DATE : ___/___/____
PAGE : _____

DATA MINING
ASSIGNMENT

NAME - NIKITA SINGH

ROLL NO. - 20570040

YEAR - 3RD

SEMESTER - 6TH

DATE: __/__/____
PAGE: _____

2018 QP No. - 9426 A

1 (a) What is the difference b/w Data mining and KDD?

Data Mining	KDD
• Data mining is the process of automatically discovering useful info in large data repositories. Data mining techniques are deployed to scour large db in order to find novel and useful patterns that might otherwise remain unknown. They also provide capabilities to predict the outcome of a future observation	• Knowledge discovery in database. It is the overall process of converting raw data into useful info. This process consists of a series of transformation steps, from data preprocessing to postprocessing of data mining results.

(b) Identify attribute types for the following:

i) eye color - Nominal
ii) grades - Ordinal
iii) dates in a calendar - Interval
iv) age - Ratio

(c) What are the maximum and minimum values of Gini index? Find Gini index for the following node:

Node N	Count
class = 0	1
class = 1	5

Maximum $= 1 - \frac{1}{c}$, when records are equally distributed amg all classes, least beneficial, c is total no. of classes

Minimum = 0, when all records belong to one class, most beneficial

Given :-

for class = 0, count = 1

for class = 1, count = 5

$$\text{Gini Index} = 1 - \sum_{i=0}^{c-1} p_i(t)^2$$

where $p_i(t)$ = frequency of class i at node t

c = total no. of classes

for 2 classes,

$$\text{Gini Index} = 1 - p^2 - (1-p)^2$$

$$P(0) = \frac{1}{6}, \quad P(1) = \frac{5}{6}$$

$$\text{Gini Index} = 1 - \left(\frac{1}{6}\right)^2 - \left(\frac{5}{6}\right)^2 = 1 - \frac{1}{36} - \frac{25}{36}$$

$$= \frac{36-26}{36}$$

$$= \frac{10}{36} = 0.2777$$

Good Write

(d) Give two applications where graph data structure is used to model data.

Applications where graph data structure is used to model data :-

1. Data with relationship among objects : The relationships amg objects frequently convey imp. info. In such cases, the data is often represented as a graph. In particular, the data objects are mapped to nodes of the graph, while the relationships among objects are captured by the links blw objects and link properties, such as direction and weight. Eg:- Web pages on world wide web, which contain both text and links to other pages.

2. Data with Objects that are graph : If objects have structure, that is, the objects contain subobjects that have relationships, then such objects are frequently represented by a graph. Eg:- the structure of chemical compounds can be represented by a graph, where the nodes are atoms and the links blw nodes are chemical bonds.

(e) Given four pnts $p_1(0,2)$, $p_2(2,0)$, $p_3(3,1)$ and $p_4(5,1)$. Calculate Euclidian distance blw the pnts p_1 and p_2 and p_3 and p_4

Euclidian Distance Matrix

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Good Write

Euclidian Distance $= \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$

where k is dimension ie. x, y, z
x_k is pnt 1's dimension
y_k is pnt 2's dimension

p1 to p2 $= \sqrt{(0-2)^2 + (2-0)^2} = \sqrt{8} = 2.828$

p1 to p3 $= \sqrt{(0-3)^2 + (2-1)^2} = \sqrt{10} = 3.162$

p1 to p4 $= \sqrt{(0-5)^2 + (2-1)^2} = \sqrt{26} = 5.099$

p2 to p3 $= \sqrt{(2-3)^2 + (0-1)^2} = \sqrt{2} = 1.414$

p2 to p4 $= \sqrt{(2-5)^2 + (0-1)^2} = \sqrt{10} = 3.162$

p3 to p4 $= \sqrt{(3-5)^2 + (1-1)^2} = 2$

(f) Let X denote the categorical attribute having possible values {poor, good, better, best}. What is the representation of each value when X is converted to binary form?

Using Symmetric Binary Conversion

Step 1: Assign all categories value some integer values from 0 to n

poor	good	better	best
0	1	2	3

Step 2: Find no. of binary attributes required

$n = \lceil \log_2(m) \rceil$

where m is the total no. of integer values.

Good Write

$n = \lceil \log_2(4) \rceil$

$n = 2$

Step 3: Assign binary values

Category	Integer values	Binary attributes x_1	Binary attributes x_2
poor	0	0	0
good	1	0	1
better	2	1	0
best	3	1	1

(g) How are interval scaled attributes different from ratio scaled attributes? Give an example of each.

- for interval scaled attributes, the difference b/w values are meaningful ie a unit of measurement exists whereas for ratio scaled attributes, both difference and ratio are meaningful.
- Zero-pnt in an interval scale is arbitrary whereas the ratio scale has an absolute zero pnt.
- Variables measured in interval scale can be added and substracted whereas variables measured in ratio scale can be multiplied and divided.
- In interval scale, the arithmetic mean is calculated ~~other~~ whereas in ratio scale, the geometric / harmonic mean is calculated.
- Eg:- Interval – Celsius
  Ratio – Kelvin

(h) How is a eager learner different from lazy learner? Support your answer with an example from both categories of classifiers.

Good Write

Eager Learner	Lazy Learner
• Eager learner is fast as it has a pre calculated algo. attribute to the class label as soon as the training data becomes available.	• Lazy learner is slow as it calculates based on current best dataset data until it is needed to classify the test examples.
• Takes long time learning and less time classifying data.	• Takes less time learning and more time classifying data.
• Eg:- Decision tree, rule based classifiers :- They have two steps. One, an inductive step for constructing a classification model from data. Two, a deductive step for applying the data to test examples.	• Eg:- KNN. KNN is called 'lazy' because it does no training at all when you supply the training data. At training time, all it is doing is storing the complete data set but it does not do any calculations at this pt. Neither does it try to derive a more compact model from the data which it could use for scoring.

(i) State the Apriori principle. Comment on the following statement:

"If an item set $\{x,y,z\}$ is frequent, then its subset $\{y,z\}$ will be frequent."

Apriori Principle - If an itemset is frequent, then all of its subsets must also be frequent.

Good Write

"If an itemset $\{x,y,z\}$ is frequent, then its subset $\{y,z\}$ will be frequent."

According to the Apriori Principle, the above statement is true. Since set $\{x,y,z\}$ is frequent, clearly the any transaction containing $\{x,y,z\}$ must also contain its subset $\{x\}$, $\{y\}$, $\{z\}$, $\{x,y\}$, $\{x,z\}$ and $\{y,z\}$. As a result $\{y,z\}$ is also frequent.

