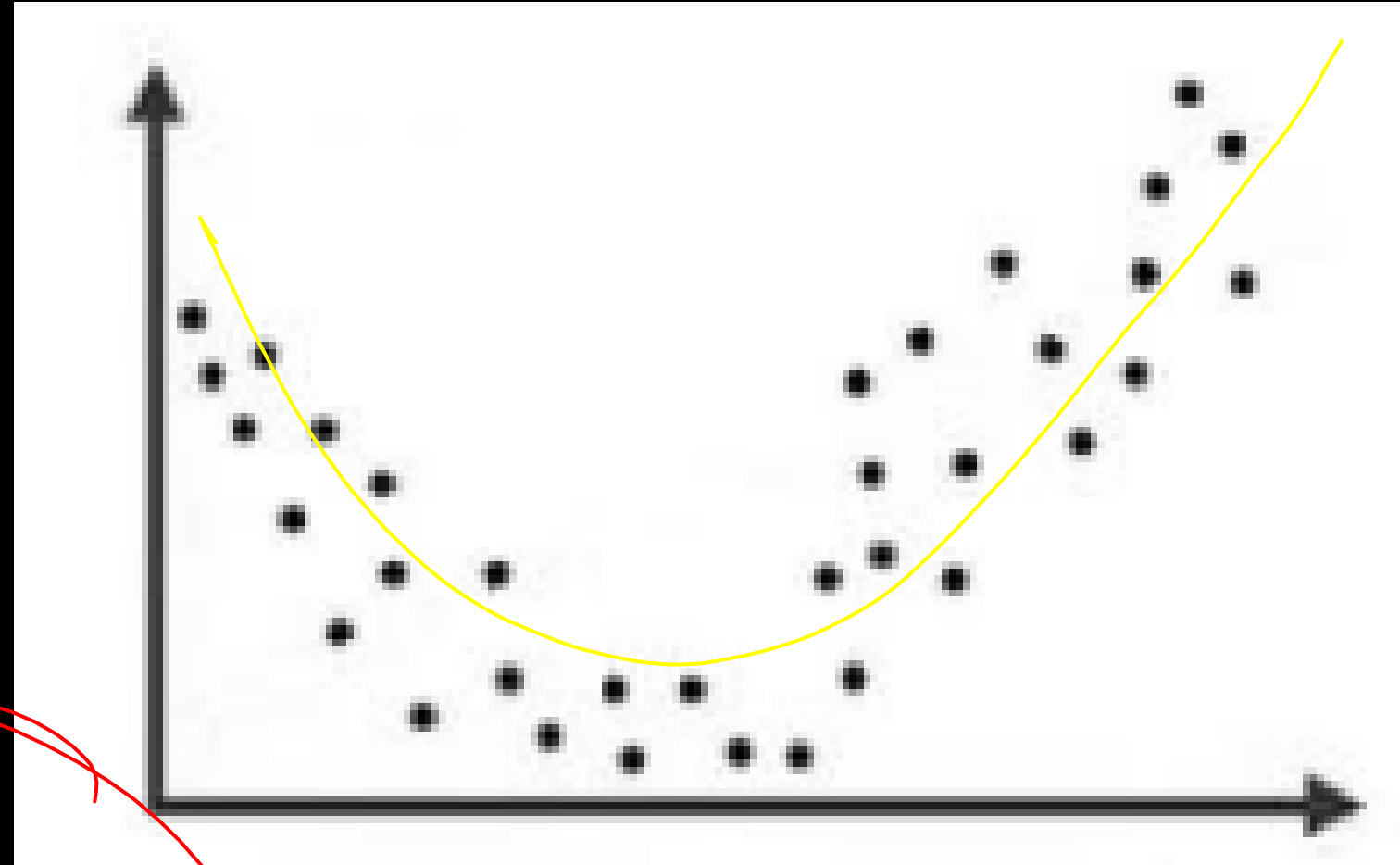
$$\lambda \rightarrow 1000 \quad \left. \begin{array}{l} w_1 = 0.001 \end{array} \right\} \rightarrow 1$$

$$\lambda = 10,000 \quad \left. \begin{array}{l} w \downarrow \end{array} \right\} \rightarrow 0$$

$$\lambda \uparrow \quad \left. \begin{array}{l} w \downarrow \end{array} \right\} \rightarrow 0$$

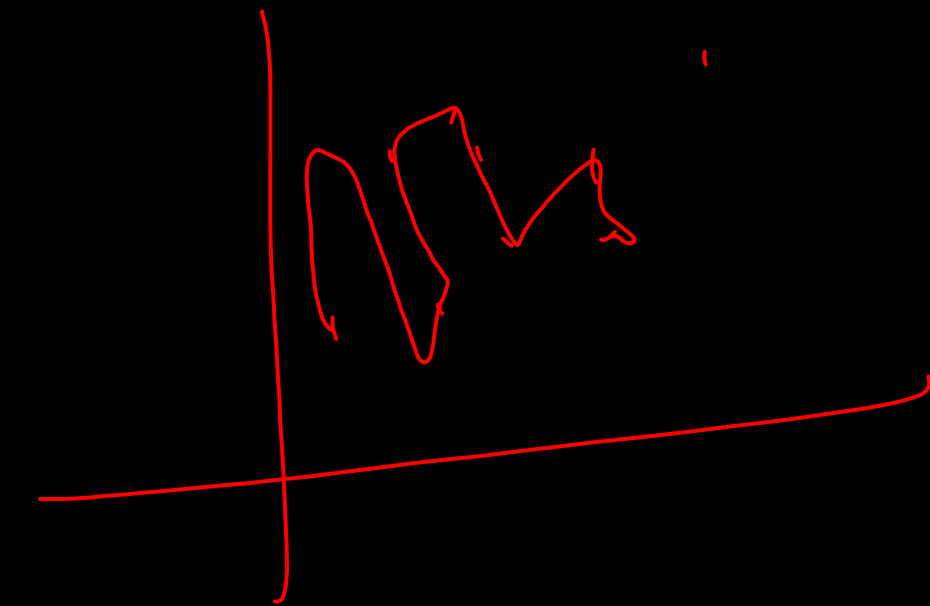
✓

# Polynomial regression



$$y = wx + b$$

$x$   $y$



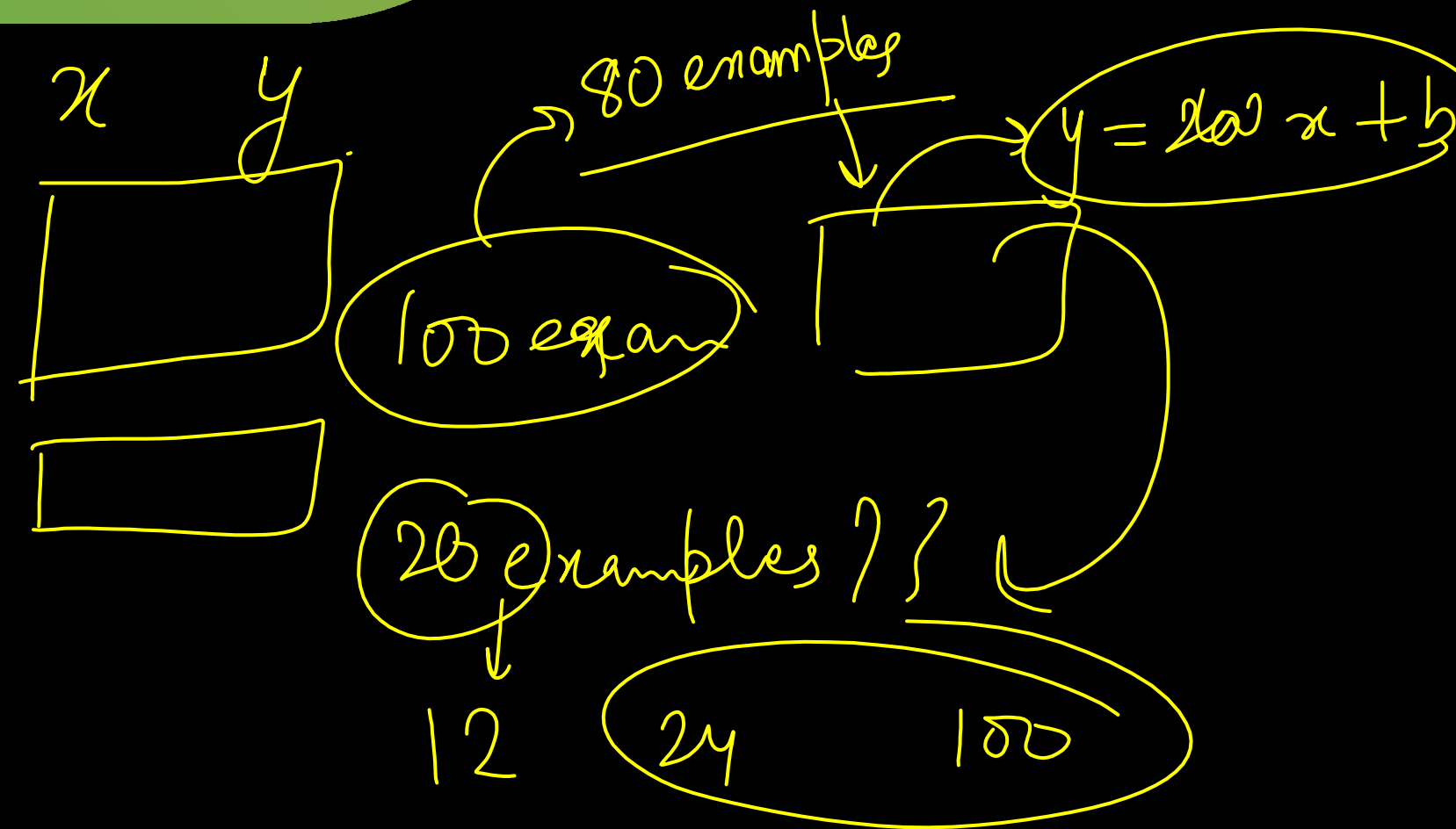
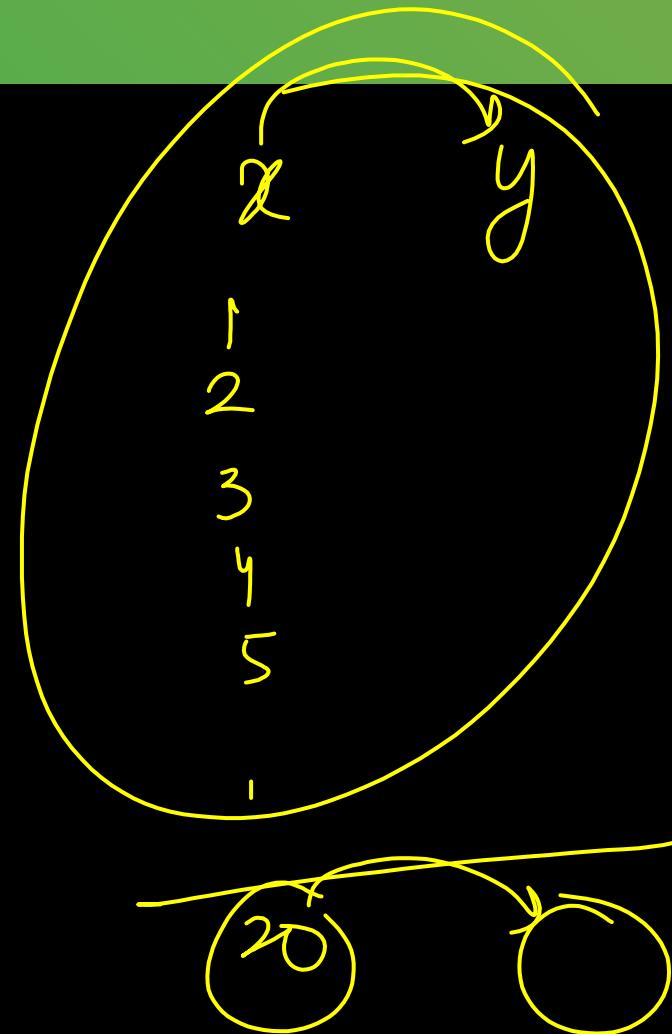
Wavy line

