21.5/30

Machine Learning CT1

Total Marks: 30 **Time: 1 Hour**

Name: Jannatul Ferdoush Jannati Roll 349

1. What type of Machine Learning approach you will follow for the following set of scenarios? 3

a) Given: training data (without desired outputs)

ML Approach: unsupervised/Clustering

b) Given: training data + desired outputs (labels)

ML Approach: supervised / Classification

c) Given: training data + a few desired outputs

ML Approach: semi-supervised

3

2. When supervised learning fails in ML? 2

for numeric ~~values~~ attributes. The issues are:

- Because of outliers it shows skewed distribution
- Non monotonous effect of attributes

When we provide completely different dataset for test

1

3. How to handle missing data in a dataset? 3

Categorical : put N/A

Numerical: any of these

- put mean
- remove the instance
- remove the entire attribute column from all instances

2.5

4. What is a class imbalanced dataset in machine learning? 3

If a dataset has so many data of one class label and a very few of another class label then the dataset is imbalanced. It is not good for ML. for example,

Happy	Sad	Play
1	0	Yes
0	1	Yes
1	1	Yes

→ no other class label given.

3

5. continued

$P(Pos) \text{ Play cricket sad}) = P(Pos) * P(Play|Pos) *$
$P(cricket|Pos) * P(sad|POS)$
$= 0.5 * \frac{2}{11} * \frac{2}{11} * \frac{2}{11} = 0.0030$

$P(Neg|\text{Play cricket sad}) = P(Neg) * P(Play|Neg) * P(cricket|Neg) *$
$P(sad|POS)$
$= 0.5 * \frac{2}{12} * \frac{2}{12} * \frac{3}{12} = 0.0035$

As, Negative probality is greater so the given new twitter is negative.

7. continued Taking Attribute = Home Owner, Attribute = Marital Status

Home Owner	Yes	No	Entropy
Yes	0	3	0
No	3	4	0.985

Marital Status	Yes	No	Entropy
Single	2	2	1
Married	0	4	0
Divorced	1	1	1

$I(\text{Home owner}) = \frac{0+3}{10} \times 0 + \frac{3+4}{10} \times 0.985$
$= 0.6895$

$Gain(\text{Homewoner}) = E(S) - I(\text{Home owner})$
$= 0.881 - 0.6895$
$= \boxed{0.1915}$

$I(\text{marital status}) =$
$= \frac{2+2}{10} \times 1 + \frac{0+4}{10} \times 0 +$
$\frac{1+1}{10} \times 1$
$= 0.6$

$G(\text{Marital Status})$
$= 0.881 - 0.6$
$= \boxed{0.281}$

Attribute = Annual Income

Annal income	Yes	No	Entropy
High	0	4	0
low	3	3	1

$I(\text{Annual income}) = \frac{0+4}{10} \times 0 + \frac{3+3}{10} \times 1$
$= 0.6$

continue lastpage.

26.5 / 30

CSE457 - Machine Learning CT-02 Time:25 Mins Marks: 30

1. 5. Solve the problem following Naïve Bayes algorithm from a labeled dataset of twitter. Find out whether the new twitter = "I am playing cricket and sad" is negative or positive. 8

Document	Label
I am happy because I am playing cricket.	Positive
I am happy, not sad.	Positive
I am sad, I am not playing cricket.	Negative
I am sad, not happy.	Negative

1

Pos = positive = 2

Neg = Negative = 2

① $P(Pos) = \frac{2}{2+2} = 0.5$

$P(Neg) = \frac{2}{2+2} = 0.5$

②

	Pos	Neg
happy	2+1	1+1
Play	1+1	1+1
cricket	1+1	1+1
not	1+1	2+1
sad	1+1	2+1
	11	12

③ $P(\text{~~Neg~~}\,Play \mid Pos) = \frac{2}{11}$, $P(happy \mid Pos) = \frac{3}{11}$

$P(cricket \mid Pos) = \frac{2}{11}$, $P(not \mid Pos) = \frac{2}{11}$

$P(sad \mid Pos) = \frac{2}{11}$

④ $P(Play \mid Neg) = \frac{2}{12}$, $P(happy \mid Neg) = \frac{2}{12}$,

$P(cricket \mid Neg) = \frac{2}{12}$, $P(not \mid Neg) = \frac{3}{12}$, $P(sad \mid Neg) = \frac{3}{12}$

~~RE~~ P(Pos| "I am playing cricket and sad") = P(Pos|play cricket sad)

← see here

6. Use the naïve Bayes method to determine whether a loan X=(Home Owner = No, Marital Status=Married, Income=High)should be classified as a Defaulted Borrower or not. (Write the calculations only in the blank) 3