(j) Given the age of four students, normalize the values $\{18, 21, 22, 25\}$.

Normalization is used to scale the data of an attribute so that it falls in a small range, such as -1.0 to 1.0 or 0.0 to 1.0.

1. Min-Max Normalization

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

2. Zero Score Normalization / Standardization

$$x' = \frac{x - \bar{x}}{\sigma}, \quad \bar{x} - \text{mean}, \ \sigma - \text{standard deviation}$$

$$\sigma = \sqrt{\frac{\Sigma(x_i - \mu)^2}{N}}$$

x	$\bar{x}$	$x-\bar{x}$	$(x-\bar{x})^2$	σ	x'
18	21.5	-3.5	12.25	6.25	-0.56
21		-0.5	0.25		-0.08
22		0.5	0.25		0.08
25		3.5	12.25		0.56
			$\Sigma = 25$		

Good Write

Normalize x : -0.56, -0.08, 0.08, 0.56

(k) Explain the following terms with reference to the DBSCAN algorithm:

i) Core point : ii) Noise point

i) Core point - These pnt are in the interior of a density based cluster. A pnt is a core pnt if the no. of pnts within a given neighborhood around the pnt as determined by the distance func and a user specified distance parameter, Eps, exceeds a certain threshold, MinPnts which is also user specified parameter.

ii) Noise point - A noise pnt is any pnt that is neither a core pnt nor a border pnt (Border point ie. falls within the neighborhood of a core pnt).

(l) What are mutually exclusive rules in a rule based classifier? What problem may arise if rules are not mutually exclusive? How can such problem be resolved?

Mutually Exclusive Rules - The rules in a rule set R are mutually exclusive if no two rules in R are triggered by the same record. This property ensures that every record is covered by at most one rule in R.

Problem that might arise if rules are not mutually exclusive :- If a rule set is not mutually exclusive, then a record can be covered by several rules, some of which may predict conflicting classes.

Solution :-

1. Ordered Rules - In this approach, the rules in a rule set are ordered in decreasing order of their priority, which can be defined in many ways. An ordered rule set is also known as a decision list. When a test record is presented, it is classified by the highest-ranked rule that covers the record. This avoids the problem of having conflicting classes predicted by multiple classification rules.

2. Unordered Rules - This approach allows a test record to trigger multiple classification rules and considers the consequent of each rule as a vote for a particular class. The votes are then tallied to determine the class label of the test record. The record is usually assigned to the class that receives the highest no. of votes.

2(a) Consider the following transaction dataset:-

Customer Id	Transaction Id	Items Bought
1	0001	{a,d,e}
1	0024	{a,b,c,e}
2	0012	{a,b,d,e}
2	0031	{a,c,de}
3	0015	{b,c,e}
3	0022	{b,d,e}
4	0029	{c,d}
4	0040	{a,b,c}
5	0033	{a,d,e}
5	0038	{a,b,e}

Good Write

Good Write

(i) Compute the support of itemsets {e}, {b, d}, {b, d, e}, {a, b, d, e}

Support - fraction of transactions that contain an itemset. (s) $s = \frac{\sigma}{|T|}$

$s(\{e\}) = \frac{8}{10}$

$s(\{b,d\}) = \frac{2}{10}$

$s(\{b,d,e\}) = \frac{2}{10}$

$s(\{a,b,d,e\}) = \frac{1}{10}$

(ii) Compute the confidence of rules {b, d} → {e} and {e} → {b, d}.

Confidence (c) measures how often in a rule $X \rightarrow Y$ items in Y appear in transaction that contain X.

$$c = \frac{\sigma(X \cup Y)}{\sigma X}$$

{b, d} → {e}

$$c = \frac{\sigma(\{b,d,e\})}{\sigma(\{b,d\})} = \frac{2}{2} = 1$$

{e} → {b, d}

$$c = \frac{\sigma(\{e,b,d\})}{\sigma(\{e\})} = \frac{2}{8} = \frac{1}{4} = 0.25$$

(iii) Is confidence a symmetric measure?

No, confidence is not a symmetric measure since the confidence for $X \rightarrow Y$ and $Y \rightarrow X$ may not be the same.

(b) Let A and B be two sets of integers. A distance measure 'd' is defined as $d(A-B) = size(A-B)$, where '-' denotes the set difference. Prove that 'd' is not a metric.

Measures that satisfy all three following property are called metric:

1) Positivity

$d(x,y) \geq 0$

$d(x,x) \geq 0$ } for all x and y

$d(x,y) = 0$ } for $x = y$

2) Symmetry

$d(x,y) = d(y,x)$ $\forall$ x and y

3) Triangle Inequality

$d(x,z) \leq d(x,y) + d(y,z)$ $\forall x,y,z$

Now, let $A = \{1,2,3,4\}$

$B = \{1,2,3\}$

size $(A-B) = \{4\}$

size $(B-A) = \phi$

$\Rightarrow$ size $(A-B) \neq$ size $(B-A)$
$\Rightarrow$ $d(A-A) \neq d(B,A)$
$\therefore$ H is asymmetric

$\therefore$ d is not a metric.

3.(a) Explain the concept of aggregation with the help of an example. List three use of aggregation.

Sometimes "less is more" and this is the case with aggregation, the combining of two or more objects into a single object. Consider a data set consisting of transactions (data objects) recording the daily sales of products in various store locations for different days over the course of a year. One way to aggregate transactions for this data set is to replace all the transactions of a single store with a single storewide transaction. This reduces the hundreds of or thousands of transactions that occur daily at a specific store to a single daily transaction, and the no. of data objects is reduced to the no. of stores.

Uses of aggregation :-

1. Data aggregation is used in many fields where a large no. of datasets are involved. It helps in making fruitful decisions in marketing or finance management.

2. Efficient use of data aggregation can help in the creation of marketing schemes. Eg. If the company is performing ad campaigns on a particular platform, they must deeply analyze the data to raise sales. The aggregation can help in analysing the execution over a respective time period of ~~com~~ campaigns or a particular cohort or a particular platform.

3. For the business analysis purpose, the data can be aggregated into summary formats which can help the head of the firm to take correct decisions for satisfying the customers. It helps in inspecting groups of peoples.

(b) What is the difference b/w noise and outliers? Answer the following questions:
i) Is noise ever interesting or desirable?
ii) Are outliers ever interesting or desirable?
iii) Are noise objects always outliers?
iv) Are outliers always noise objects?