$$y = w_1x + w_2x^2 + w_3x^3 + w_4x^4 + w_5x^5 + w_6x^6 + w_7x^7 + b$$

$$y = w_1x + w_2x^2 + b$$

# Bias Variance Tradeoff

B<sub>1</sub>





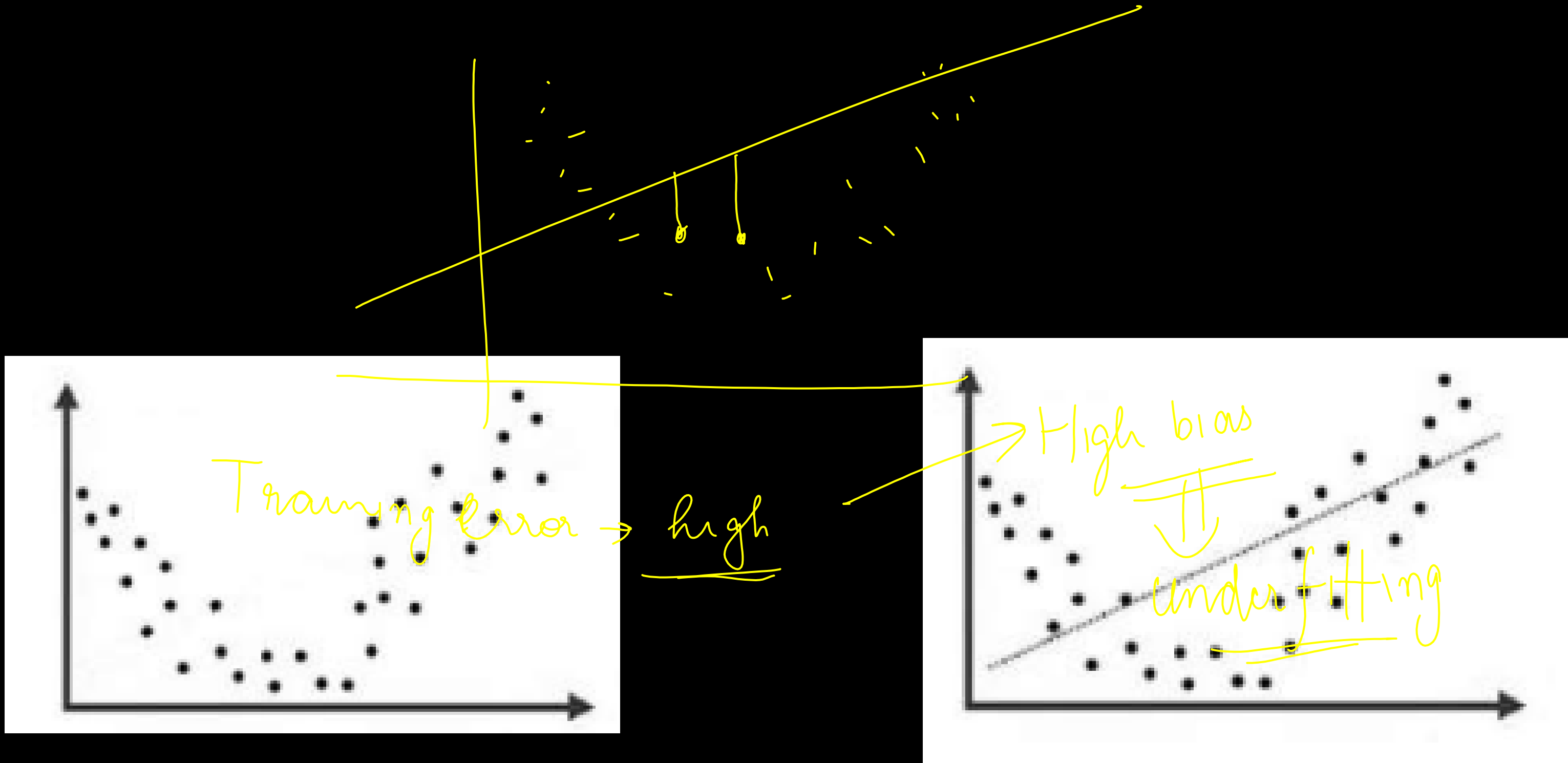
## Training & Test Data

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

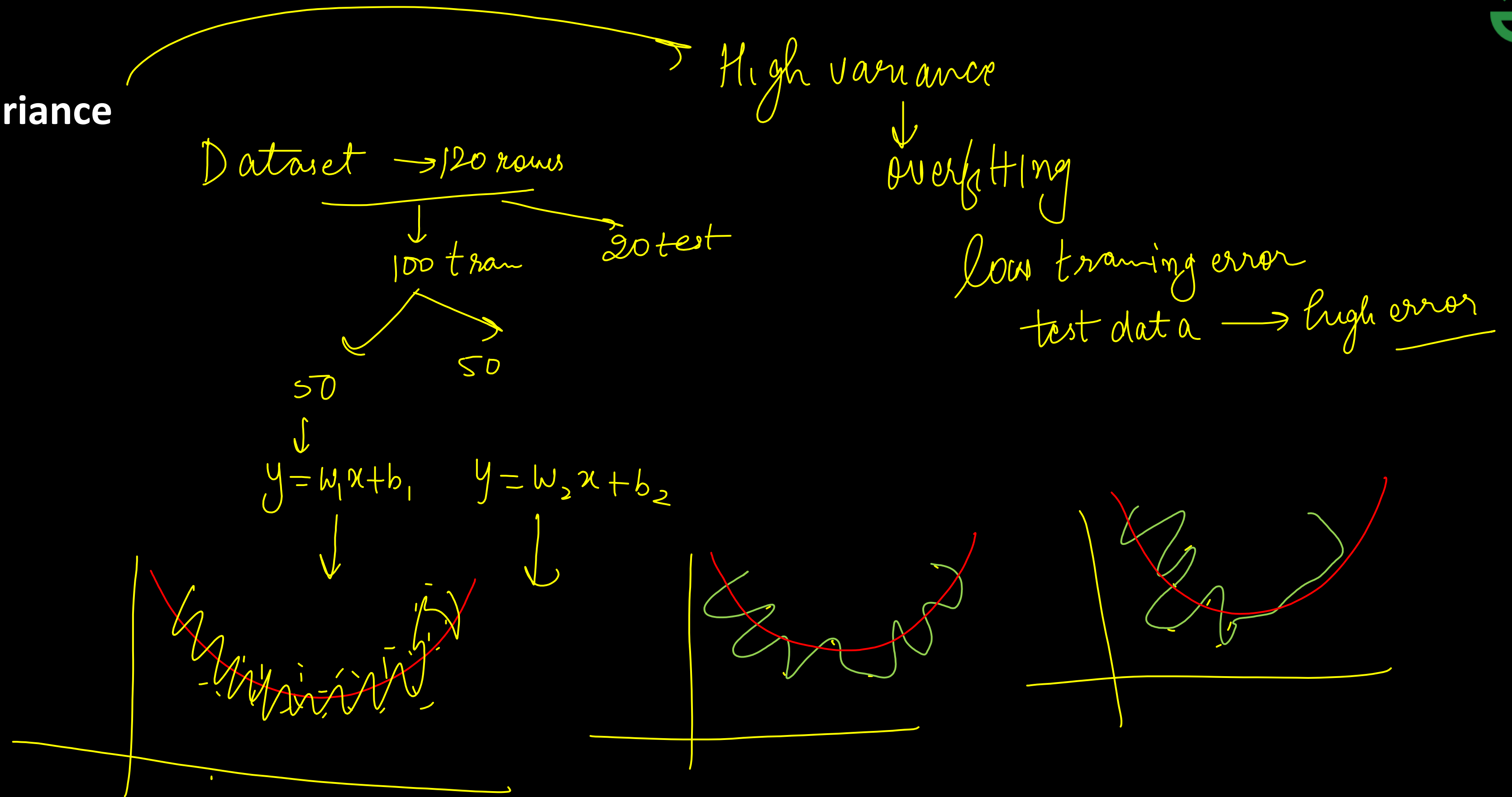
The first split of data, i.e. the initial reserve of data you use to develop your model, provides the training data.

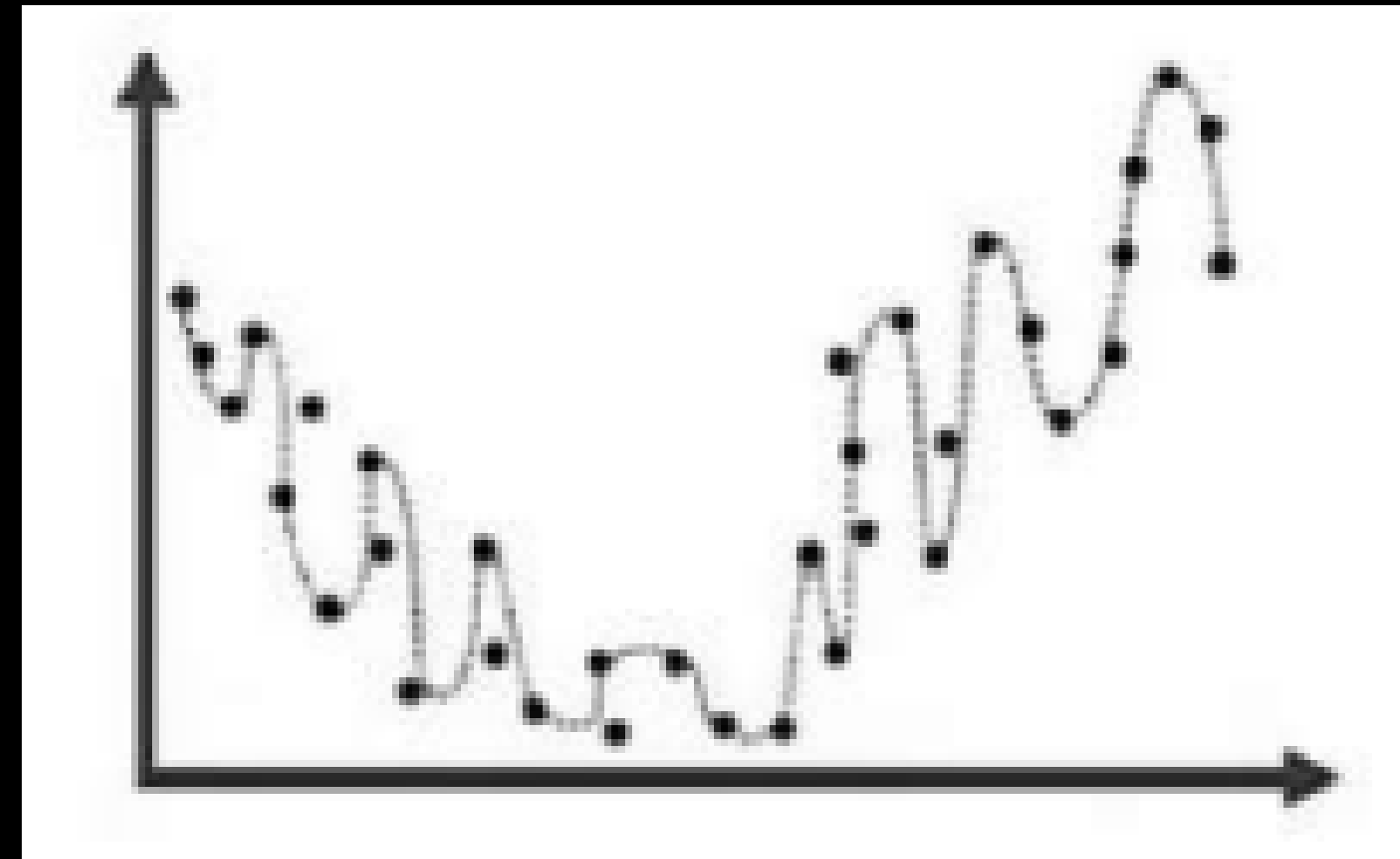
After you have successfully developed a model based on the training data and are satisfied with its accuracy, you can then test the model on the remaining data, known as the test data.