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	High	No
2	No	Married	High	No
3	No	Single	Low	No
4	Yes	Married	High	No
5	No	Divorced	Low	Yes
6	No	Married	Low	No
7	Yes	Divorced	High	No
8	No	Single	Low	Yes
9	No	Married	Low	No
10	No	Single	Low	Yes

3

i. P(Yes) = 3/10 = 0.3 and P(No) = 7/10 = 0.7

ii. P(X|No) = P(Home Owner=No|No) x P(Status=Married|No) x P(Income=High|No)

= 4/7 * 4/7 * 4/7 = 0.187

7. Solve the above problem Q.6 using Decision Tree, find the Root node only. ID3 8

Overall entropy $E(S) = -\frac{3}{3+7}\log\left(\frac{3}{3+7}\right) - \frac{7}{3+7}\log\left(\frac{7}{3+7}\right)$

$= 0.881$

Yes = 3
No = 7

← see here.

High variance. Variance

Home owner	p	n	entropy
yes	0	~~8~~ 3	0
No	3	4	0.98

marital status	p	n	entropy
single	2	2	~~0.9~~ 1
married	0	1	0
Divorced	1	1	1

Annual income	p	n	entropy
High	0	4	0
low	3	3	~~0.968~~ 1

∴ I (Home owner) = $\frac{7}{10} \times 0.98 = 0.686$

I (marital status)

$= \frac{4}{10} \times$ ~~0.9~~ $1 + \frac{2}{10} \times 1$

= ~~0.473~~ 0.6

I (Annual income)

$= \frac{6}{10} \times$ ~~0.968~~ 1 = ~~0.484~~ 0.6

∴ Gain (Home owner)