Noise	Outliers
• Noise is the random component of a measurement error.	• They are data objects which have characteristics different/ unsual than other objects in the data set.
• It is used in connection with data that has a spatial or temporal component.	• They can be legimate data objects or values.
• Noise is an undesirable signal.	• They may sometimes be of interest.

i) Noise denotes mistakes / errors / wrong items. It is never interesting or desirable.

ii) Outliers are anomalies. They are sometimes interesting or desirable.

iii) No, noise object are not always outliers as random distortions can result in an object or value much like a normal one.

iv) No, outliers are not always noise as they merely represent a class of objects that are different from normal objects.

4. (a) for the following two-class problem, draw a Confusion Matrix and compute the Accuracy and Error from it.

Instance id	A	B	Predicted Class	Actual Class
1	T	F	+	+
2	T	T	+	+
3	T	T	+	−
4	T	F	−	−
5	T	T	+	+
6	F	F	−	+
7	F	F	−	−
8	F	F	−	−
9	T	T	−	+
10	T	F	−	+

Confusion Matrix

TP = 3
TN = 3
FP = 1
FN = 3

		PREDICTED CLASS	
		Class=+	Class=−
ACTUAL CLASS	Class=+	3	3
	Class=−	1	3

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{3+3}{10} = \frac{6}{10} = 0.6$$

$$\text{Error} = \frac{FP+FN}{TP+TN+FP+FN} = \frac{1+3}{10} = \frac{4}{10} = 0.4$$

(b) What is k-fold cross validation? How is it different from the hold-out method?

An alternative to random subsampling is cross-validation. In this approach, each ~~one~~ record is used the same number of times for training and exactly once for testing. The k-fold cross validation method generalizes this approach by segmenting the data into k equal-sized partitions. During each run, one of the partitions is chosen for testing, while the rest of them are used for training. This procedure is repeated k times so that each partition is used for testing exactly once. The total error is found by summing up the errors for all k runs.

How is K-fold cross validation different from Hold Out method :-

- Cross-validation is usually the preferred method b/c it gives your model the opportunity to train on multiple train-test splits. This gives you a better indication of how well your model will perform on unseen data. Hold-out, on the other hand, is dependent on just one train-test split. That makes the hold-out method score dependent on the how the data is split into train and test sets.

- The hold-out method is good to use when you have a very large dataset, you're on a time crunch, or you are starting to build an intial model is your data science project. Keep in mind that b/c cross-validation uses multiple train-test splits, it takes more computational power and time to run than using the holdout method.

5.(a) What is the difference b/w hierarchical and partition based clustering? Enumerate two advantages and disadvantages of hierarchical clustering.

Hierarchical Clustering	Partition-Based Clustering
• Set of clusters is nested.	• set of clusters is unnested
• It is a set of nested clusters that are organized as a tree. Each node (cluster) in the tree (except for leaf node), is the union of its children (subclusters) of the root of the tree is the cluster containing all the objects.	• It is a division of the set of data objects into non-overlapping subsets (clusters), such that each data object is in exactly one subset.

Hierarchical clustering	Partition Based clustering
• It can be viewed as a sequence of partitional clustering	• It can be obtained by cutting the tree hierarchical tree at a particular level
• Eg :- Dendogram	• Eg :- K-means

Advantages of Hierarchical Clustering

- We can obtain the optimal number of clusters from the model itself, human intervention not required.
- Dendrograms help us in clear visualization, which is practical and easy to understand.

Disadvantages of Hierarchical clustering

- Not suitable for large datasets due to high time and space complexity.
- There is no mathematical objective for hierarchical clustering.
- All the approaches to calculate the similarity b/w clusters has their own disadvantages.

(b) What is simple random sampling?

The simplest type of sampling is simple random sampling. For this type of sampling, there is an equal probability of selecting any particular item. There are two variations on random sampling:

1. Sampling without replacement — as each item is selected, it is removed from the set of all objects that together constitute the population, and

2. sampling with replacement - objects are not removed from the population as they are selected for the sample.

When the population consist of different types of objects, with widely different numbers of objects, simple random sampling can fail to adequately represent those types of objects that are less frequent. This can cause problems when the analysis requires proper representation of all objects types.

(c) Consider the following rule set:

R1: (Give Birth = no) ∧ (Can fly = yes) → Birds
R2: (Give Birth = no) ∧ (Live in water = yes) → Fishes
R3: (Give Birth = yes) ∧ (Blood Type = warm) → Mammals
R4: (Give Birth = no) ∧ (Can fly = no) → Reptiles
R5: (Live in Water = sometimes) → Amphibians

Which rules cover the following tuples:

Name	Blood Type	Give Birth	Can fly	Live in water	class
hawk	Warm	No	Yes	No	?
bear	warm	Yes	No	No	?

Hawk is covered by R1:
(Give Birth = no) ∧ (can fly = yes) → Birds

⇒ Hawk has class = Birds

Bear is covered by R3:
(Give Birth = yes) ∧ (Blood Type = warm) → Mammals

⇒ Bear has class = Mammals.

6.(a) List the rules that can be generated from the 3-itemset ABE using the following transactional data set. Compute the confidence.

T_id	Itemset
t1	ACD
t2	BCE
t3	ABCE
t4	BDE
t5	ABCE
t6	ABCD

for 3 itemset {ABE}
No. of rules that can be generated = $2^3 - 2 = 8 - 2 = 6$

R1 = {A, B} → {E}
R2 = {A, E} → {B}
R3 = {B, E} → {A}
R4 = {A} → {B, E}
R5 = {B} → {A, E}
R6 = {E} → {A, B}

Confidence = $c(X \to Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$

$c(R_1) = \frac{\sigma(\{A,B,\varepsilon\})}{\sigma(\{e\})} = \frac{2}{4} = 0.5$

$c(R_2) = \frac{\sigma(\{A,\varepsilon,B\})}{\sigma\{\}}$

$c(R_1) = \frac{\sigma(\{A,B,\varepsilon\})}{\sigma(\{A,B\})} = \frac{2}{3} = 0.66$

$c(R_2) = \frac{\sigma(\{A,B,\varepsilon\})}{\sigma(\{A,\varepsilon\})} = \frac{2}{2} = 1$

$c(R_3) = \frac{\sigma(\{A,B,\varepsilon\})}{\sigma(\{B,\varepsilon\})} = \frac{2}{4} = 0.5$