# Bias



# Variance

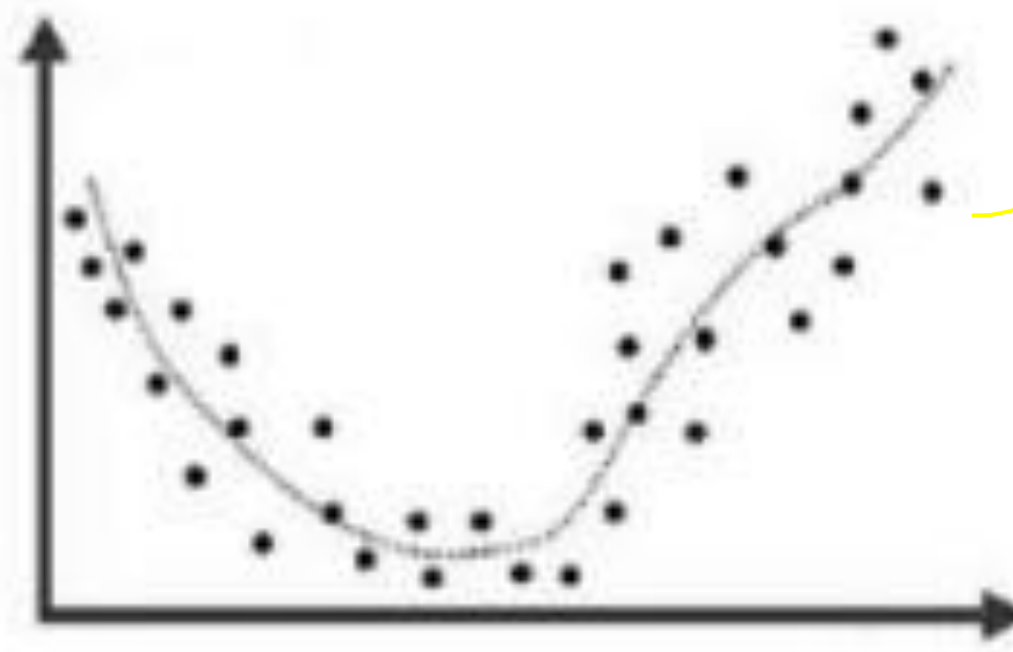




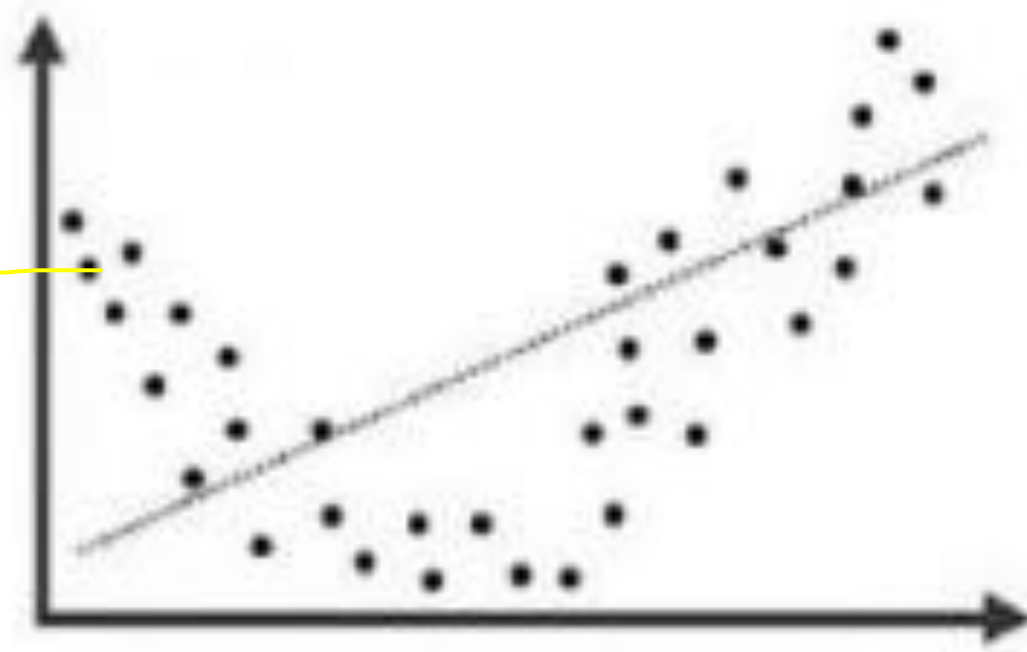




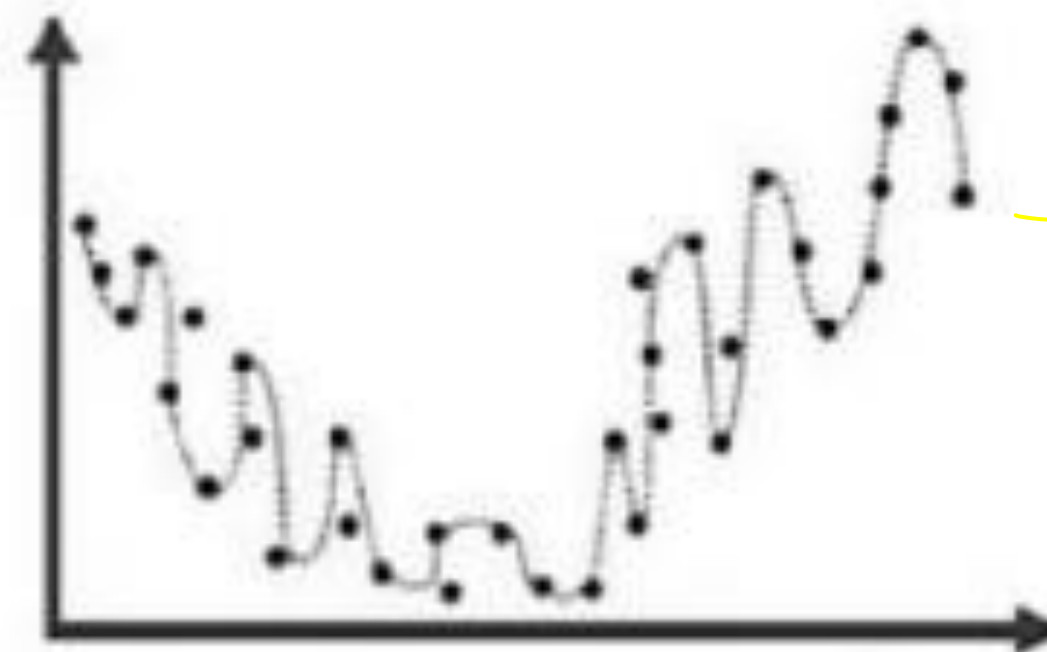
(a) Given dataset



(b) "Just right" model



(c) Underfitting model

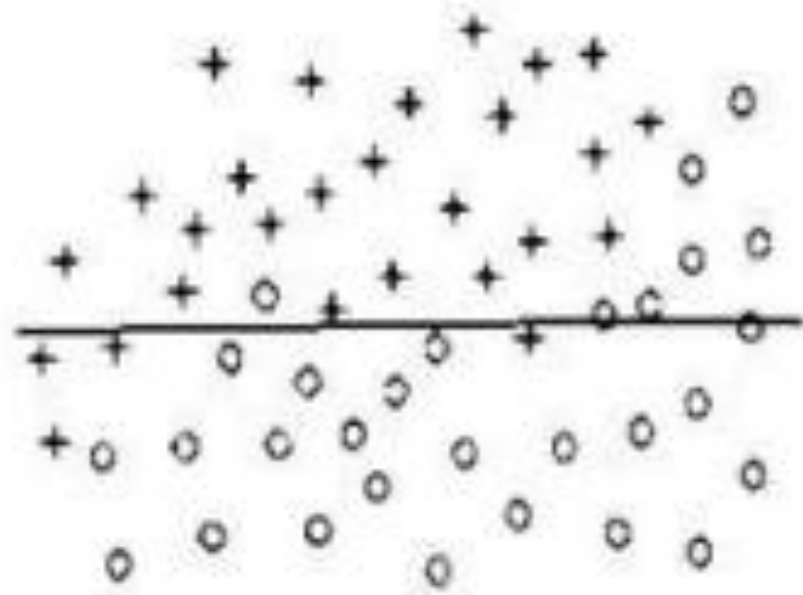


(d) Overfitting model

High bias

High variance  
low bias

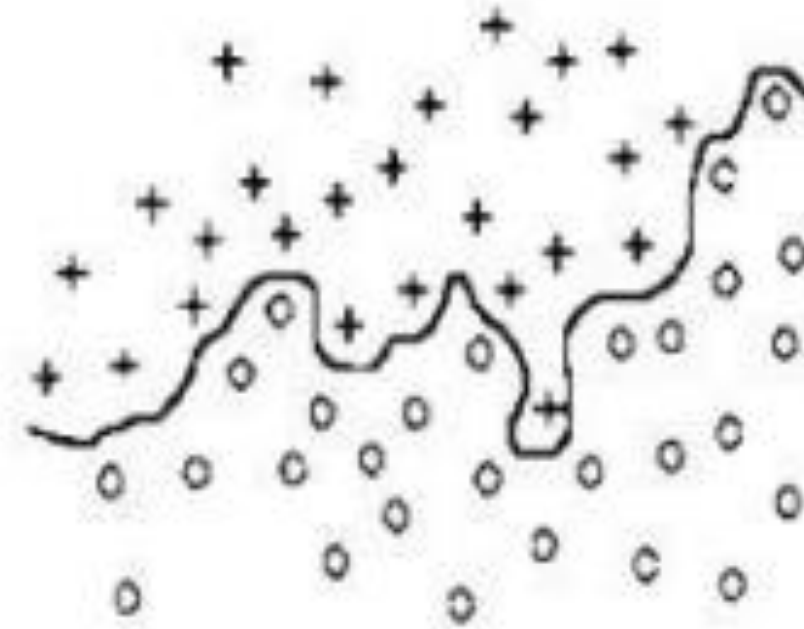




(a) Underfitting



(b) Right fitting



(c) Overfitting

## Tradeoff:

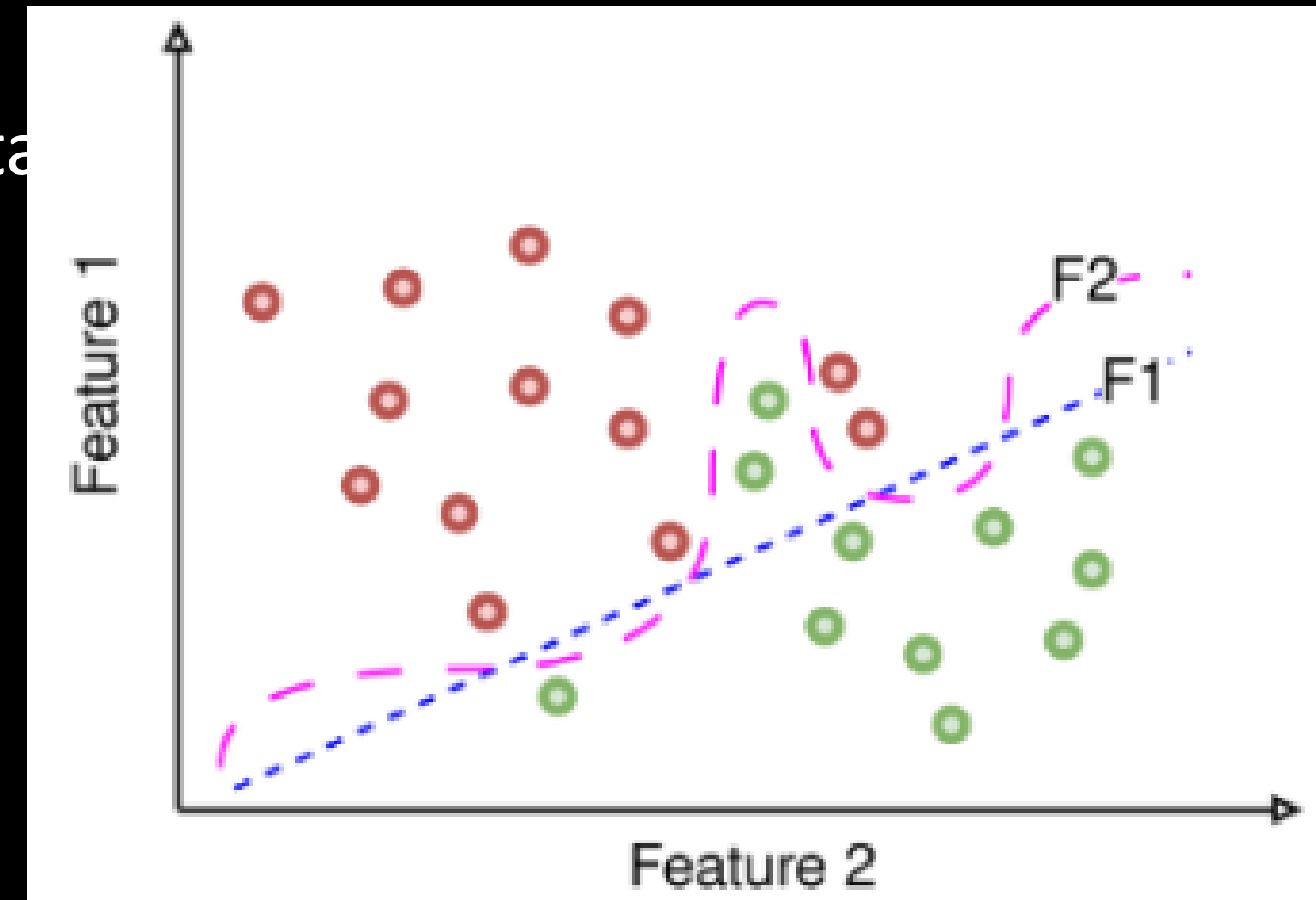
- The bias-variance tradeoff suggests that there is a balance to be struck between bias and variance. As you decrease bias (by increasing model complexity), you typically increase variance, and vice versa.
- The goal is to find the right level of model complexity that minimizes the combined error due 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

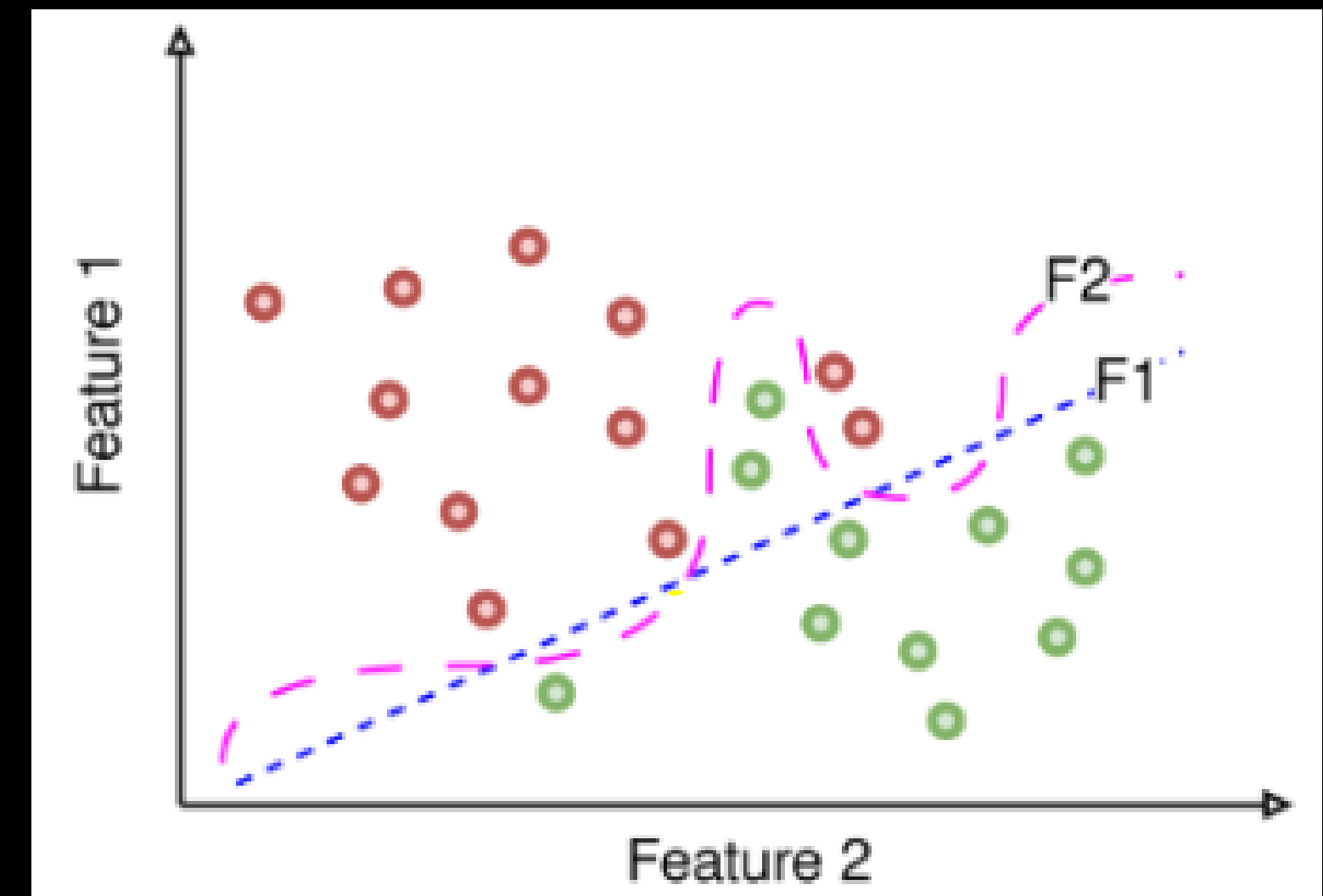
Answer: (C)



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)

$$w_0x^3 + w_1x^2 + w_2x + b + w_3x^2 + w_4x$$

Handwritten equation with annotations: The term  $w_0x^3$  is circled with a yellow circle and has a yellow arrow pointing down from it. The entire expression  $w_0x^3 + w_1x^2 + w_2x + b + w_3x^2 + w_4x$  is circled with a yellow oval. There are also yellow checkmarks above the first three terms and a yellow checkmark above the entire expression.

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)

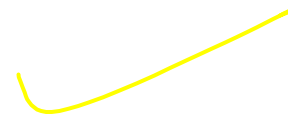