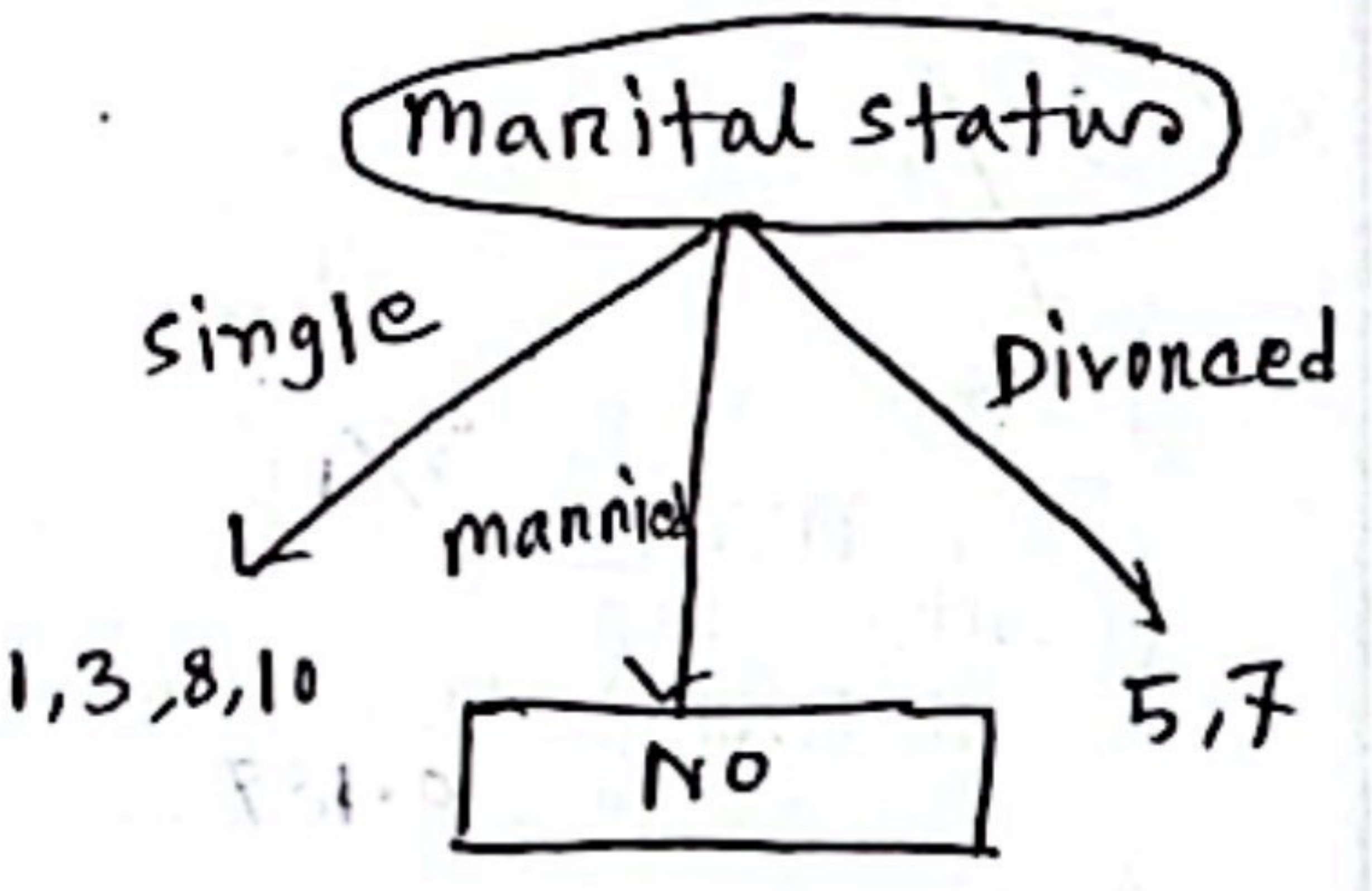
= 0.88 − 0.686 = 0.194

∴ Gain (marital status)

= 0.88 − ~~0.473~~ 0.6 = ~~0.407~~ 0.28 ✓

∴ Gain (Annual ~~pa~~ income)

= 0.88 − ~~0.484~~ 0.6 = ~~0.396~~ 0.28