$c(R_4) = \frac{\sigma(\{A,B,\varepsilon\})}{\sigma(\{A\})} = \frac{2}{4} = 0.5$

$c(R_5) = \frac{\sigma(\{A,B,\varepsilon\})}{\sigma(\{B\})} = \frac{2}{5} = 0.4$

$c(R_6) = \frac{\sigma(\{A,B,\varepsilon\})}{\sigma(\{e\})} = \frac{2}{4} = 0.5$

(b) Enumerate strong association rules if min.conf = 0.5.

min.conf = 0.5

∴ all rules with confidence ≥ 0.5 are strong association rules

∴ Strong association Rules are:-

R_1 : $\{A,B\} \rightarrow \{\varepsilon\}$

R_2 : $\{A,\varepsilon\} \rightarrow \{B\}$

R_3 : $\{B,\varepsilon\} \rightarrow \{A\}$

R_4 : $\{A\} \rightarrow \{B,\varepsilon\}$

R_6 : $\{\varepsilon\} \rightarrow \{A,B\}$

7. (a) Given the following points: 2, 4, 10, 12, 3, 20, 30, 11, 25. Given k=3, and the initial means, $\mu_1 = 2$, $\mu_2 = 4$ and $\mu_3 = 6$. Show the clusters obtained and the new mean after each iteration using the k-means algorithm.

k = 3

$\mu_1 = 2$, $\mu_2 = 4$, $\mu_3 = 6$

Iteration 1:

points x	dis. from $\mu_1 = 2$ $\sqrt{(x-2)^2}$	dis. from $\mu_2 = 4$ $\sqrt{(x-4)^2}$	dist. from $\mu_3 = 6$ $\sqrt{(x-6)^2}$
2	0 cluster 1	2	4
4	2	0 cluster 2	2
10	8	6	4 cluster 3
12	10	8	6 cluster 3
3	1 cluster 1	1	3
20	18	16	14 cluster 3
30	28	26	24 cluster 3
11	9	7	5 cluster 3
25	23	21	19 cluster 3

Clusters obtained :-

cluster 1 = 2, 3

new mean $\mu_1 = \frac{2+3}{2} = 2.5$

cluster 2 = 4

new mean, $\mu_2 = \frac{4}{1} = 4$

cluster 3 = 10, 12, 20, 30, 11, 25

new mean, $\mu_3 = \frac{10+12+20+30+11+25}{6} = \frac{108}{6} = 18$

Iteration 2:

points x	dis. from $\mu_1 = 2.5$ $\sqrt{(x-2.5)^2}$	dis. from $\mu_2 = 4$ $\sqrt{(x-4)^2}$	dist from $\mu_3 = 18$ $\sqrt{(x-18)^2}$
2	0.5 cluster 1	2	16
4	1.5	0 cluster 2	14
10	7.5	6 cluster 2	8
12	9.5	8	6 cluster 3
3	0.5 cluster 1	1	15
20	17.5	16	2 cluster 3
30	27.5	26	12 cluster 2
11	8.5	7	7 cluster 2
25	22.5	21	7 cluster 3

clusters obtained:

cluster 1, = 2, 3

new mean, $\mu_1 = \frac{2+3}{2} = 2.5$

cluster 2, = 4, 10

new mean, $\mu_2 = \frac{4+10}{2} = 7$

cluster 3 = 12, 20, 30, 11, 25

new mean, $\mu_3 = \frac{12+20+30+11+25}{5} = 19.6$

Iteration 3:

points x	dis from $\mu_1 = 2.5$ $\sqrt{(x-2.5)^2}$	dis. from $\mu_2 = 7$ $\sqrt{(x-7)^2}$	dis from $\mu_3 = 19.6$ $\sqrt{(x-19.6)^2}$
2	0.5 cluster 1	5	17.6
4	1.5 cluster 1	3	15.6
10	7.5	3 cluster 2	9.6
12	9.5	5 cluster 2	7.6
3	0.5 cluster 1	4	16.6
20	17.5	13	0.4 cluster 3
30	27.5	23	10.4 cluster 3
11	8.5	4 cluster 2	8.6
25	22.5	18	5.4 cluster 3

Clusters obtained:

cluster 1, = 2, 4, 3

new mean, $\mu_1 = \frac{2+4+3}{3} = 3$

cluster 2 = 10, 12, 11

new mean, $\mu_2 = \frac{10+12+11}{3} = 11$

cluster 3 = 20, 30, 25

new mean, $\mu_3 = \frac{20+30+25}{3} = 25$

Iteration 4:

Points x	dis from $\mu_1 = 3$ $\sqrt{(x-3)^2}$	dis from $\mu_2 = 11$ $\sqrt{(x-11)^2}$	dis from $\mu_3 = 25$ $\sqrt{(x-25)^2}$
2	1 cluster 1	9	23
4	1 cluster 1	7	21
10	7	1 cluster 2	15
12	9	1 cluster 2	13
3	0 cluster 1	8	22
20	17	9	5 cluster 3
30	27	19	5 cluster 3
11	8	0 cluster 2	14
25	22	14	0 cluster 3

Clusters obtained:

cluster 1 = 2, 4, 3

new mean, $\mu_1 = \frac{2+4+3}{3} = 3$

cluster 2 = 10, 12, 11

new mean, $\mu_2 = \frac{10+12+11}{3} = 11$

cluster 3 = 20, 30, 25

new mean, $\mu_3 = \frac{20+30+25}{3} = 25$

New set of means ie. μ_1, μ_2 and μ_3 are same as the before ie. means obtained in Iteration 3 therefore algorithm ends here succesfully.

Good Write

(b) What is the difference b/w Partial and Complete clustering scheme?

Partial clustering	Complete clustering
• Does not assign a cluster to every object	• Assign every object to a cluster
• Only desired in case where some object in a data set may not define to well defined groups	• Desirable in most cases.
• Eg:- To find important topics in a newspaper's particular month's stories, one may want to search only clusters of documents that are tightly related to a common theme.	• Eg:- An application that uses clustering to organize documents for browsing needs to guarantee that all documents can be browsed.

Good Write

2019 QP No. - 2780

1.(a) find the Euclidean distance b/w data points X(0, -1, 0, 1) and Y(1, 0, -1, 0).

$$\text{Euclidean dis} = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

$$= \sqrt{(0-1)^2 + (-1-0)^2 + (0-(-1))^2 + (1-0)^2}$$

$$= \sqrt{1+1+1+1}$$

$$= \sqrt{4}$$

$$= 2$$

Euclidean distance b/w X and Y = 2 units

(b) If recall and precision are 0.5 and 0.6 respectively, compute the value of F1 measure.

$$\text{Precision} = \frac{TP}{TP+FP} = 0.6$$

$$\text{Recall} = \frac{TP}{TP+FN} = 0.5$$

$$F1 = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} = \frac{2 \times 0.5 \times 0.6}{0.5 + 0.6}$$

$$= \frac{0.6}{1.1} = 0.5454 \Rightarrow$$

Good Write

(c) In a given dataset, it is found that an itemset {a, b}, is infrequent. will itemset {abc} be infrequent or frequent? Explain why.