Suppose, you got a situation where you find that your linear regression model is under fitting the data. In such situation which of the following options would you consider?

- 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

9:50 pm  
↓

## 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

Classification

\_\_\_\_\_ PCa  
\_\_\_\_\_  $x, p$   
\_\_\_\_\_ Place



KNN  $\rightarrow$   $K=3$

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	???

$(x_1, y_1)$  &  $(x_2, y_2)$

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

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

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	???	

2 Intel & 1 Leader

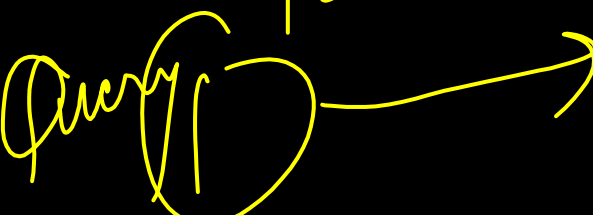
→ Intel

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

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
Query 

State whether the statement is True/False:

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

1. True
2. False

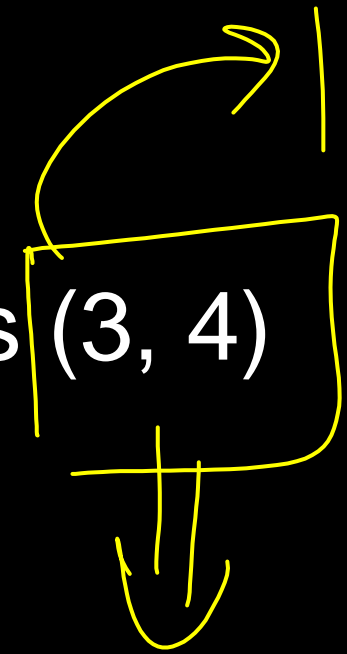


Consider a dataset with the following three data points in a two-dimensional space:

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

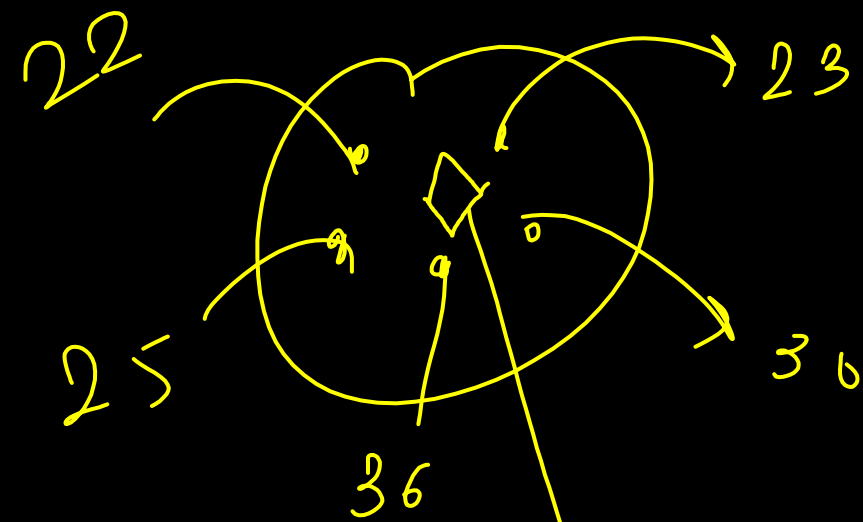
If  $K=2$ , what is the predicted class of a new data point at coordinates (3, 4) using the Euclidean distance metric?

$$\left. \begin{aligned} (3, 4) \text{ \& } A &= \sqrt{(3-2)^2 + (4-3)^2} = \sqrt{2} \\ (3, 4) \text{ \& } B &= \sqrt{(3-4)^2 + (4-6)^2} = \sqrt{5} \\ (3, 4) \text{ \& } C &= \sqrt{(3-5)^2 + (4-2)^2} = \sqrt{8} \end{aligned} \right\} \begin{array}{c} A \text{ \& } B \\ \text{---} \\ 1 \end{array}$$