∴ the root node is Marital status as it's gain is highest

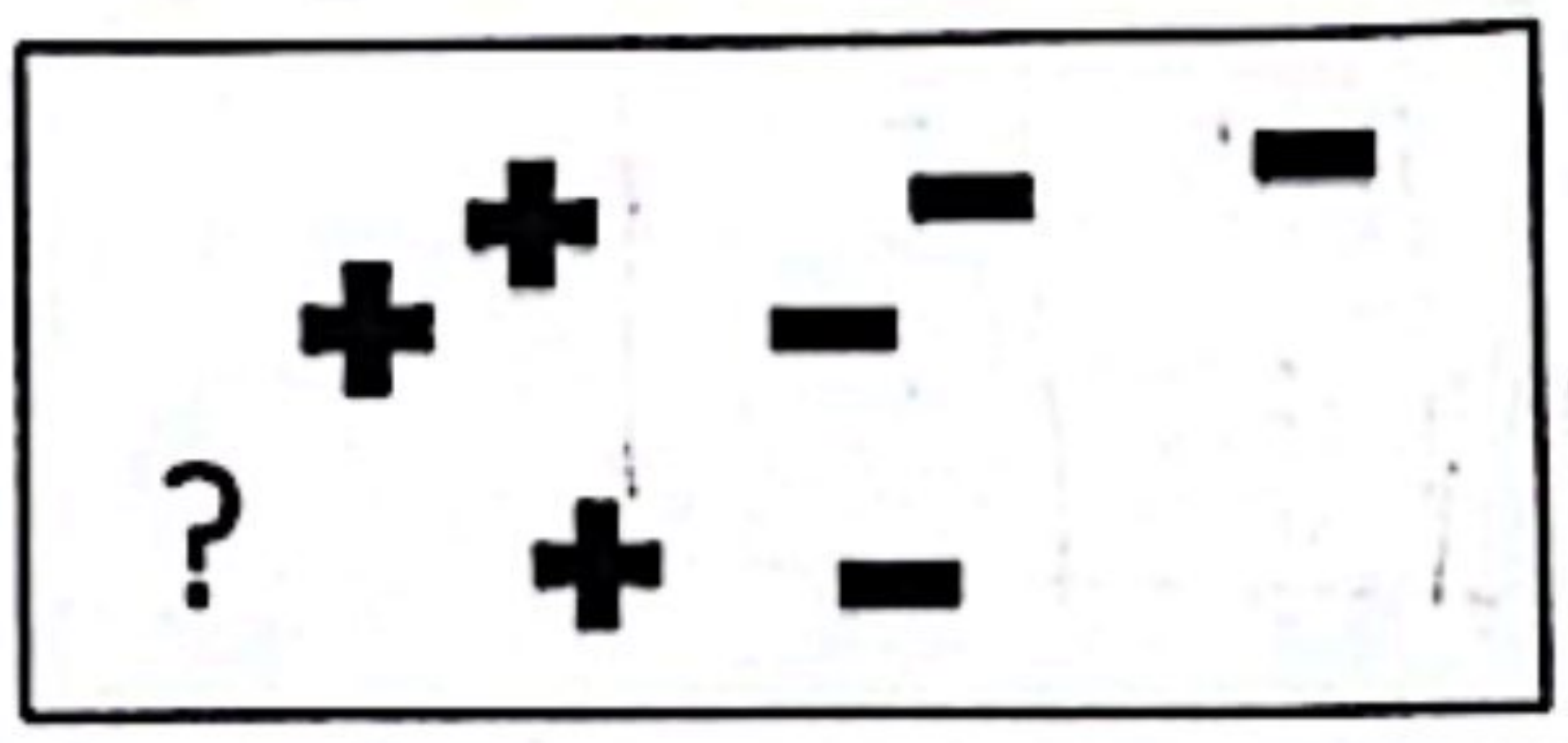
25.75/30

CSE457 - Machine Learning CT-02 Time:25 Mins Marks: 30

Name: Shreya Nag Riya Roll : 338

1. [10] Given the following labeled dataset:

For what value of k(minimal) will be the query point "?" be negative? Show it thoroughly using K-NN. (ties are broken at random, do not consider ties output)



2. [5] How can you identify a *High Bias model*? How can you fix it? [5]

⇒ A model has high bias if the model is unable to cover the true relationship between variables. for ex.: In linear regression method, a straight line has high bias as it is inflexible and unable to cover the true relationship. High bias can be fixed by finding a sweet spot between two models. 3 methods: i) regularization ii) Bagging iii) Boosting

5

3. [5] What happens if you use a very large k on the dataset in K-NN? Why might too small values of k also be bad?

⇒ If I use very large value of k, then a category with a few samples will always be out voted by other categories.

too small values of k (such as k=1 or 2) ~~can~~ might cause noise and subject to the effects of outliers.

5

4. [5] Suppose, we use a logistic regression model to predict whether or not 400 different college basketball players get drafted into the NBA. From the predicted and actual test outcome the confusion matrix formed as the following. Calculate Precision, Recall andAccuracy and F1-measure.

	Actual "Yes"	Actual "No"
Predicted "Yes"	120 TP	40 ~~FN~~ FP
Predicted "No"	70 F~~P~~N	170 TN

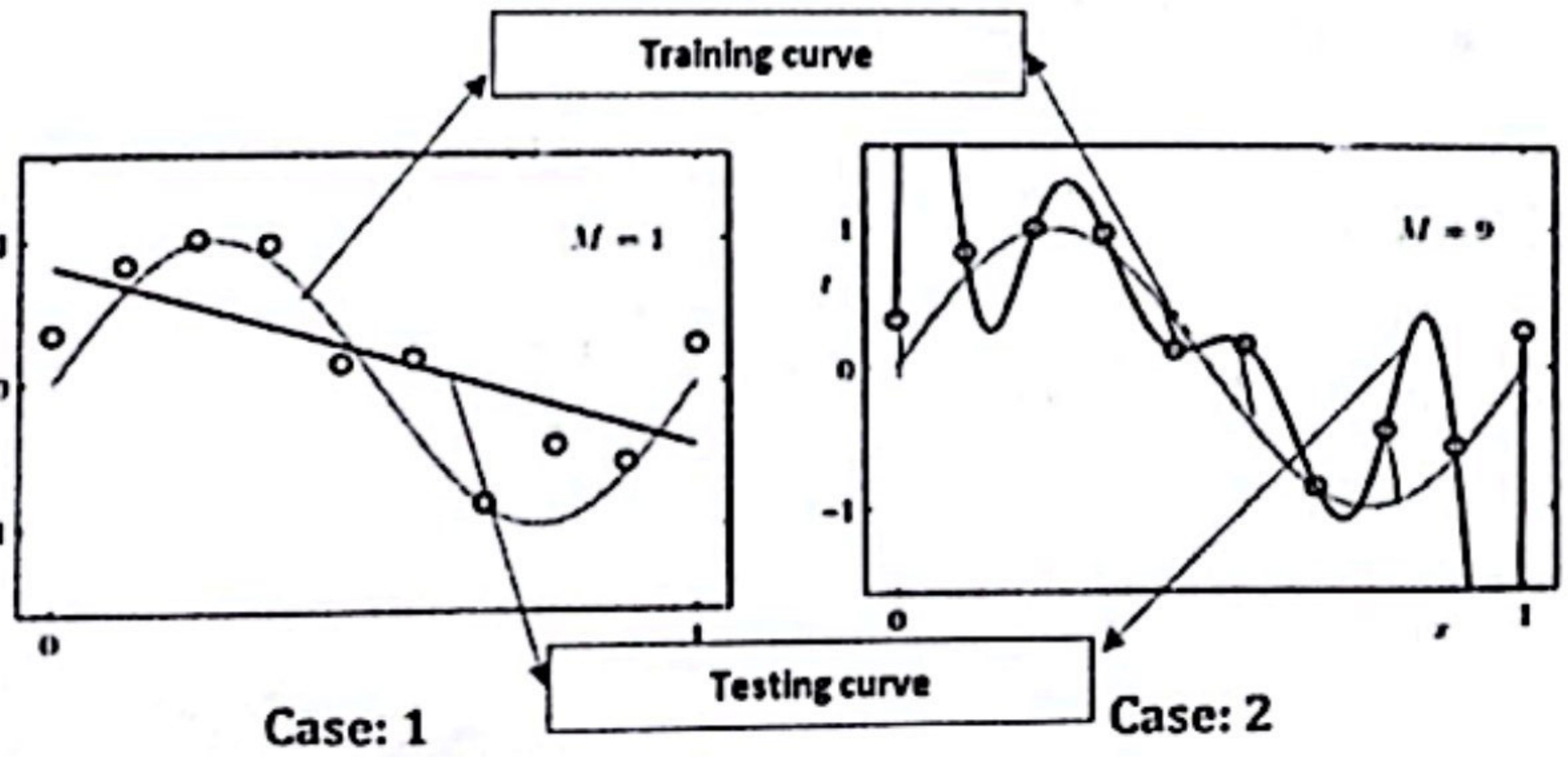
$$\therefore \text{Precision} = \frac{TP}{TP+FP} = \frac{120}{120+40} = \cancel{0.632}\ 0.75$$