If an itemset {a b} is infrequent, then all of its supersets must be infrequent too as a converse of Apriori Principle.
In addition, according to anti-monotone property of support, if at least one of the subset of an itemset is infrequent, then it is guarantee to be infrequent.
so using these two statements,

{a, b} superset {a, b, c}
infrequent ⇒ infrequent

{a, b, c} subset {a, b}
infrequent ⇒ infrequent

{a, b, c} is infrequent.

(d) What are the three strategies for handling missing values in a dataset?

i) Eliminate data object or attributes
If a data set has only a few objects that have missing values, then it may be expedient to omit them.

ii) Estimate missing values
Sometimes missing data can be reliable estimated. If the attribute is continuous, then the average attribute value of the nearest neighbor is used. If categorical, then the most commonly occuring attribute value can be taken.

Good Write

iii) Ignore missing values during analysis
for instance, suppose that objects are being clustered and the similarity b/w pairs of data objects need to be calculated. If one or both objects of a pair have missing values for some attributes, then the similarity can be calculated by using only the attributes that do not have missing values.
∴ A data mining approach is modified here.

(e) Differentiate b/w precision and bias on the basis of the quality of the measurement process.

Precision	Bias
1. The closeness of repeated measurement (of the same quality) to one another is called precision.	1. A systematic variation of measurements from the quantity being measured is called Bias.
2. It is measured by the standard duration of a set of values.	2. It is measured by taking the difference b/w the mean of set of values and the known values of the quantity being measured.
3. Precision can be determined for other objects too.	3. Bias can only be determined for objects whose measured quantity knows by means external to the current situation.

for a set: 1.015, ~~1.05~~ 0.990, 1.013, 1.001, 0.986

Precision = 0.013 Bias = 0.001

(f) What is meant by variable transformation? what are its advantages?

A variable transformation refers to a transformation that is applied to all the values of a variable. In other words, for each object, the transformation is applied to the value of the variable for that object. for eg., if only the magnitude of a variable is important, then the values of the variable can be transformed by taking the absolute value.

Advantages:-
i) ~~the~~ It helps in making the relationship b/w variables more linear.
ii) The data after variable transformation is much easier to model as it will be less skewed and outlier will be less extreme.

(g) If support of an association rule $X \to Y$ is 80% & confidence is 75% can we derive support and confidence of the rule $Y \to X$? If yes, list down the values. If no, state the reasons.

If $s(X \to Y) = 80\%$,
then $s(Y \to X) = 80\%$
as support is a symmetric measure

$c(X \to Y) = 75\%$

$\neq c(Y \to X)$

as confidence is an asymmetric measure and as no other info is provided for X and Y, we cannot derive confidence.

(h) List down two advantages and two disadvantages of leave-one-out approach used in cross-validation for evaluating the performance of the classifier?

Advantages

i) It utilizes as much data as possible for training.

ii) Test sets are mutually exclusive and they effectively cover the entire data set.

Disadvantages

i) It is computationally expensive to repeat the procedure N times.

ii) Since each test set contains only one record the variance of the estimated performance metric tends to be high.

(i) Differentiate b/w agglomerative and divisive methods of hierarchical clustering with the help of a diagram.

Agglomerative hierarchical clustering method:

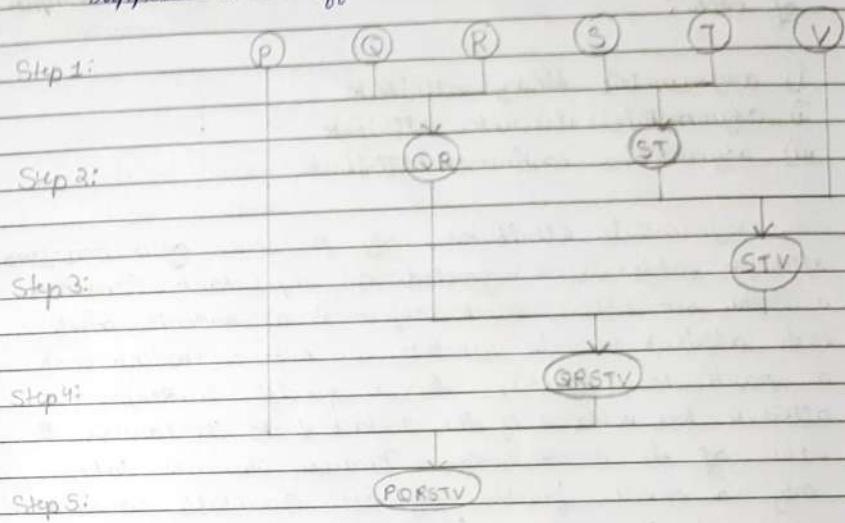
→ Start with points as individual clusters

→ At each step, merge the ~~closet~~ closest pair of clusters

→ It requires defining a notion of cluster proximity.

Good Write

Suppose six different data point PQRSTV:

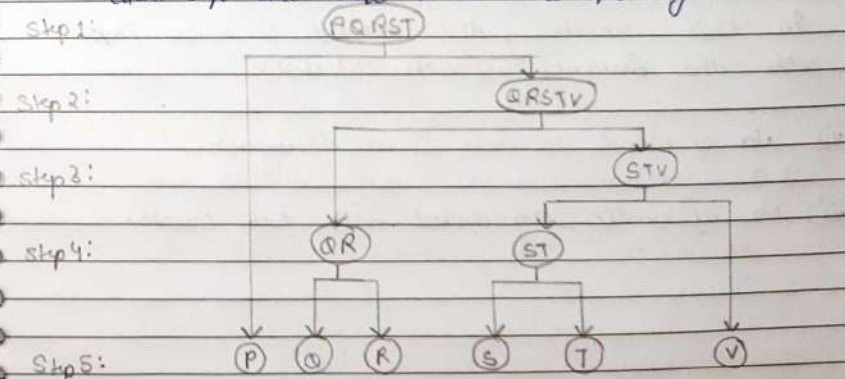


Division Method:

→ Start with one all inclusive cluster

→ At each step, split a cluster until only singleton clusters of individual points remain.

