Final Examination

2020

5) b) solu:-

Given,

$K = 2$

we select two random representative objects,

$C_1(2,6)$, $C_2(6,5)$

Base on this two objects, we have to do the clustering

. we have to compare by finding minimum distance of both C_1 and C_2 and then make the cluster.

- Through manhattam distance formula, we have to find the distance / cost.

$m = (a,b)$

$n = (c,d)$

$$\text{Distance} = |a-c| + |b-d|$$

- Calculating the distance / cost of ~~C2 (6, 5)~~ object - $c_1(2,6)$

i	x	y	c_1		Distance / cost	C
x_1	a 7	b 6	c 2	d 6	$\lvert 7-2 \rvert + \lvert 6-6 \rvert$	5
x_3	3	8	2	6	1 + 2	3
x_4	8	5	2	6	6 + 1	7
x_5	7	4	2	6	5 + 2	7
x_6	4	7	2	6	2 + 1	3
x_7	6	2	2	6	4 + 4	8
x_8	7	3	2	6	5 + 3	8
x_{10}	3	5	2	6	1 + 1	2

Now,

Similarly, calculate the distance / cost of $c_2(6,5)$ object.

P.T.O

P	x	y	C_2	Distance | cost	c
x_1	7	6	6 5	1 + 1	2
x_3	3	8	6 5	3 + 3	6
x_4	8	5	6 5	2 + 0	2
x_5	7	4	6 5	1 + 1	2
x_6	4	7	6 5	2 + 2	4
x_7	6	2	6 5	0 + 3	3
x_8	7	3	6 5	1 + 2	3
x_{10}	3	5	6 5	3 + 0	3

Now,

compare the cost (c_1) and cost (c_2) for every i, and select the minimum one —

The clusten are —

cluster -1 : $\{(3,8), (4,7), (3,5)\}$

Cluster -2 : $\{(7,6), (8,5), (7,4), (6,2), (7,3),\}$

Now, calculate the cost —

$$T_{cost}(x,c) = \sum_{i=1}^{d} |x_i - c_i|$$

Total cost $= (3+3+2) + (2+2+2+3+3)$

$= 8 + 12 = 20$

Now, Select one of the non-medoids O'.

let $O' = (7,6)$ i.e x_1

Now,

medoids are $C_1(2,6)$ and O' $(7,6)$

Now.

Calculate cost of both C_1 and O' then compare the cost if Cost (C_1) and cost (O') for every i and Select the minimum one.

i	x	y	O'		Distance / cost	C
x_3	3	8	7	6	4+2	6
x_4	8	5	7	6	1+1	2
x_5	7	4	7	6	0+2	2
x_6	4	7	7	6	3+1	4
x_7	6	2	7	6	1+4	5
x_8	7	3	7	6	0+3	3
x_9	6	5	7	6	1+1	2
x_{10}	3	5	7	6	3+1	4

i	x	y	C_1	Distance / Cost	c
x_3	3	8	2 6	1 + 2	3
x_4	8	5	2 6	6 + 1	7
x_5	7	4	2 6	5 + 2	7
x_6	4	7	2 6	2 + 1	3
x_7	6	2	2 6	4 + 4	8
x_8	7	3	2 6	5 + 3	8
x_9	6	5	2 6	4 + 1	5
x_{10}	3	5	2 6	1 + 1	2

Again, Create the cluster —

cluster-1 : {(8,5), (7,4), (6,2), (7,3), (6,5)}

cluster-2 : {(3,8), (4,7), (3,5)}

∴ Current total cost = (2 + 2 + 5 + 3 + 2) + (3 + 3 + 2)

= 14 + 8

= 22

Now,

cost of swapping medoid from C_2 to O's

δ = Current total cost − prev total cost

$= 22 - 20$

$= 2 > 0$

So, moving 0' would be a bad idea. So, previous choice was good.