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

Yes



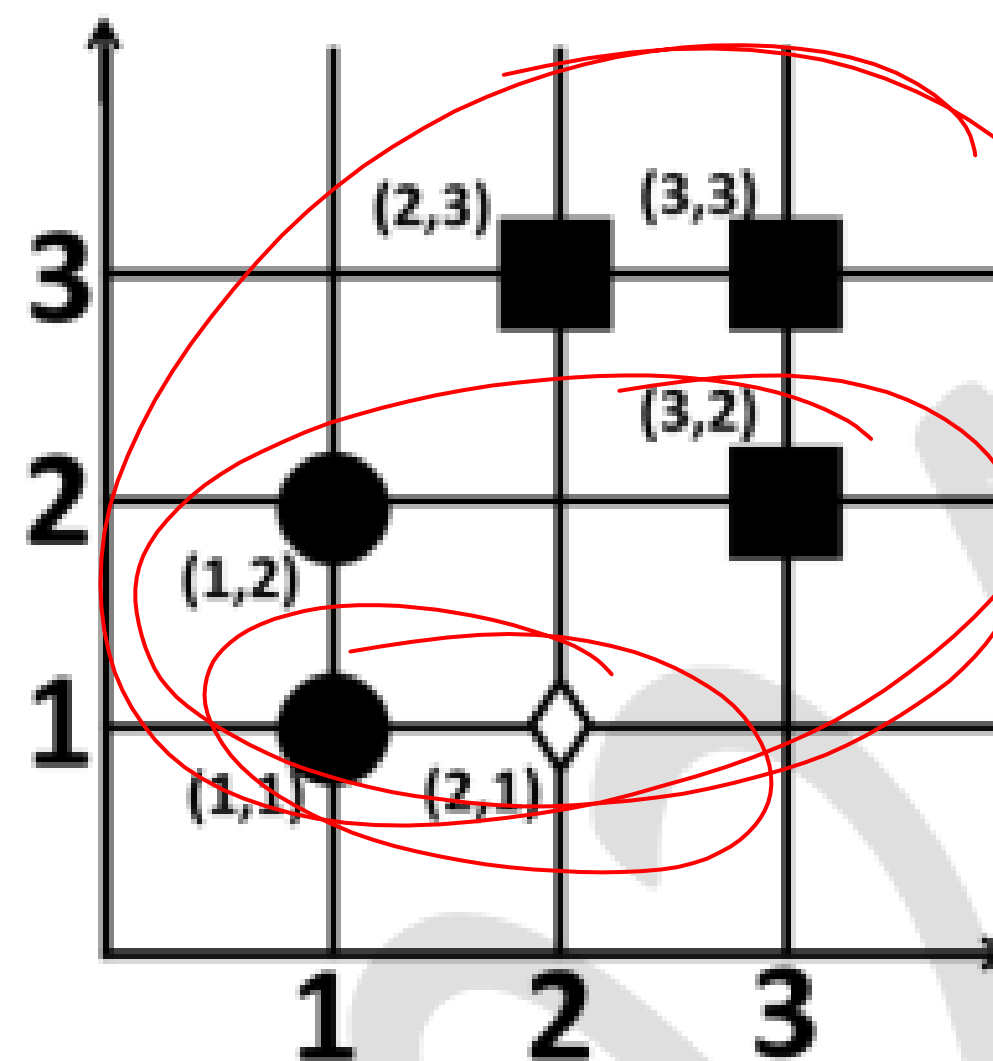
$$\frac{22 + 23 + 25 + 36 + 30}{5}$$

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

1.  $1\text{-NN} > 2\text{-NN} > 3\text{-NN}$
2.  $1\text{-NN} < 2\text{-NN} < 3\text{-NN}$
3.  $1\text{-NN} \sim 2\text{-NN} \sim 3\text{-NN}$  ✓
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 ( $\diamond$ ) shaped data point is assigned the label square is \_\_\_\_\_.

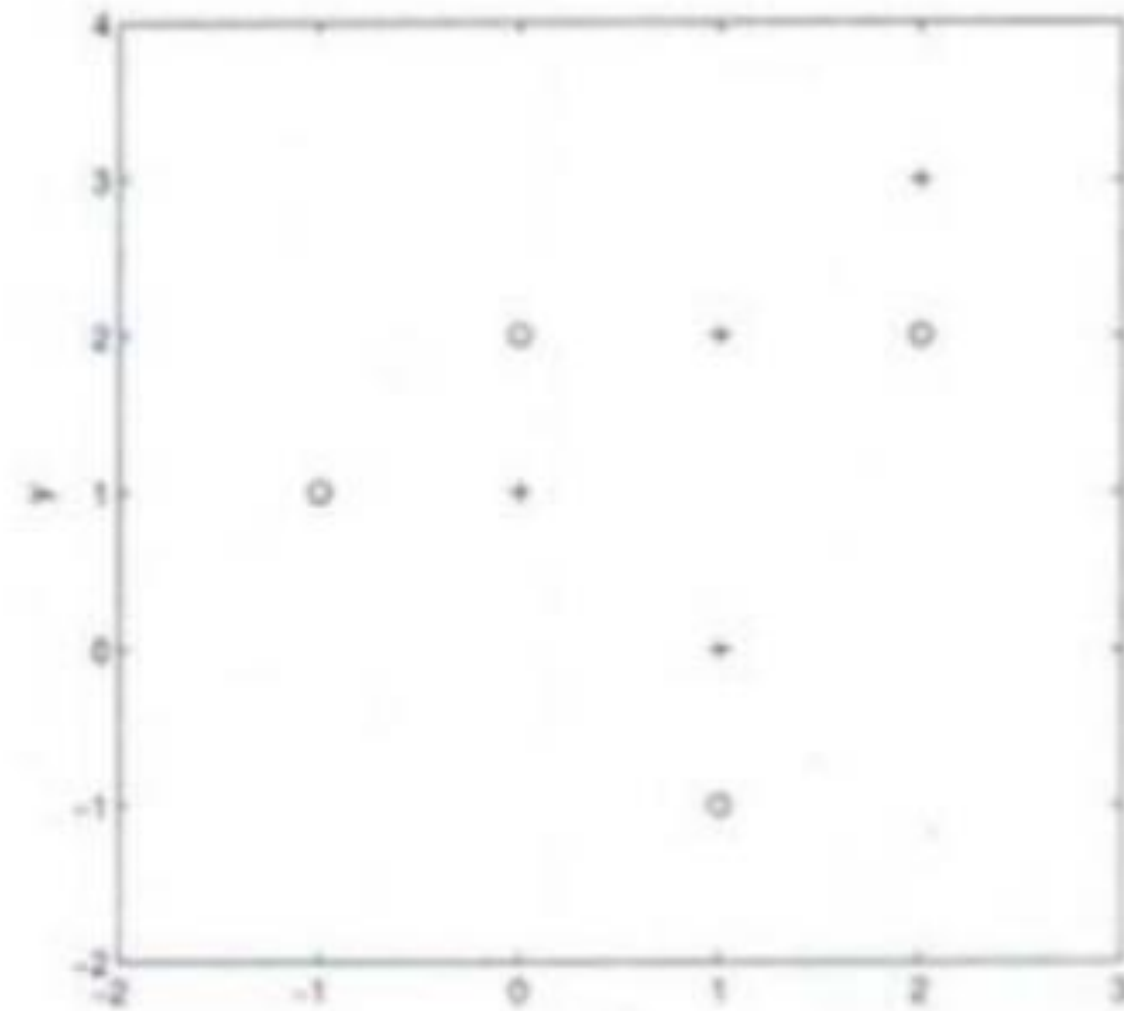
$k=1$   $\diamond$   
 $k=3$   $\diamond$   
 $k=5$   $\square$



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 Euclidean 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:

$$\begin{aligned}
 &((1,1), -1) \longrightarrow 2+5 = 7 \\
 &\{((1,7), +1) \longrightarrow 2+1 = 3 \\
 &\{((3,3), +1) \longrightarrow 0+3 = 3 \\
 &((5,4), -1) \longrightarrow 2+2 = 4 \\
 &\{((2,5), -1) \longrightarrow 1+1 = 2
 \end{aligned}$$