→ Here, we need to decide which cluster to split at each step and how to do the splitting.



Good Write

(j) What are asymmetric attributes? Give an example of each:

i) asymmetric binary attribute
ii) asymmetric discrete attribute
iii) asymmetric continuous attribute

for asymmetric attributes, only presence of a non-zero attribute values is regarded as important. Consider a data set where each object is a student and each attribute records whether or not a student took a particular course. for a specific student, an attribute has value 1 if the student took the course & value of is 0 otherwise. Because students take only a small fraction of all available courses, most of the values will be zero ∴ it is more meaningful and more efficient to focus on nonzero values.

i)

Attribute	Value
HIV	Yes / No

In this situation, HIV detected is more imp. than the instance of not detected.

ii) No. of words present in a document

iii) No. of credits associated with each course.

(k) The confusion matrix for a 2-class problem is given below:

		Predicted Class	
		class = 1	class = 0
Actual class	class = 1	400	100
	class = 0	200	300

Calculate the Accuracy, Sensitivity, Specificity, True Positivity Rate, and false Positive rate.

Given:

TP = 400 TN = 300
FP = 200 FN = 100

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{400+300}{400+300+200+100}$$

$$= \frac{700}{1000} = 0.7$$

$$\text{Sensitivity} = \frac{TP}{TP+FN} = \frac{400}{400+100} = \frac{400}{500} = 0.8$$

$$\text{Specificity} = \frac{TN}{TN+FP} = \frac{300}{300+200} = \frac{300}{500} = 0.6$$

True Positive Rate = Recall = Sensitivity = 0.8

False Positive rate = $\alpha = \frac{FP}{TN+FP}$ = 1 − specificity = 1 − 0.6

= 0.4

2.(a) What are the difference b/w noise and outliers? Are noise objects always outliers? Are outliers always noise objects?

~~Difference b/w noise and outliers : 2018 3(b)~~

Done in 2018 3(b)

(b) Let A and B be two sets of integers. A distance measure 'd' is defined as follows:
$d(A-B)$ = size $(A-B)$ + size $(B-A)$ where '−' denotes set difference. Size denotes the number of elements in the set.
Prove that the distance measure 'd' is a metric.

Measures that satisfy all three following properties are called metrics:

1) Positivity

$d(x,x) \geq 0$ } $\forall x, y$
$d(x,y) = 0$ } $\forall x = y$

2) Symmetry

$d(x,y) = d(y,x)$, $\forall x y$

3) Triangle inequality

$d(x,z) \leq d(x,y) + d(y,z)$, $\forall x, y, z$

Good Write

Now, let $A = \{1,2,3,4\}$
$B = \{1,2,3\}$
$C = \{1,2\}$

size $(A-B) = \{4\}$
size $(B-A) = \phi$

(I) $d(A-B)$ = size $(A-B)$ + size $(B-A)$
$= \{4\} + \phi$
$= \{4\} = 1$ element

$\therefore$ Positivity satisfied

(II) $d(B-A)$ = size $(B-A)$ + size $(A-B)$
$= \phi + \{4\}$
$= \{4\} = 1$ element

$\Rightarrow d(A-B) = d(B-A)$

$\therefore$ symmetry satisfied

(III) $d(A-C)$ = size $(A-C)$ + size $(C-A)$
$= \{3,4\} + \phi$
$= 2$ elements

$d(C-B)$ = size $(C-B)$ + size $(B-C)$
$= \phi + \{3\}$
$= 1$ element

Now,
$d(A-C) + d(C-B)$
$= 3$ elements

Good Write

and $d(A-B) = 1$ element

$\therefore d(A,B) < d(A,C) + d(C,B)$

$\therefore$ Triangle inequality satisfied.

As all three properties are satisfied

$\Rightarrow d$ is a metric.

(c) What is unsupervised learning? Explain with the help of an example application.

Unsupervised learning is the training of a machine using information that is neither classified nor labelled and allowing algorithm to act on that information w/o guidance. Here, the task of the machine is to group unsorted information according to similarity, patterns, differences w/o any prior training of data.

for instance: Take a collection of 1000 essays written on National Parks and find a way to automatically group these into a small number that are somehow similar or related by different variables such as sentence length, page count etc.

3.(a) Consider the following dataset for a 2-class problem:

i) Calculate the gain in the Gini Index when splitting on A and B.

ii) Which attribute would the decision tree induction algo choose?

iii) Draw the decision tree after spliting showing the number of instances of each class.

iv) How many instances are misclassified by the resulting decision tree?

A	B	Class label
T	F	+
T	T	+
T	T	+
T	F	−
T	T	+
F	F	−
F	F	−
F	F	−
T	T	−
T	F	−

Contingency tables after splitting on attribute A, B:

A = T	A = F
4 +	0 +
3 −	3 −

B = T	B = F
3 +	1 +
1 −	5 −

$$\text{Gini Index} = 1 - \sum_{i=0}^{c-1} [p(i|t)]^2$$

Gini Index before splitting

Node	Count
+	4
−	6

$$\text{Gini} = 1 - \left(\frac{4}{10}\right)^2 - \left(\frac{6}{10}\right)^2$$

$$= 1 - (0.4)^2 - (0.6)^2 = 0.4898$$

* On splitting A

$Gini_{A=T} = 1 - \left(\frac{4}{7}\right)^2 - \left(\frac{3}{7}\right)^2$

$= 1 - 0.3265 - 0.1836$

$= 0.4898$

$Gini_{A=F} = 1 - \left(\frac{3}{3}\right)^2 - \left(\frac{0}{3}\right)^2$

$= 0$

* On splitting B

$Gini_{B=T} = 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2$

$= 1 - 0.5625 - 0.0625$

$= 0.375$

$Gini_{B=F} = 1 - \left(\frac{1}{6}\right)^2 - \left(\frac{5}{6}\right)^2$

$= 1 - 0.0256 - 0.6889$

$= 0.2855$

* Gain on splitting A :-

$Gain = Gini - \frac{7}{10} Gini_{A=T} - \frac{3}{10} Gini_{A=F}$

$= 0.4898 - (0.7)(0.4898) - (0.3)(0)$

$= 0.4898 - 0.3428$

$= 0.1371$

* Gain on splitting B

$Gain = Gini - \left(\frac{4}{10}\right) Gini_{B=T} - \left(\frac{6}{10}\right) Gini_{B=F}$

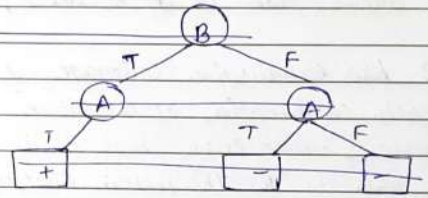
$= 0.4898 - (0.4)(0.375) - (0.6)(0.2855)$

$= 0.4898 - 0.15 - 0.1713$

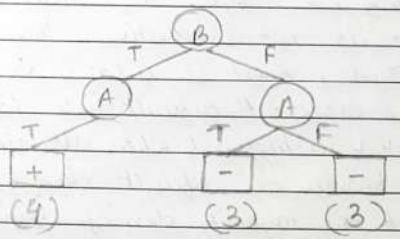
$= 0.1635$

ii) Attribute B will be choosen as $Gain_A < Gain_B$

iii)



iii)



(iv) 2

(b) Why is K-nearest neighbor classifier a lazy learner ?