2021

Q-2

Can K-means algorithms find the global optimum? Justify your ans using an example -

Ans:-

The K-means algorithm is not an ideal option to find the global optimum because to make sure it reaches the global optimum we have to repeat the clustering exercise multiple times i.e. we have to iterate over all possible clusterings and then take the most common solution out of it.

· But this gives the results in exponential runtime.

Example - for justification -

Assume a cluster, $A = \{1, 2, 3, 4\}$

Atleast two different settings satisfy the stationary condition of K-means -

Setting 1 -

center 1 = 1, cluster 1 = $\{1\}$

center 2 = 3 cluster 2 = $\{2, 3, 4\}$

As the objective is 2, & it is a saddle point consider epsilon.

Setting 2

center 1 = 1.5 , cluster1 = {1, 2}

center 2 = 3.5 cluster 2 = {3, 4}

$\Rightarrow$ objective = $\frac{1}{4}$

- If K-means was used as the first (setting 1), it would be stuck without giving us the global optimum.

Hence,

justified.

2020 (End Sem)

5) c) ii)

ii) Solu:— How would DBSCAN handle such data —

Typically,

- empty cluster indicate that, K (number of clusters) is greater than the no. of points.

- If this is happening for 100 points in the given range for K, then, all the points in the ~~given~~ data set ~~range for K~~, must be identical, giving only a single non-

empty cluster.

- Since all points are identical, the similarity / distance matrix would be a matrix of one / zeroes.

- For single link agglomerative clustering, clusters can merge with the points in the order that they appear in the data set.

- Given,

that the distance matrix is zero, all points lie within a single circle with an arbitrary radius, ε.

- with the appropriate value of MinPts, all points will be core points.

- Otherwise, they will all be considered noise

2020

5) c) i)

i) How can use DBSCAN algorithms to detect the outliers?

Ans:- The DBSCAN algorithm is a density based gorithm. It looks at the density of data poin

in a neibourhood to decide whether they belong to the same cluster or not.

- If a point is too ~~fam~~ far from all other points then it is considered as Outlier and is assigned a label of -1.

- DBSCAN algorithm takes multi-dimensional data as inputs and clusters them according to the model parameters — e.g. epsilon & minimum samples.

- Based on these parameters, the algorithm determine whether certain values in the cluster are outliers or not.

— o —

2019 1) 2) → 2020

2) b) with K-fold cross-validation, larger K is always better —

Soln:-

As, we know, when a model is trained ~~along~~ using all of the data in a single shot and give the

best performance accuracy, To resist this, K-fold cross validation helps us to build the model as a generalised one.

Thumb rules Associated with K-fold :

→ K should be always >= 2 & = number of records

- if 2 then 2 iterations only

=> if K = no. of records in dataset (say n), then 1 for testing & (n-1) for training

→ optimized value of K is 10 and commonly used with the data of good size.

- If K values is too large, it will lead to less variance across the training set.

- The no. of folds is indirectly proportional to the size of the data set i.e. folds ↑ $\propto \frac{1}{\text{dataset size} \downarrow}$

→ Larger values of K eventually, increase the running time of the cross-validation process.

The ans is No.

Good choice of K (say 10) is good for large data, but that's not the case always, as increasing folds reduces biases or variance in training sets but it increases the run-time of solution -

Hence,

inspite of this slow-execution runtime K-fold cross validation, plays a critical role in ML world, But with K-fold CV, larger K is not always better

1) a) vi)

2) a) Overfitting is a danger when Learning a classifier, but not when doing unsupervised learning

Solu:-

Yes, it's true. Overfitting has little to do with whether the setting is supervised or unsupervised. Essentially, you can break your data points into two components - pattern + stochastic noise.

The goal of M.L is to model the pattern & ignore the noise.

- In the supervised setting, we typically want to match the output of a prediction function to our training labels.

- In the supervised setting, since the noise is stochastic, this doesn't generalize well to unseen data & we have low training error but high test error.

- In unsupervised setting, we have some notion of the quality of the solution.