$$\therefore \text{Recall} = \frac{TP}{TP+FN} = \frac{120}{120+70} = \cancel{0.75}\ 0.632$$

$$\therefore \text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{120+170}{120+170+70+40} = 0.725$$

3.75

5. [5] Define Overfitting and Underfitting for the following two figures (Case1 and Case 2). What is wrong with them?



Case: 1

Case: 2

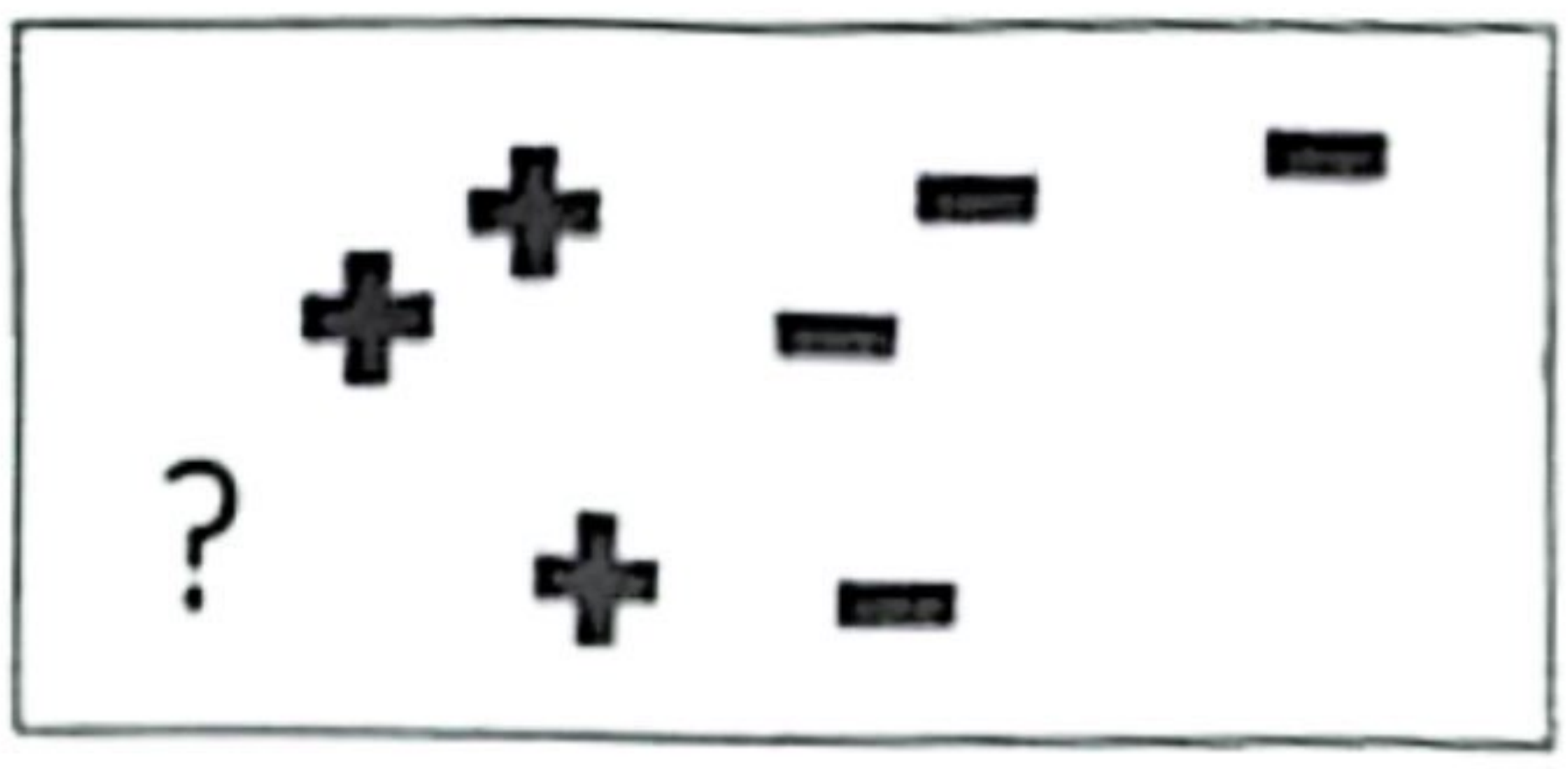
26.5/30

CSE457 - Machine Learning CT-02 Time:25 Mins Marks: 30

Name: Jannatul Ferdowsh Jannati Roll : 349

1. [10] Given the following labeled dataset:

For what value of k(minimal) will be the query point "?" be negative? Show it thoroughly **using K-NN.** (ties are broken at random, do not consider ties output)



2. [5] How can you identify a *High Bias model*? How can you fix it? [5]

Bias means the inability to identify the true relationship of a dataset. For different data points we can calculate the predicted values and ~~see~~ ~~of~~ determine high bias with error mean square. If error mean square is more, bias is high. To fix, we can use squiggly line rather than straight line.

1.5

3. [5] What happens if you use a very large k on the dataset in K-NN? Why might too small values of k also be bad?

If we take large value of K, then cluster with low data points will always be ignored.

To small value for K, can identify ~~noise~~ a data point in a cluster of noise, also can be affected by outliers

5

4. [5] Suppose, we use a logistic regression model to predict whether or not 400 different college basketball players get drafted into the NBA. From the predicted and actual test outcome the confusion matrix formed as the following. Calculate Precision, Recall andAccuracy and F1-measure.

	Actual "Yes"	Actual "No"
Predicted "Yes"	120	40
Predicted "No"	70	170

Precision: $\frac{120}{120+40} = \frac{120}{160} = 0.075$

2

Recall : $\frac{120}{120+70} = 0.63$

Accuracy: $\frac{120+170}{120+40+70+170} =$?

2

5. [5] Define Overfitting and Underfitting for the following two figures (Case1 and Case 2). What is wrong with them?

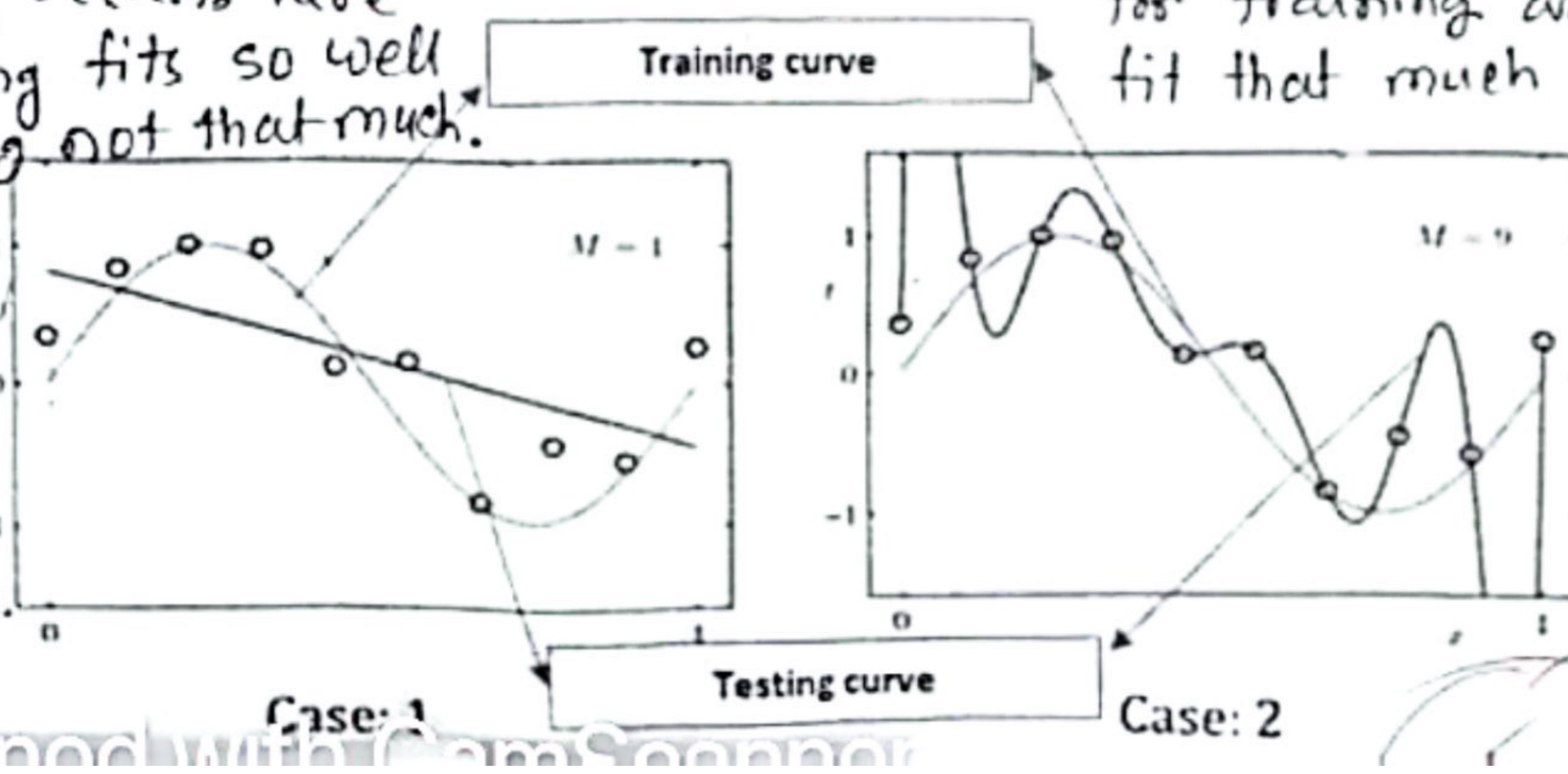
Overfitting occurs here as training fits so well but testing not that much.