A lazy learner does not have a training phase. KNN is a lazy learner b/c it doesn't learn a discriminative function from the training data but "memorize" the training data instead. for eg:- the logistic regression algo learns the model weight during training time. Although it sounds convinient, it doesn't come w/o a consequence. The prediction step in KNN is relatively expensive.

4 (a) What is exhaustive rule-sets in Rule based classification? If the rule-set is not exhaustive, what problem arises? How is it resolved?

A rule set R has exhaustive coverage if there is a rule for each combination of attribute values. This property ensures that every record is covered by at least one rule in R. The problem that may arise is that a record may not trigger any rule out of the ruleset. If the rule set is not exhaustive, then a default rule, $r_d : () \rightarrow y_d$, must be added to cover the remaining ~~case~~ cases. A default rule has an empty antecedent and is triggered when all other rules have failed. y_d is known as default class and is typically assigned to the majority class of training records not covered by the existing rules.

(b) What is progressive sampling? What are its advantages?

A proper sample size can be difficult to ~~find~~ determine so progressive scheme is used. It starts with a small sample, and then increase the sample size until a sample of sufficient size has been obtained.

Advantages:

1. Data set is compressed substantially in the first step thereby reducing the size of i/p used.
2. Accuracy of predictive models increases as the sample size increases.

(c) State Bayes' theorem. What assumptions is used by the Naive Bayes classifier?

The formula below is known as Bayes theorem

$$P(Y|X) = \frac{P(X|Y)\,P(Y)}{P(X)}$$

Bayes theorem describes the probability of an event based on precedent knowledge of conditions which might be related to the event.

Assumptions of Naive Bayes ~~theorem~~ classifier:
It estimates the class-conditional probability by assuming that the attributes are conditionally independent, given class label y.

$$P(X|Y=y) = \prod_{i=1}^{d} P(X_i|Y=y)$$

where each attribute set $X = \{X_1, X_2, \ldots X_d\}$ consists of d attributes.

Good Write

5(b).

5. (a) Consider the following set of frequent 3-itemset:

{1,2,3}, {1,2,4}, {1,2,5}, {1,3,4}, {1,3,5}, {2,3,4}, {2,3,5}, {3,4,5}.

Assume that there are only five items in the dataset.

(i) List all ~~con~~ candidate 4-itemsets obtained by a candidate generation procedure using the $F_{k-1} \times F_1$ merging strategy.

Using $F_{k-1} \times F_1$ strategy

Here $k = 4$.

Frequent 3 itemset
{1,2,3}
{1,2,4}
{1,2,5}
{1,3,4}
{1,3,5}
{2,3,4}
{2,3,5}
{3,4,5}

Frequent 1-itemset
{1}
{2}
{3}
{4}
{5}

Using lexicographical order pruning

Candidate :-
{1,2,3,4}
{1,2,3,5}
{1,2,4,5}
{1,3,4,5}
{2,3,4,5}

(ii) List all candidate 4-itemset obtained by a candidate generation procedure in Apriori.

Frequent 3 itemset	Frequent 4 Itemset	Itemset count
{1,2,3}	{1,2,3,4}	4
{1,2,4}	{1,2,3,5}	3
{1,2,5}	{1,2,4,5}	1
{1,3,4}	{1,3,4,5}	2
{1,3,5}	{2,3,4,5}	1
{2,3,4}		
{2,3,5}		
{3,4,5}		

1	frequency	support
1	5	5/8 x100 = 62.5%
2	5	5/8 x100 = 62.5%
3	6	6/8 x100 = 75%
4	3	3/8 x100 = 37.5%
5	4	4/8 x100 = 50%

6. Consider the following transactional dataset:

Transaction ID	Items Bought
0001	{a, d, e}
0002	{a, b, c, e}
0003	{a, b, d, e}
0004	{a, c, d, e}
0005	{b, c, e}
0006	{b, d, e}
0007	{c, d}
0008	{a, b, c}
0009	{a, d, e}
0010	{a, b, e}

(i) find out the support of items {e}, {b,d}, {a,d} and {b,d,e}. Are these itemsets frequent if min support threshold is 30%?

$$\text{Suppor} = \frac{\sigma(\{\})}{|T|}$$

$$s(\{e\}) = \frac{\sigma(\{e\})}{10} = \frac{8}{10} = 0.8$$

$$s(\{b,d\}) = \frac{2}{10} = 0.2$$

$$s(\{a,d\}) = \frac{4}{10} = 0.4$$

$$s(\{b,d,e\}) = \frac{2}{10} = 0.2$$

Min support = 30% = $\frac{30}{100} = 0.3$

Itemset {e} and {a,d} are frequent.
Itemset {b,d} and {b,d,e} are not frequent.

(ii) find all the rules generated from the 3-itemset {b,d,e}. List down the strong rules and those rule if min confidence threshold is 60%.