- The distinction b/w pattern & noise is not obvious in most cases. So, we don't have foolproof methods to model only the pattern & ignore the noise completely.

δ = Current total cost − past total cost

$= 22 - 20$

$= 2 > 0$

So, moving O' would be a bad idea. So, previous choice was good.

2020

1) b) solu:-

i) A rule that has high support and high confidence −

Milk → Bread. / Cereal → Milk

Such obvious rule tends to be uninteresting. Nothing to learn from this rule, as this items are far too common.

ii) A rule that has reasonably high support but low confidence −

Milk → Cereal.

Same support as above, but milk has other uses. Not subjectively interesting.

iii) A rule that has low support and low confidence —

Batteries → Toothpaste.

Not as frequent. No common use cases. Not subjectively interesting.

iv) A rule that has low support and high confidence —

Strawberries → chocolate.

Occur relatively, frequently together.

More interesting.

1) a) ii) Outliers are influential observations —

An outliers is a point with large residual. Outliers are the data points those diverge by good margin from the overall pattern. It can have extreme X or Y values or both compared to other values.

An influential points is a point that has a large

impact on the regression.

- It is an outlier that impacts the slope of the regression line.

- To ~~test~~ test the influence of an outlier is to compute the regression equation with and without the outlier.

So,

~~the~~ an outlier is a data point that diverges from an overall pattern in a sample. Therefore, an outlier can certainly influence the relationship b/w the variables ~~and~~ may also exert an influence on the slope of the regression line.

4) b) Solu:-

Instance	a_1	a_2	a_3	Target class
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	−
4	F	F	4.0	+
5	F	T	7.0	−
6	F	T	8.0	−
7	F	F	9.0	−
8	T	F	7.0	+
9	F	T	5.0	−

a) There are four positive examples, and 5 negative examples. Thus,

$P(+) = \frac{4}{9}$ and $P(-) = \frac{5}{9}$

The entropy of the training example is

$$-\frac{4}{9}\log_2\left(\frac{4}{9}\right) - \frac{5}{9}\log_2\left(\frac{5}{9}\right)$$

$$= 0.9911$$

b) For attribute a_1, the corresponding counts and probabilities are -

a_1	+	−
T	3	1
F	1	4

The entropy for a_1 is,

$$\frac{4}{9}\left[-\left(\frac{3}{4}\right)\log_2\left(\frac{3}{4}\right)-\left(\frac{1}{4}\right)\log_2\left(\frac{1}{4}\right)\right]$$

$$+\frac{5}{9}\left[-\left(\frac{1}{5}\right)\log_2\left(\frac{1}{5}\right)-\left(\frac{4}{5}\right)\log_2\left(\frac{4}{5}\right)\right]$$

$$= 0.7616$$

So, The information gain for a_1 is

$0.9911 - 0.7616 = 0.2294$.

For attribute a_2, the corresponding counts and probabilities are –

a_2	+	-
T	2	3
F	2	2

The **Entropy for a_2** is,

$$\frac{5}{9}\left[-\left(\frac{2}{5}\right)\log_2\left(\frac{2}{5}\right)-\left(\frac{3}{5}\right)\log_2\left(\frac{3}{5}\right)\right]+\frac{4}{9}\left[-\left(\frac{2}{4}\right)\log_2\left(\frac{2}{4}\right)\right.$$

$$\left.-\left(\frac{2}{4}\right)\log_2\left(\frac{2}{4}\right)\right] = \underline{\underline{0.9839}}$$

$\therefore$ Information gain for a_2 is $0.9911 - 0.9839$

$= 0.0072$

5)a) Solu:-

i) {18,40} and {42,48,

clusters are {6,12,18,24,30} ~~42,48~~ ~~54~~

with total square error

5)a) solu:- Given, one-dimensional points : {6,12,18,24,30,42,48}

• {18,45}

clusters are : {6,12,18,24,30} and {42,48}

with total square error $((12)^2 + 6^2 + 0^2 + 6^2 + 12^2)$

$+((3)^2 + 3^2)$

$= 360 + 18 = 378$

• {15,40}

clusters are {6,12,18,24} and {30,42,48}

with total square error $(9^2 + 3^2 + 3^2 + 9^2) + (10^2 + 2^2 + 8^2)$

$= 180 + 168 = 348$

, 3, 4, 5, 6, 7, 7.