$$(x_1, y_1) \neq (x_2, y_2)$$

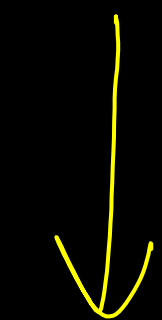
$$|y_2 - y_1| + |x_2 - x_1|$$

$$\longrightarrow +1$$

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



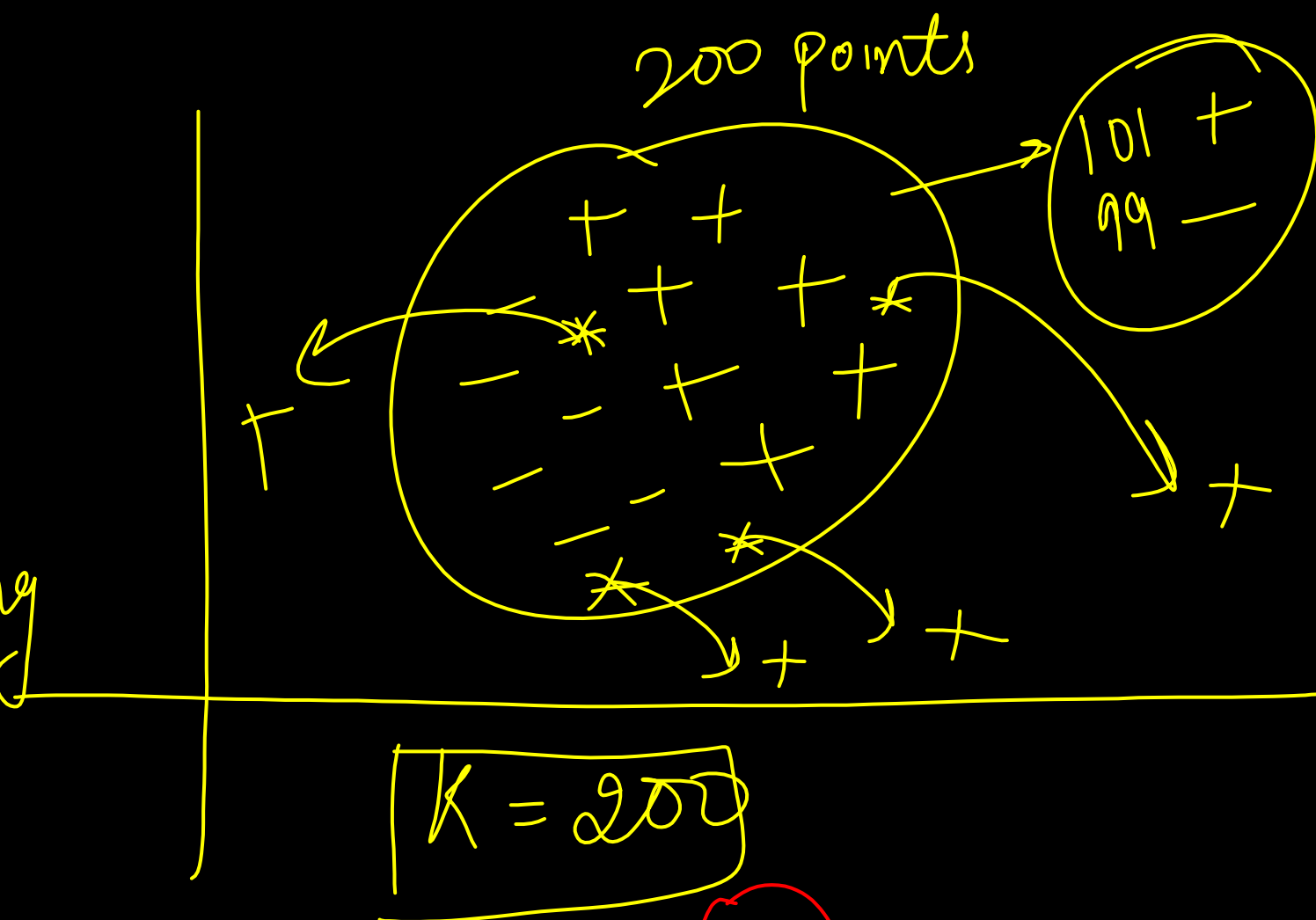
Large value of  $k$



underfitting



high bias



Small value of  $k$



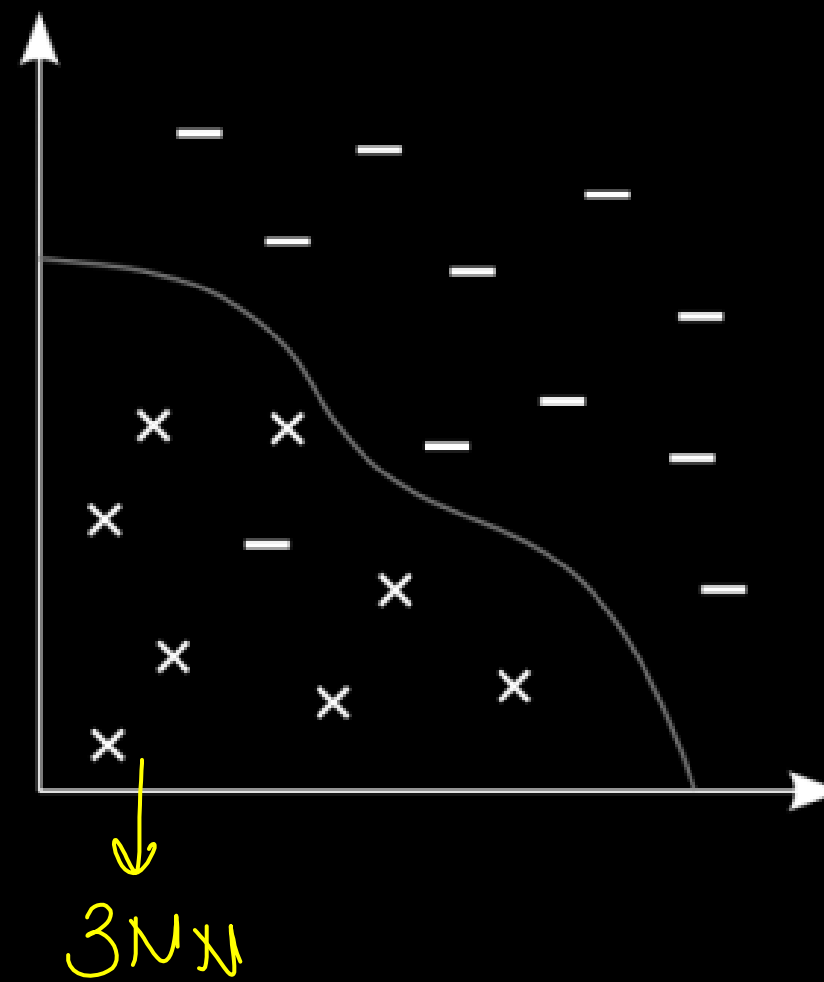
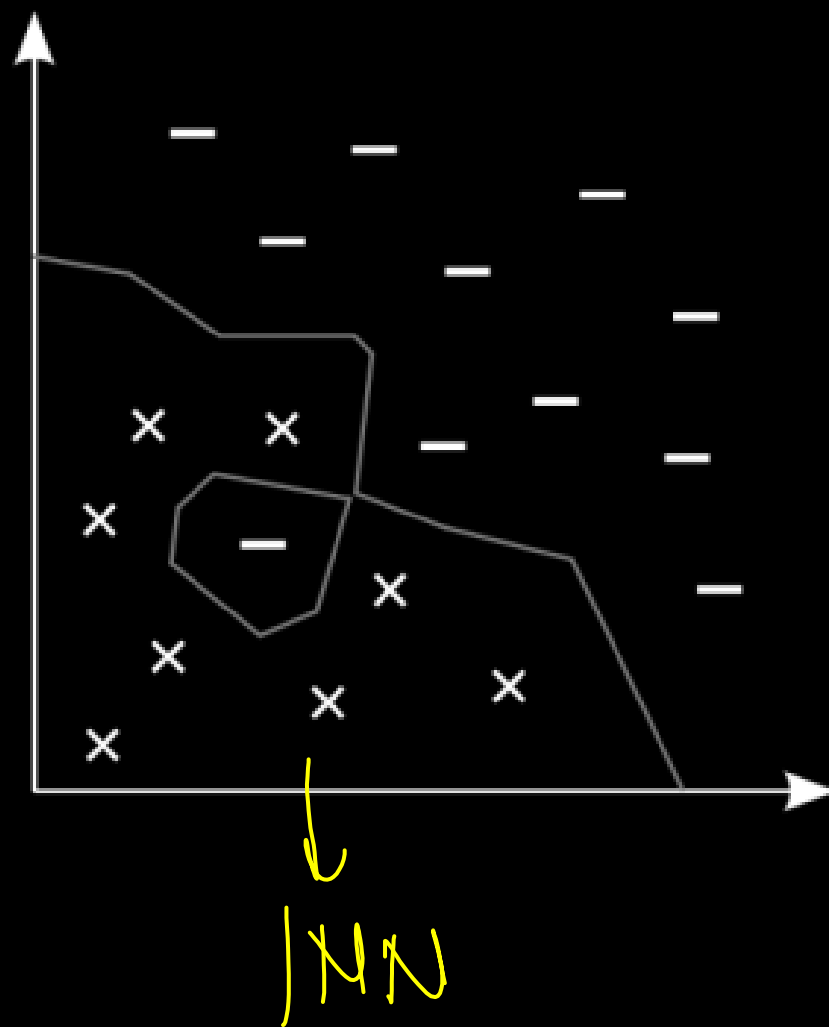
Overfitting



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?

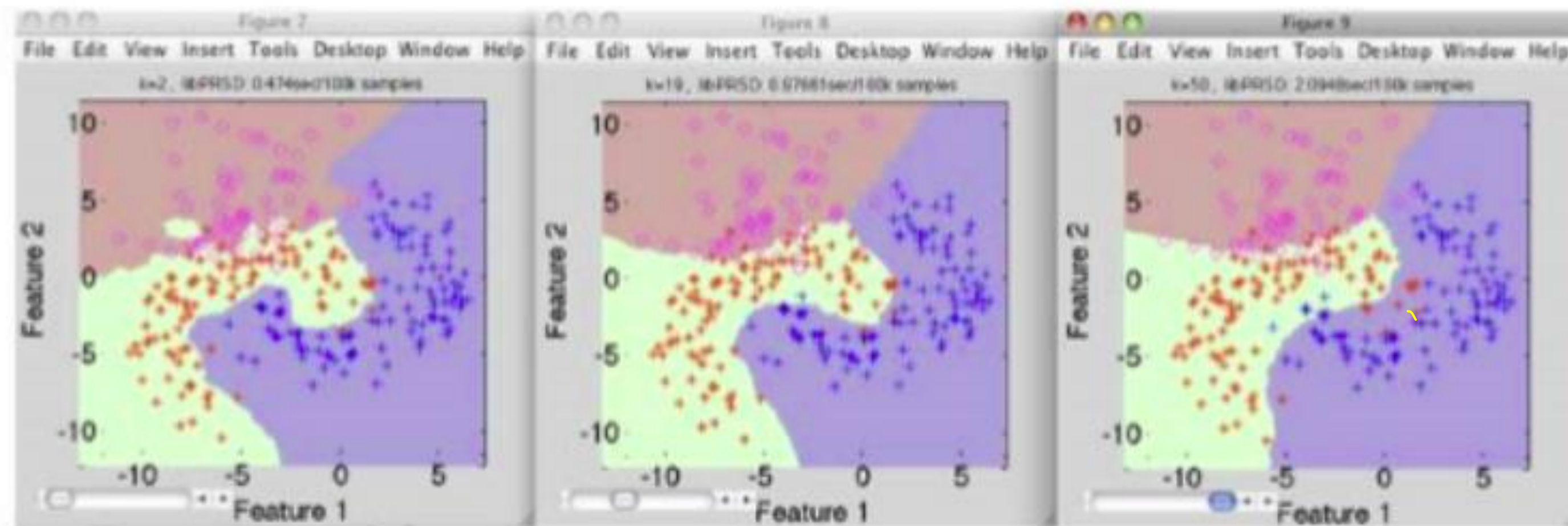
- a) Increasing  $K$  increases bias and decreases variance.
- b) Increasing  $K$  decreases bias and increases variance.
- c) Increasing  $K$  has no effect on bias or variance.
- d) Increasing  $K$  decreases both bias and variance.

overfitting  $\rightarrow$  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  $k_1$  is for 1<sup>st</sup>,  $k_2$  is for 2<sup>nd</sup> and  $k_3$  is for 3rd figure.



- a.  $k_1 > k_2 > k_3$
- b.  $k_1 < k_2 > k_3$
- c.  $k_1 < k_2 < k_3$
- 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

Handwritten calculations for distances from (2,3):

- Distance to (2,5):  $\sqrt{0^2 + 2^2} = 2$
- Distance to (5,5):  $\sqrt{3^2 + 2^2} = \sqrt{13}$
- Distance to (3,3): 1
- Distance to (6,3): 4
- Distance to (2,2): 1
- Distance to (4,1):  $\sqrt{8}$

Closest points are (2,5), (3,3), and (2,2) with Y values 3.4, 3, and 2 respectively.

A. 2.0

B. 2.6

☒ C. 2.8

D. 3.2

$$\frac{3 \cdot 4 + 3 + 2}{3} = \frac{8 \cdot 4}{2} = 2.8$$

Answer: (C)