~~we can~~

The wrong thing is it has high variance.

Here underfitting occurs for training as ~~it~~ it doesn't fit that much but for testing part, it fits so well.

The wrong thing is that it also has high variance

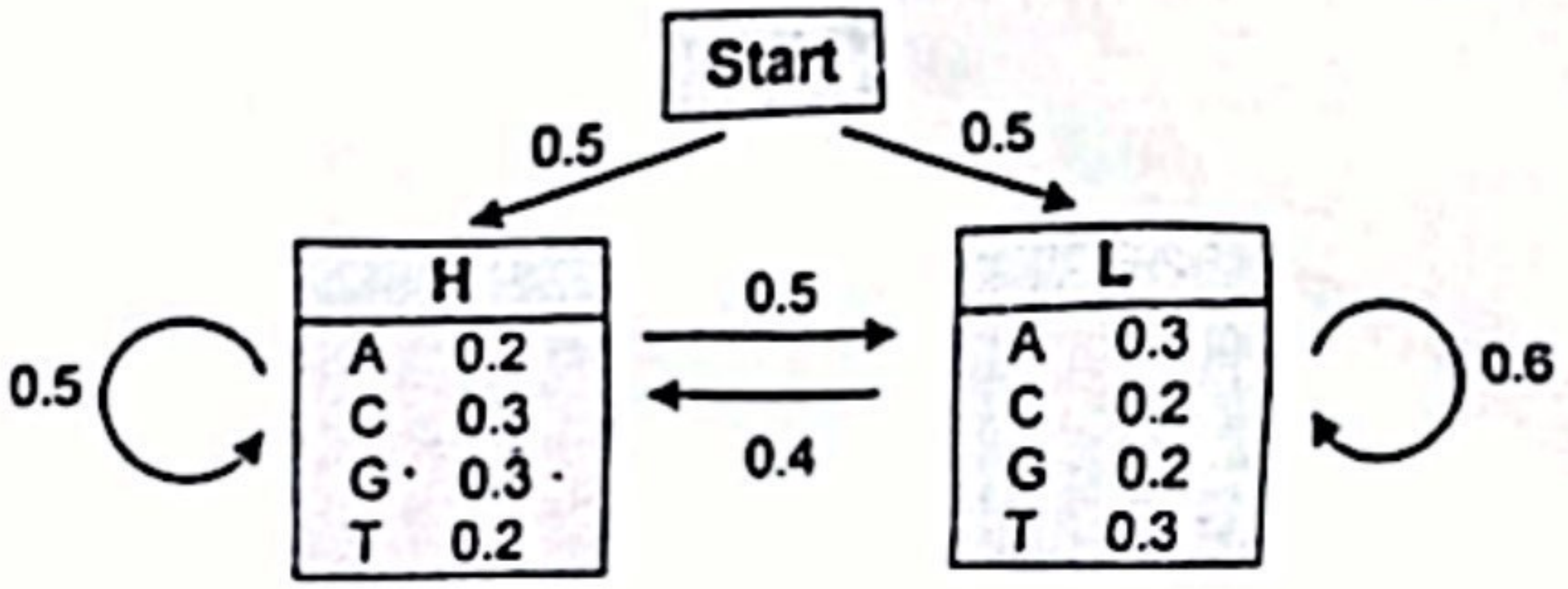


Case: 1

Case: 2

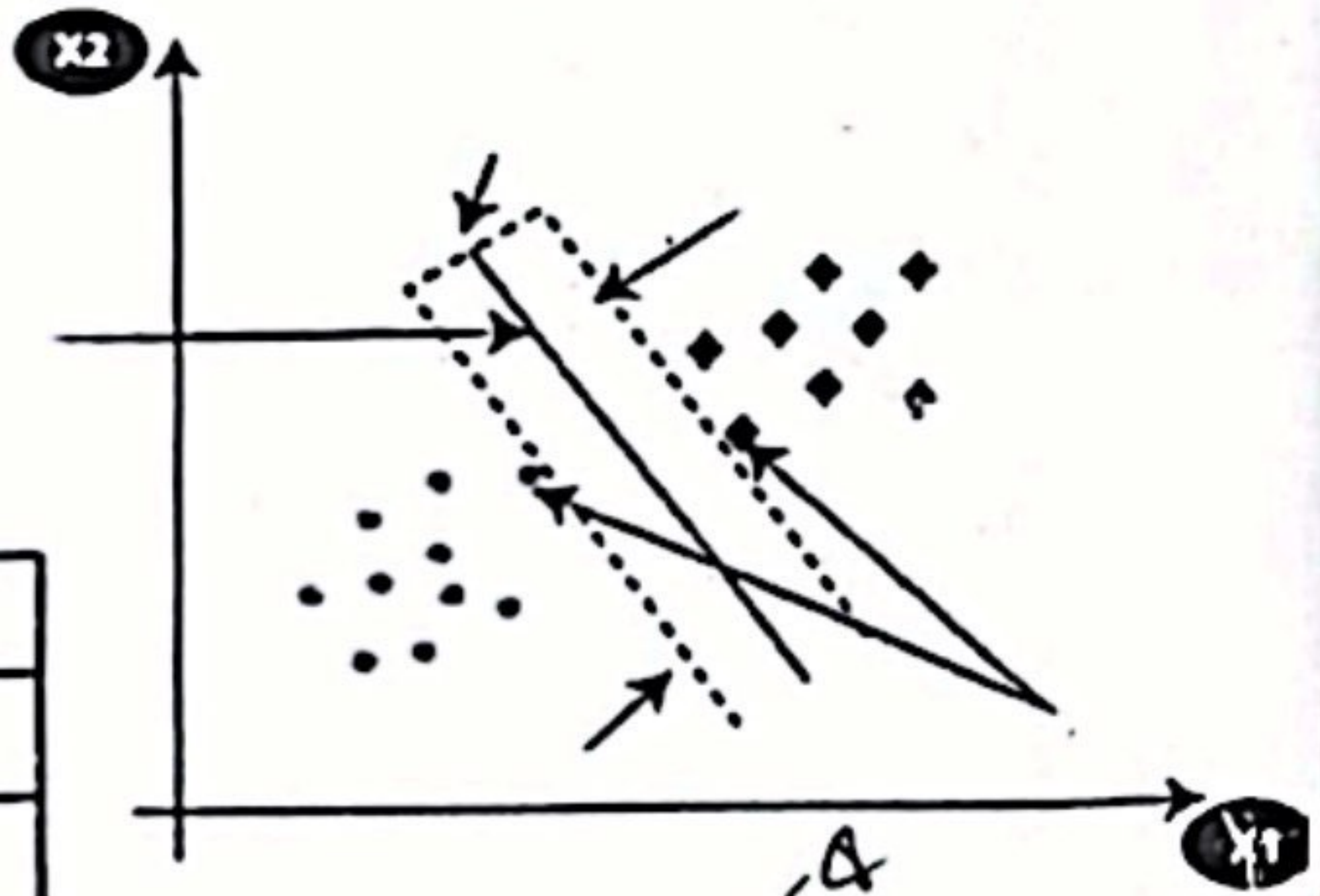
CT-03(Machine Learning) **Time: 25 minutes** **Marks:30**

1. [5points] Consider the Sequence, S = GGCA. There are several paths through the hidden states (H and L) that lead to the given sequence. Find the probabilities of getting outcomes (G, G,C,A) using the Forward Algorithm generating of the following combination. i) (H,L,L,H) ii) (H,H,L,L)



2. [5 points] While clustering analysis in Machine Learning, Inter-cluster distances are minimized and Intra-cluster distances are maximized,"- agree or disagree? Justify your answer.

3. [5 points] *Identify the indicated arrow sign using the SVM algorithm from the following diagram.* →



4. Given the following labeled dataset:

Determine the centers of the new clusters and show the clusters after the first epoch and the new centroids.

EmployeeID	YearService	Income (K)
✓ E01	4	9 ✓
E02	8	4
E03	2	10 ✓
✓ E04	5	8
E05	6	4
E06	7	6

Table 1

1.44×10^{-4}

9.05×10^{-8}

(3, 9.5)

(6.5, 5.25)