5)a)

Set one one-dimensional points: $\{6, 10, 18, 24, 32, 42, 48, 54\}$

i) (18, 40)

cluster are - $\{6, 10, 18, 24\}$ and $\{32, 42, 48, 54\}$

with total square error

$$= (12^2 + 8^2 + 0^2 + 6^2) + (8^2 + 2^2 + 8^2 + 14^2)$$

$$= 244 + 328$$

$$= 572$$

ii) (15, 45)

cluster are - $\{6, 10, 18, 24, 32\}$ and $\{42, 48, 54\}$

with total square error

$$= (9^2 + 5^2 + 3^2 + 9^2 + 17^2) + (3^2 + 3^2 + 9^2)$$

$$= 485 + 99$$

$$= 584$$

iii) (10, 32).

Clusters are : {6, 10, 18, 24} and (32, 42, 48, 54)

with total square error

$$\left(4^2 + 0^2 + 8^2 + 14^2\right) + \left(0^2 + 10^2 + 16^2 + 22^2\right)$$

$= 276 + 840$

$= 1176$

—o—

3) a) solu:-

There are four positive examples i.e $Y = 4$,

and 5 no i.e $N = 5$

So, Entropy of the training example is

$$-\frac{4}{9}\log_2\left(\frac{4}{9}\right) - \frac{5}{9}\log_2\left(\frac{5}{9}\right)$$

$= 0.9911$

For attribute a_1, corresponding counts and probabilities are -

a_1	Y	N
T	3	1
F	1	4

The entropy for a_1,

$$\frac{4}{9}\left[-\frac{3}{4}\log_2\left(\frac{3}{4}\right)-\left(\frac{1}{4}\right)\log_2\left(\frac{1}{4}\right)\right]$$

$$+\frac{5}{9}\left[-\frac{1}{5}\log_2\left(\frac{1}{5}\right)-\left(\frac{4}{5}\right)\log_2\left(\frac{4}{5}\right)\right]$$

$= 0.7616$

$\boxed{\text{The IG} = 0.9911 - 0.7616 = 0.2294}$

For **attribute a_2**, the corresponding counts and probabilities are —

a_2	Y	N
T	2	3
F	2	2

The entropy for a_2,

$$\frac{5}{9}\left[-\frac{2}{5}\log_2\left(\frac{2}{5}\right)-\frac{3}{5}\log_2\left(\frac{3}{5}\right)\right]+\frac{4}{9}\left[-\left(\frac{2}{4}\right)\log_2\left(\frac{2}{4}\right)-\frac{2}{4}\log_2\left(\frac{2}{4}\right)\right]$$

$= 0.9839$

So, **I.G = 0.9911 − 0.9839**
= 0.0072

4)c) Solv: —

Given, classifying arribute = "Class"

out of 12,

Play, P = 8, No, N = 4

[I.G = Entrops (S) - [weight avg.] × Entropy (each attribute)]

Now,

$$E(S) = -P(Yes)\log_2 P(Yes) - P(No)\log_2 P(No)$$

$$= -\frac{8}{12}\log_2\left(\frac{8}{12}\right) - \left(\frac{4}{12}\right)\log_2\left(\frac{4}{12}\right)$$

$$= 0.918$$

Now,

For finding the root node, we have to find the IG of each Attribute

- The attribute with highest IG will be the Root node.