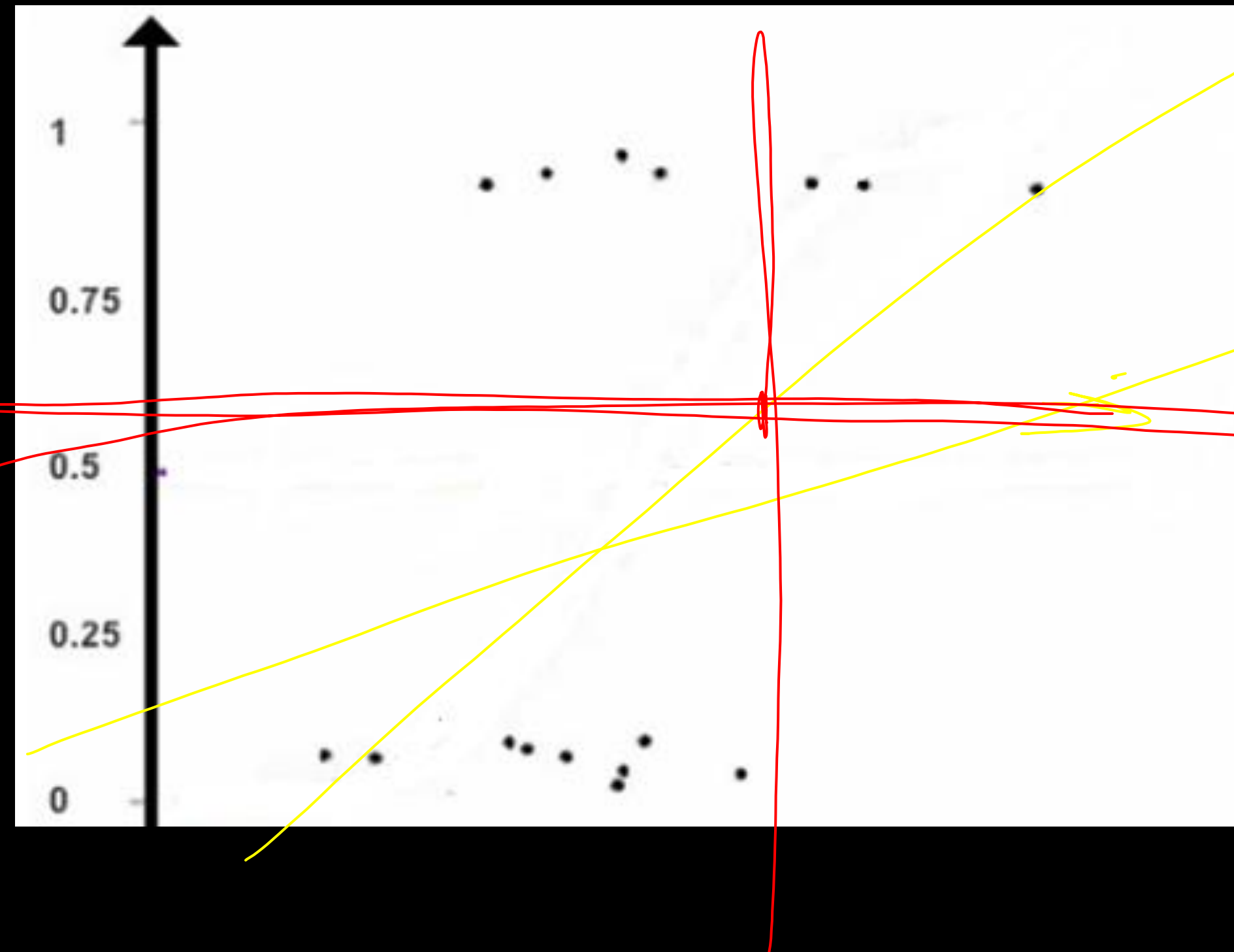
# Classification



# Logistic Regression

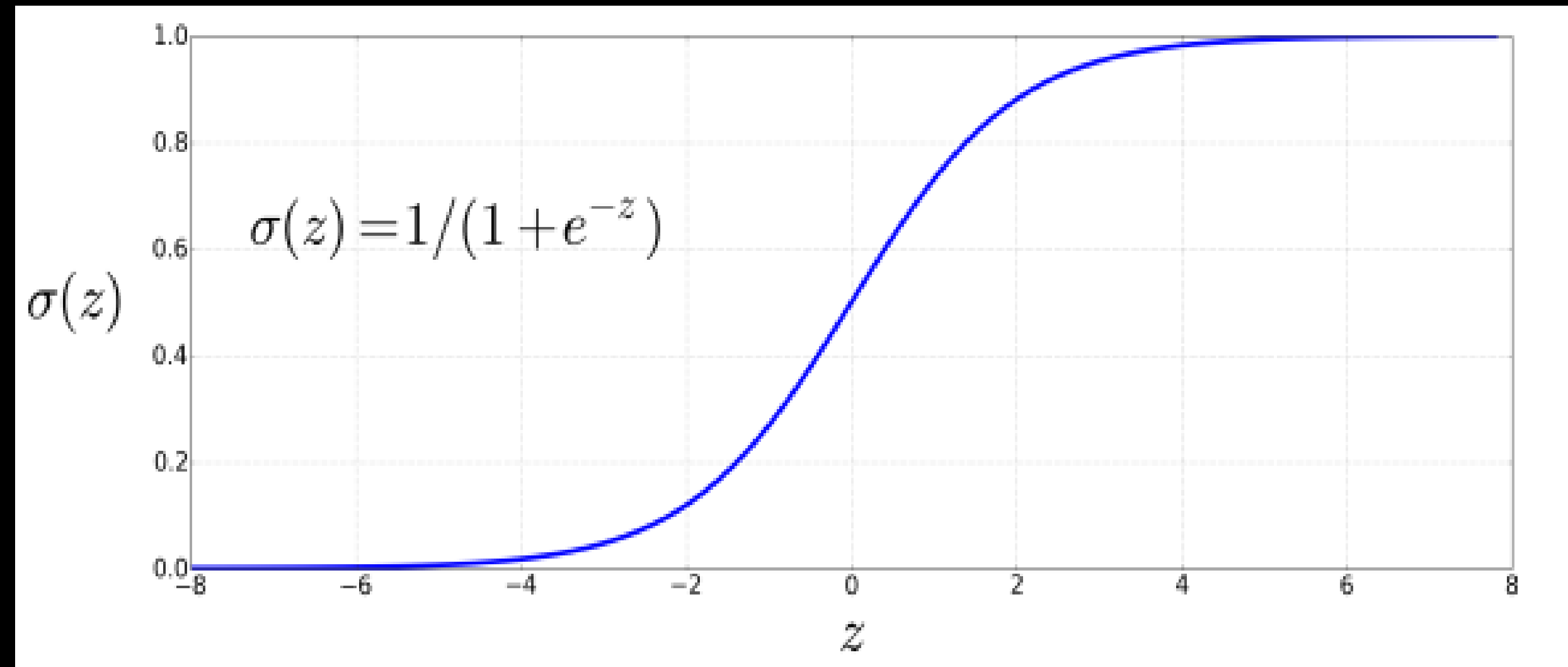
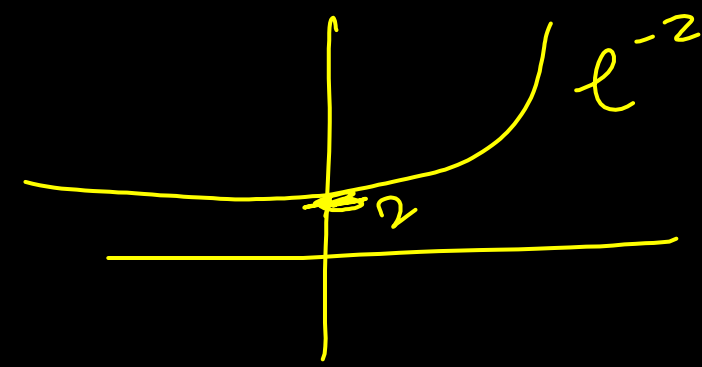
$x$

$f \geq 0.5 \rightarrow 1$   
 $f < 0.5 \rightarrow 0$



$$f(x) = wx + b$$

# Sigmoid Function



$$z = wx + b$$

$$\begin{aligned}\sigma(z) &= \frac{1}{1 + e^{-z}} \\ &= \frac{1}{1 + e^{-(wx+b)}}\end{aligned}$$

$$\begin{aligned}\sigma(z) &= P(y=1) \\ 1 - \sigma(z) &= P(y=0)\end{aligned}$$

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

$$\begin{aligned}w &= 2, b = 1 \\ \sigma(z) &= \frac{1}{1 + e^{-2x+1}} \\ &= \frac{1}{1 + e^{-1}}\end{aligned}$$

$$P(y=1) = \sigma(\mathbf{w} \cdot \mathbf{x} + b)$$

$$= \frac{1}{1 + \exp(-(\mathbf{w} \cdot \mathbf{x} + b))}$$

$\Rightarrow \sigma(z)$

$$P(y=0) = 1 - \sigma(\mathbf{w} \cdot \mathbf{x} + b)$$

$$= 1 - \frac{1}{1 + \exp(-(\mathbf{w} \cdot \mathbf{x} + b))}$$

$$= \frac{\exp(-(\mathbf{w} \cdot \mathbf{x} + b))}{1 + \exp(-(\mathbf{w} \cdot \mathbf{x} + b))}$$

$\Rightarrow 1 - \sigma(z)$

Goal  $\mathbf{w} \cdot \mathbf{x} + b$

$$(y - y')^2$$

1	0
0	1

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

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

## Decision boundary

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

## Loss Function

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

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

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

$$-L = y_1 \log z_1 + (1-y_1) \log z_2$$

$$y=1$$

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

$$y=0$$

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

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

$$z = wx + b$$

## Gradient Descent Implementation

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

$$\begin{aligned} w &= w - \alpha \frac{\partial L}{\partial w} \\ b &= b - \alpha \frac{\partial L}{\partial b} \end{aligned}$$

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

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

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

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

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

$$(y - \hat{y})$$

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

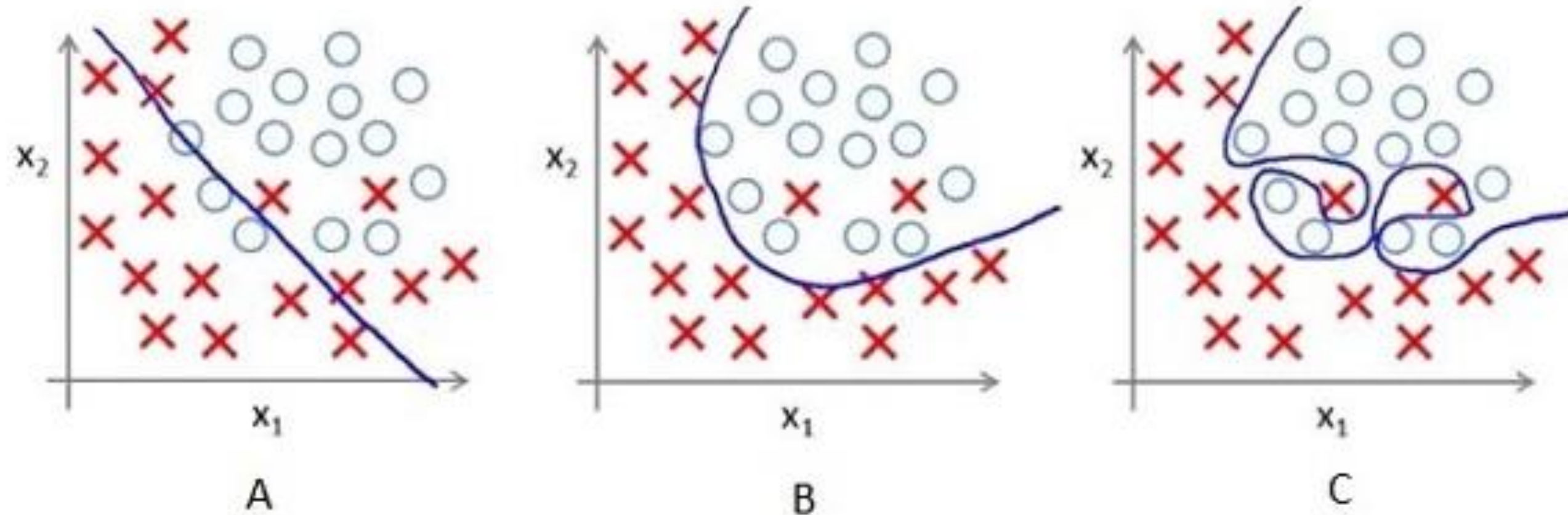
a.  $g(z)(g(z)-1)$

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

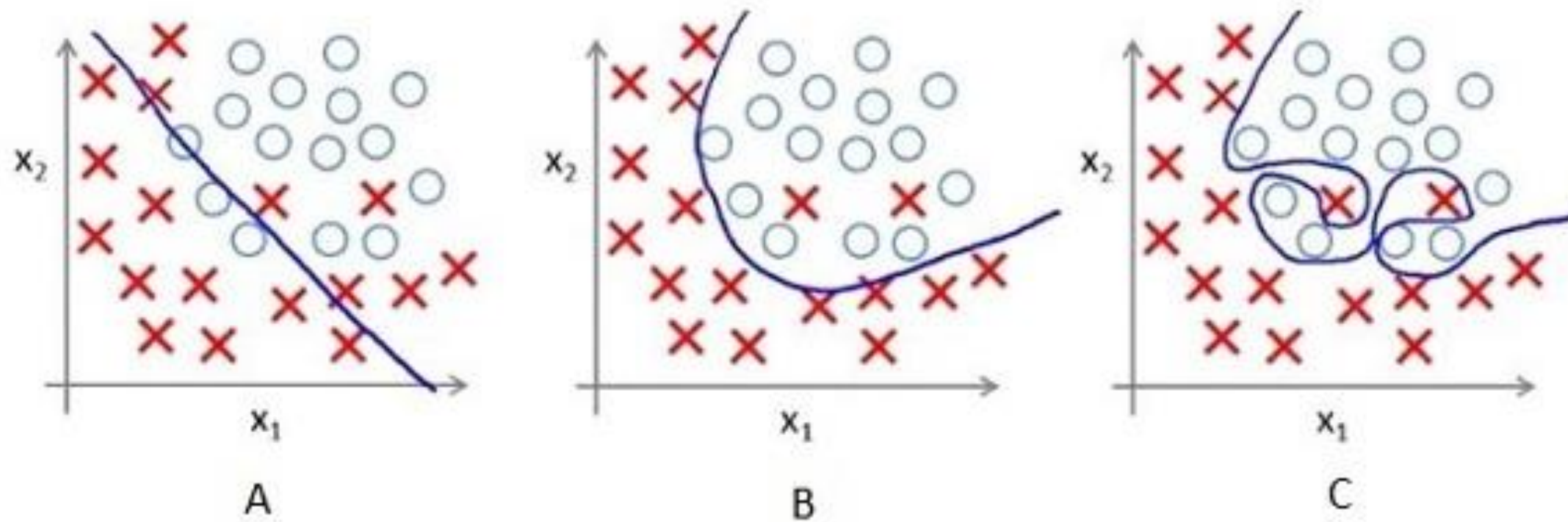
B) B

C) C

D) None of these

Answer: (C)