Calculating Information Gain (IG) of Each Attribute

Outlook

Sunny	Overcast	Rainy
$\frac{4}{12}$ P = 1, N = 3	$\frac{4}{12}$ P = 4, N = 0	$\frac{4}{12}$ P = 3, N = 1

Now,

Entropy for each feature —

$$E(\text{outlook} = \text{sunny}) = -\frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{3}{4}\log_2\left(\frac{3}{4}\right) = 0.811$$

$$E(\text{outlook} = \text{overcast}) = -\frac{4}{4}\log_2\left(\frac{4}{4}\right) - \frac{0}{4}\log_2\left(\frac{0}{4}\right) = 0$$

$$E(\text{outlook} = \text{Rainy}) = -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) = 0.811$$

Now,

$$\text{Information from outlook} = \frac{4}{12} \times 0.811 + \frac{4}{12} \times 0 + \frac{4}{12} \times 0.811$$

$$= 0.540$$

So, $IG = E(S) - 0.540$

$$= 0.918 - 0.540 = 0.378$$

For windy —

windy

$\frac{8}{12}$ False — P = 6, N = 2

True — P = 2, N = 2 $\frac{4}{12}$

Now,

$$E(\text{windy} = \text{true}) = -\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)$$

$$= 1$$

$$E(\text{windy} = \text{false}) = -\frac{6}{8}\log_2\left(\frac{6}{8}\right) - \frac{2}{8}\log_2\left(\frac{2}{8}\right)$$

$$= 0.811$$

Now,

Information from windy — $\frac{4}{12} \times 1 + \frac{8}{12} \times 0.811$

$$= 0.874$$

So,

Information Gain $= E(S) - 0.874$

$$= 0.918 - 0.874$$

$$= 0.044$$

For Humidity —

Humidity

$\frac{6}{12}$ High

P : 3

N : 3

Normal $\frac{6}{12}$

P : 5

N : 1

Now,

$$E(\text{Humidity} = \text{high}) = -\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right)$$

$$= 1$$

$$E(\text{Humidity} = \text{Normal})$$

$$= -\frac{5}{6}\log_2\left(\frac{5}{6}\right) - \frac{1}{6}\log_2\left(\frac{1}{6}\right)$$

$$= 0.650$$

Now,

Information from Humidity —

$$= \frac{6}{12} \times 1 + \frac{6}{12} \times 0.650$$

$$= 0.825$$

Now

Information Gain $= E(S) - 0.825$

$= 0.918 - 0.825 = 0.093$

For Temperature —

Temperature

Hot	Mild	Cool
$\frac{4}{12}$	$\frac{4}{12}$	$\frac{4}{12}$
P = 2	P = 3	P = 3
N = 2	N = 1	N = 1

Now,

$$E(\text{Temperature} = \text{Hot}) = -\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)$$
$$= 1$$

$$E(\text{Temperature} = \text{Mild})$$
$$= -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) = 0.811$$

$$E(\text{Temperature} = \text{Cool})$$
$$= -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)$$
$$= 0.811$$

$$\text{Information from Temperature} = \frac{4}{12}\times 1 + \frac{4}{12}\times 0.811 + \frac{4}{12}\times 0.811$$
$$= 0.874$$

$$\therefore IG = 0.918 - 0.874$$
$$= 0.044$$

Comparing -

Outlook -

I.G = 0.378

windy -

I.G = 0.044

Humidity -

IG = 0.093

Temperature -

IG = 0.044

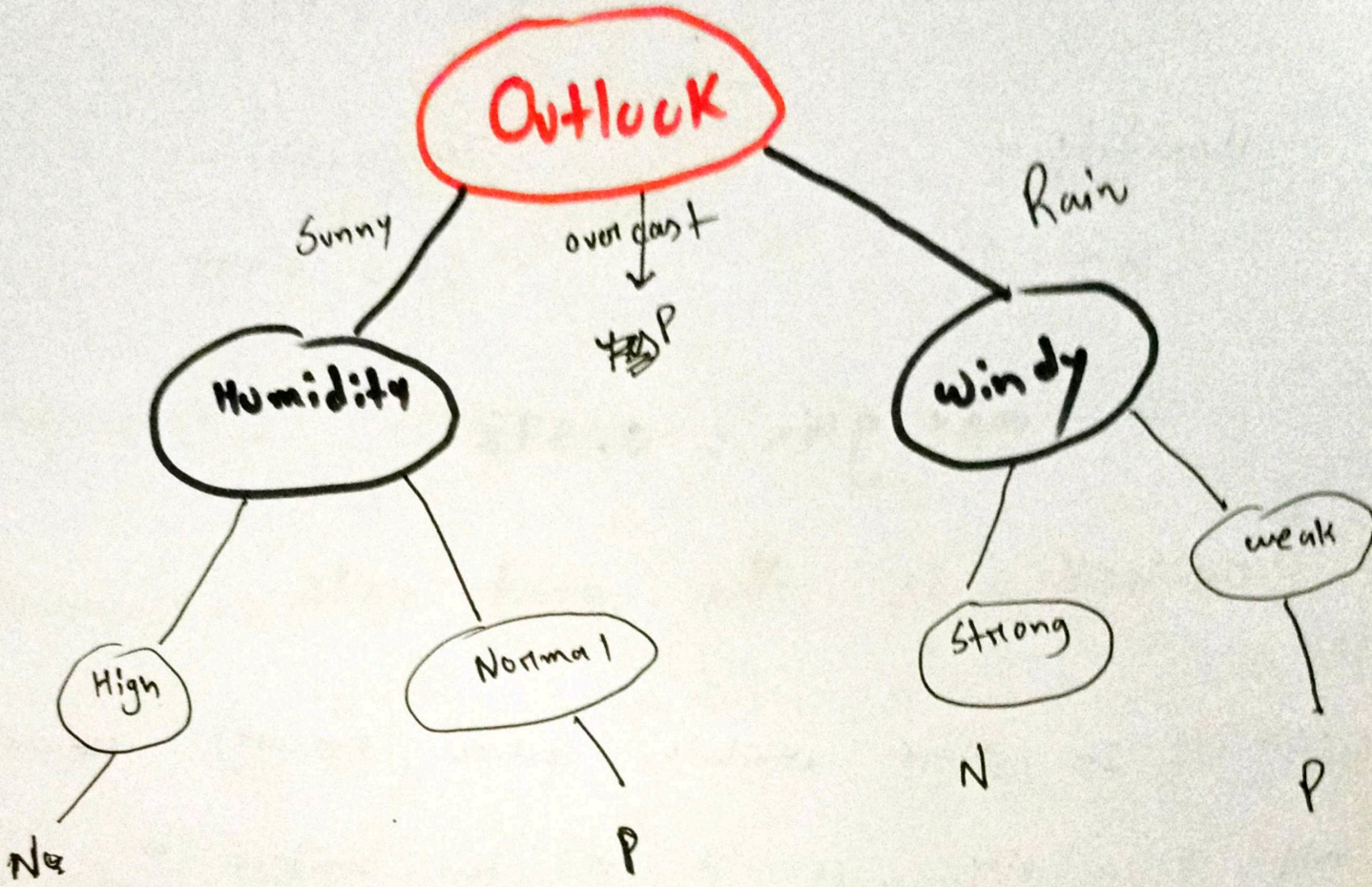
so, **max. gain : 0.378**

∴ Outlook is the root node.

· Since, In outlook attribute, feature (over cast) ~~can~~ containing only P i.e Play, so it can be considered as root node.

For sunny, and Rain, we have to recalculate the things and select the attribute which is having maximum information gain —

Fulling all the requirements, The required decesion Tree -



- when outlook ⇒ overcast we can always play.

- If the outlook is sunny, we will drill down to check humidity condition,

  if normal ⇒ P

  if high → N

- At the end, we will do prunning to get the optimal solution, which reduces the complexity.

i) c) Solu:-

FP tree -

Given, minimum support = 3

Now,

A	1 3 5 6
B	2 3 4 5 6
C	1 2 3 5 6
D	1 4 6
E	2 3 4 5

- First we will list out the individuals items and then find the frequency.
- Then list out the priorities.