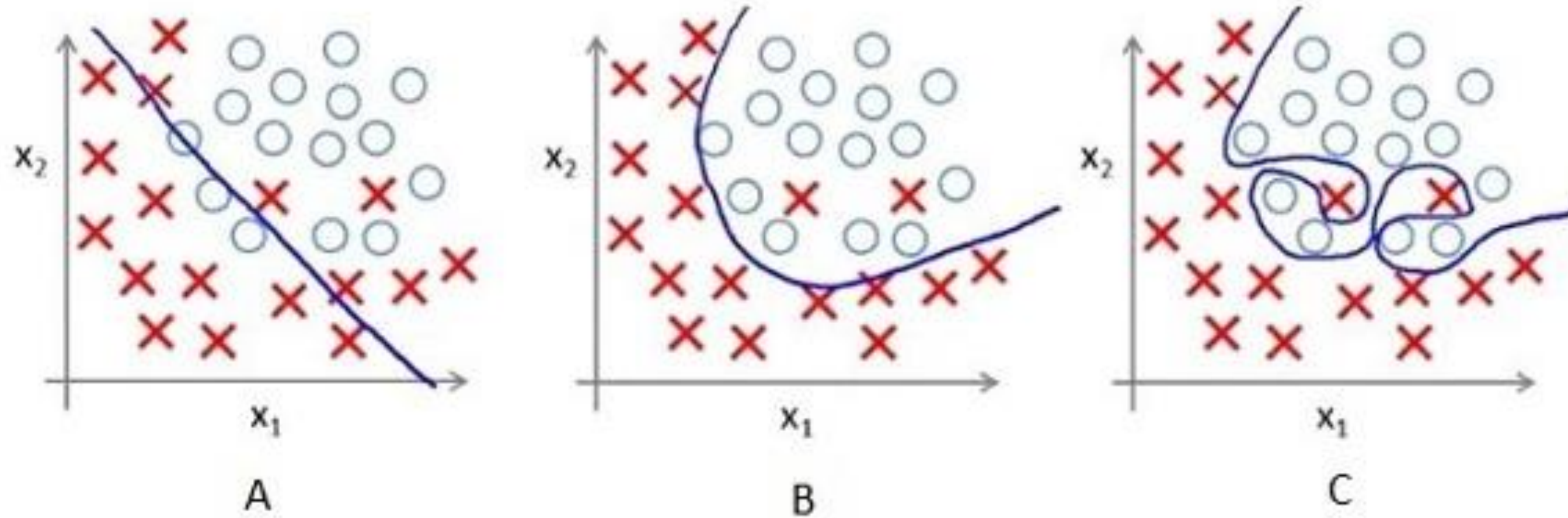




**What do you conclude after seeing this visualization?**

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)

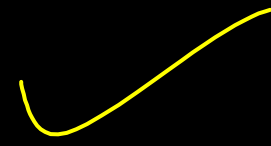
Which of the following are symptoms of a logistic regression model being overfit? Select all that apply

- ☒ (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)

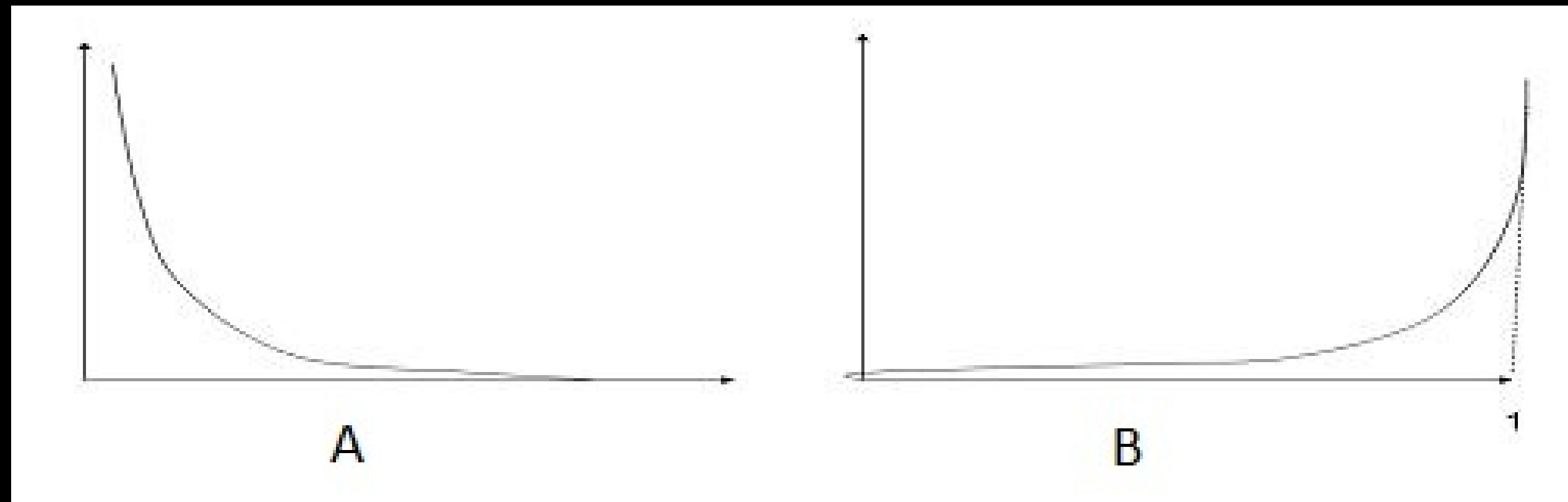
Suppose you are using a Logistic Regression model on a huge dataset. One of the problem you may face on such huge data is that Logistic regression will take very long time to train.

- 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

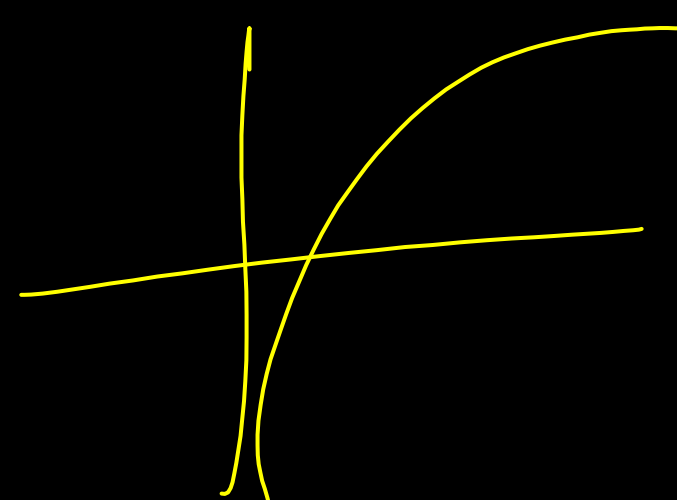
**Solution: D**



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



logistic regression  
 $\mathcal{L}(\hat{y}) = -\log \hat{y}$



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

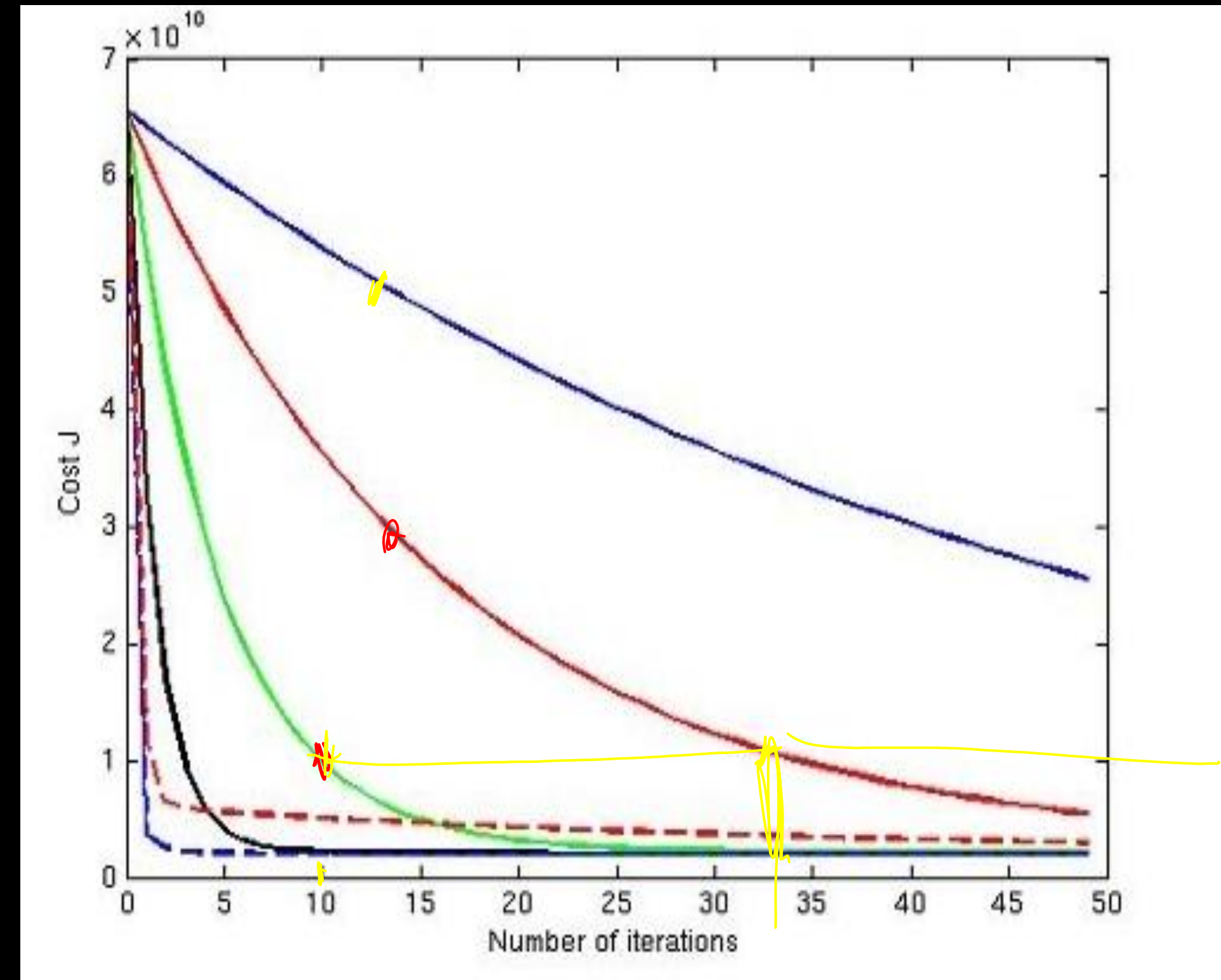
**Solution: A**

Imagine, you have given the below graph of logistic regression which 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  $l_1$
2. The learning rate for red is  $l_2$
3. The learning rate for green is  $l_3$

- A)  $l_1 > l_2 > l_3$
- B)  $l_1 = l_2 = l_3$
- C)  $l_1 < l_2 < l_3$
- D) None of these

**Solution: C**





**Thank you**