P.T.O

itemset	Frequency	Priority
1	3	4th
2	3	5th
3	4	1st
4	3	6th
5	4	2nd
6	4	3rd

More frequency

→ give more priority

→ same frequency
↓
FCFS

In the order of priority -

3 5 6 1 2 4

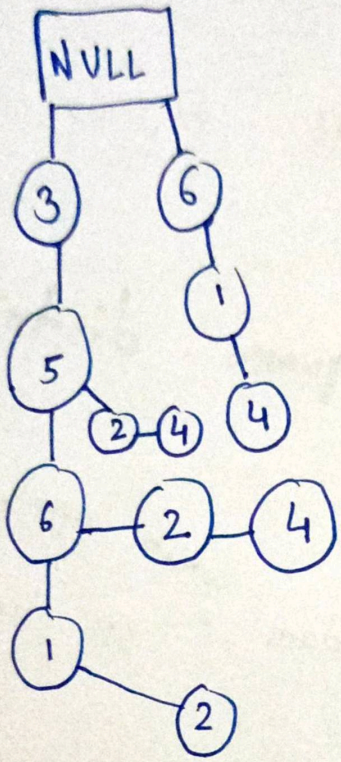
Now, order items according to priority —
↓
means reorder

Trans ID	Items	Ordered Items
A	1 3 5 6	3 5 6 1
B	2 3 4 5 6	3 5 6 2 4
C	1 2 3 5 6	3 5 6 1 2
D	1 4 6	6 1 4
E	2 3 4 5	3 5 2 4

Now, we will construct the tree -

Initially, we will take NULL.

- root node will be NULL



3 = 1, 2, 3, 4

5 = 1, 2, 3, 4

6 = 1, 2, 3

1 = 1, 2

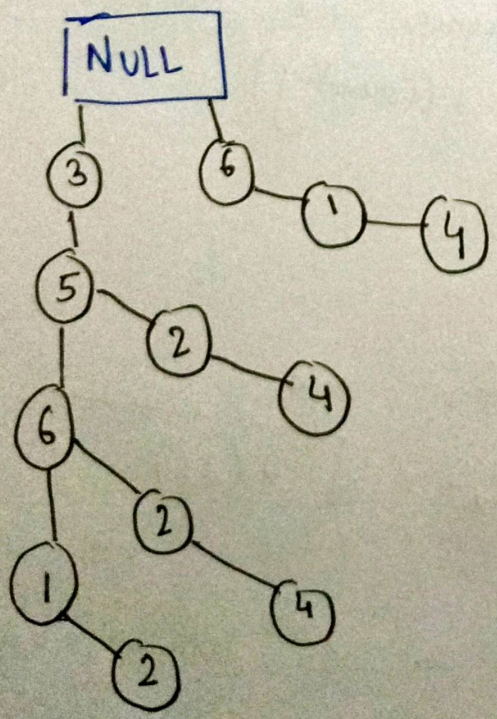
2 = 1, 2, 3

4 = 1, 2

6 = 1

1 = 1

4 = 1



1) a) The chi-square distribution is left skewed —

No,

The chi-square distribution curve is skewed to the right, and its shape depends on the degrees of freedom. df.

• The notion for the chi-square distribution is

$\chi \sim \chi^2_{df}$, where

df = degrees of freedom, which depends on how chi-square is being used.

(if you want to practice calculating chi-square probabilities then use $df = n-1$. The degrees of freedom for the major uses are each calculated differently).

For the χ^2 distribution,

the population mean is $\mu = df$ and

population standard deviation is, $\sigma_{\chi^2} = \sqrt{2(df)}$.

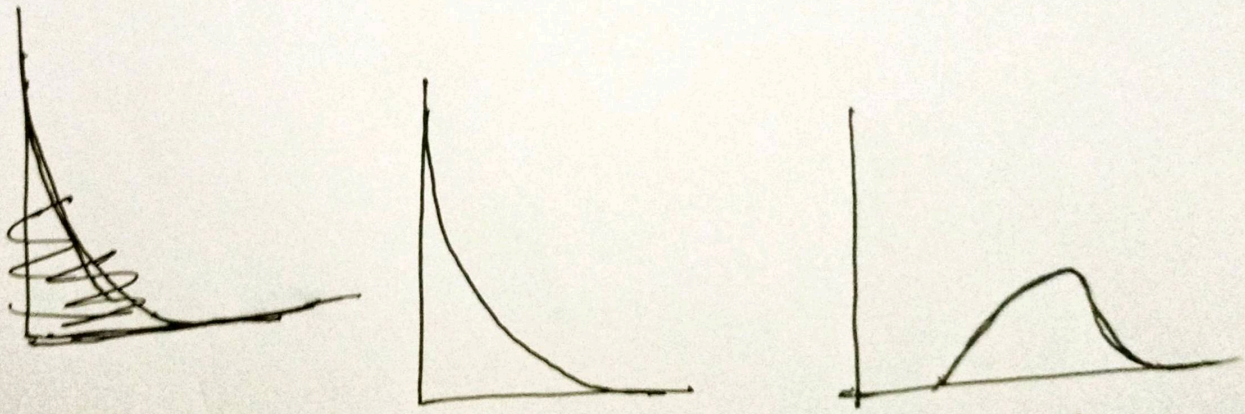
The random variable is shown as χ^2, but may be any upper case letter.

The random variable for a chi-square distribution with K degrees of freedom is the sum of K independent, squared standard normal variables.

$$X^2 = (Z_1)^2 + (Z_2)^2 + \cdots + (Z_K)^2$$

1. The curve is nonsymmetrical and skewed to the right.

2. There is diff. chi-square curve for each df.



df = 2 df = 24

3. The test static for any test is always greater than or equal to zero.

4. when df > 90, the chi-square curve approximates the normal distribution.

1) a) iii) solu:-

The principal components are always mutually exclusive and exhaustive of the variables.

Ans:-

As per MECE method,

Principal components are mutually exclusive and collectively ~~exato~~ exhaustive.

→ no one individual can appear in more than one category and,

→ all categories combined include all individuals

∴ Given statements is True.

Hence,

every set in maximal is also closed but every closed set is not maximal.

Hence,

maximal set is subset of closed set.

$M \subseteq C$.

3)b) ii) Solu:-

Boosting is an ensemble modeling techniques, that attempts to an that can be used to improve the accuracy of a classifier.

- It attempts to build a strong classifier from the number of weak classifiers.
- It is done by building a model by using weak models in series.
- Firstly, a model is built from the training data. Then, the second model is built which tries to

correct the errors present in the first model.

- This procedure is continued and models are added until either the complete training data set is predicted correctly or the maximum number of models is added.

Boosting Algorithms —

An example of how boosting can improve the accuracy of a classifier is through the use of **the AdaBoost** (Adaptive Boosting) algorithm.

- First, really successful boosting algorithm developed for the purpose of binary classification.

→ It is a very popular boosting technique that combines multiple "week classifiers" into a single "strong classifier".

Algorithm

1) Initialise the dataset and assign equal weight to each of the data point.

2) Provide this as input to the model and identify the wrongly classified data points.

3. Increase the weight of the wrongly classified data points and decrease the weights of correctly classified data points. And then normalize the weights of all data points.

4. If (got required results)

Goto step 5

else

Goto step 2

5. End

3)b) i) solu:

Can't be determined.

The question says that it achieves an accuracy of 95% on the test set. It doesn't tell anything about the test set. It may be possible that the test data is also skewed, but it's also possible that the data isn't skewed

Hence, we can't say anything