No. of rules = $2^k - 2 = 2^3 - 2 = 8 - 2 = 6$

R_1 : {b,d} → {e}
R_2 : {b,e} → {d}
R_3 : {d,e} → {b}
R_4 : {b} → {d,e}
R_5 : {d} → {b,e}
R_6 : {e} → {b,d}

confidence of $X \to Y = \frac{\sigma(X \cup Y)}{\sigma X}$

$$c(R_1) = \frac{\sigma(\{b,d,e\})}{\sigma(\{b,d\})} = \frac{2}{2} = 1$$

$$c(R_2) = \frac{\sigma(\{b,d,e\})}{\sigma\{b,e\}} = \frac{2}{5} = 0.4$$

$$c(R_3) = \frac{\sigma(\{b,d,e\})}{\sigma(\{d,e\})} = \frac{2}{5} = 0.4$$

$$c(R_4) = \frac{\sigma(\{b,d,e\})}{\sigma(\{b\})} = \frac{2}{6} = 0.33$$

$$c(R_5) = \frac{\sigma(\{b,d,e\})}{\sigma(\{d\})} = \frac{2}{6} = 0.33$$

$$c(R_6) = \frac{\sigma(\{b,d,e\})}{\sigma(\{e\})} = \frac{2}{8} = 0.25$$

minconf = 60% = 0.6
∴ Strong rules are :-
R1

Good Write

(b) What is the difference b/w nominal attribute and ordinal attribute ? Give an example of each.

Nominal Attribute	Ordinal Attribute
• The values of a nominal attribute are just different names i.e nominal values provide only enough info to distinguish one object from another. ($=, \neq$)	• The values of ordinal attribute provide info to order objects. ($<, >$)
• Operations : Mode, entropy, contingency correlation, χ^2 test.	• Operations : Median, Rank, percentiles, correlations, run tests.
• Eg: Zip codes, employee id, eye color, gender.	• Eg: hardness of minerals, {good, better, best}, grades.

7.(a) Explain the following terms with reference to the DBSCAN clustering algo.

i) Core point
ii) Noise point
iii) Border point

~~Core~~ i) Done in 2018 1.(k) (i)
ii) Done in 2018 1.(k) (ii)

iii) Border point – A border pnt is not a core point, but falls within its neighborhood. A border pnt can fall within the neighborhood of several core pnts.

(b) Given the following data pnts : 2, 4, 10, 12, 3, 20, 30, 11, 25. Assume K=3 and initial means 2, 4, 6. Show the clusters obtained using K-means algo after two iterations and show the new means for the next iterations.

Done in 2018 7.(a)

Clusters obtained after two iterations :-

cluster 1 = 2, 3
cluster 2 = 4, 10
cluster 3 = 12, 20, 30, 11, 25

new means for next iterations :-

$\mu_1 = 2.5$
$\mu_2 = 7$
$\mu_3 = 19.6$

2020 QP No.

1. Given the following table, classify all the attributes appearing in the table as binary, discrete or continuous. Also classify them as qualitative (nominal or ordinal) or quantitative (ratio or interval). Justify your answer in each case. Show the normalization (scaling data b/w 0 and 1) of values in the age attribute column. How can you handle missing values in Age column and Height column? Replace Y and N respectively by 1 and 0 in the first column. Replace M and F respectively by 0 and 1 in the second column. You will get one binary vector each for Smoker attribute and Gender attribute. Find out the similarity measure b/w these two vectors using Jaccard coefficient.

Smoker	Gender	Age	Height	Marital Status
Y	F	32	Tall	Married
Y	M	34	Medium	Maries
N	F	39	Medium	Single
Y	M	41	Tall	Single
Y	M	25	Tall	Divorcee
N	M	36	Tall	Single
Y	F	45	Short	Married
Y	M	31	Tall	Single
N	M	29	Medium	Divorcee
N	F	51	Tall	Single
Y	F	38	short	Married

Good Write

i) Smoker - Binary, Nominal
Gender - Binary, Nominal
Age - Continuous, Ratio
Height - Discrete, Ordinal
Marital Status - Discrete, Ordinal

Nominal attributes provide only enough info to distinguish one object from other.
Eg: Smoker (Y/N), Gender (F/M)

for Ratio variables, both difference b/w values and ratios are meaningful.
Eg: Age.

for Ordinal attributes, the value provides enough info to order objects.
Eg: Height {Tall, Medium, short}, Marital Status {Married, Single, Divorcee}

ii) Using Min-Max Normalization

$$x' = \frac{x - min\_x}{max\_x - min\_x}$$

$min\_x = 25$
$max\_x = 51$

Age = {32, 34, 39, 41, 25, 36, 45, 31, 29, 51, 38}
a_1 a_2 a_3 a_4 a_5 a_6 a_7 a_8 a_9 a_{10} a_{11}

$$a_1' = \frac{32-25}{51-25} = \frac{7}{26} = 0.269$$

Good Write

$a_2' = \frac{34-25}{51-25} = \frac{9}{26} = 0.346$

$a_3' = \frac{39-25}{51-25} = \frac{14}{26} = 0.538$

$a_4' = \frac{41-25}{51-25} = \frac{16}{26} = 0.615$

$a_5' = \frac{25-25}{51-25} = 0$

$a_6' = \frac{36-25}{51-25} = \frac{11}{26} = 0.423$

$a_7' = \frac{45-25}{51-25} = \frac{20}{26} = 0.769$

$a_8' = \frac{31-25}{51-25} = \frac{6}{26} = 0.23$

$a_9' = \frac{29-25}{51-25} = \frac{4}{26} = 0.153$

$a_{10}' = \frac{51-25}{51-25} = 1$

$a_{11}' = \frac{38-25}{51-26} = \frac{13}{26} = 0.5$

Data is normalized b/w 0 & 1

Good Write

New Age = { 0.269, 0.346, 0.538, 0.615, 0.0423, 0.769, 0.23, 0.153, 1, 0.5 }

(iii) Replace Y & N with 1 and 0 respectively.
M and F with 0 and 1 respectively.

Smoker	Gender	Age	Height	Marital Status
1	1	0.269	Tall	Married
1	0	0.346	Medium	Married
0	1	0.538	Medium	Single
1	0	0.615	Tall	Single
1	0	0	Tall	Divorcee
0	0	0.423	Tall	Single
1	1	0.769	Short	Married
1	0	0.23	Tall	Single
0	0	0.153	Medium	Divorcee
0	1	1	Tall	Single
1	1	0.5	Short	Married

Calculating Jaccard coefficient:

$x = (1, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1)$
$y = (1, 0, 1, 0, 0, 0, 1, 0, 0, 1, 1)$

$f_{xy} = f_{01} = 2$ $\quad f_{xy} = f_{00} = 2$

$f_{xy} = f_{10} = 4$ $\quad f_{xy} = f_{11} = 3$

$J = \frac{f_{11}}{f_{01}+f_{10}+f_{11}} = \frac{3}{2+4+3} = \frac{3}{9} = 0.33$

Good Write

iv) Handle missing values in age and height column:

i) Estimate data attributes

Here we will estimate the object with missing age and height values.

ii) Estimate missing values

Sometimes, missing data can be reliably estimated. In age, the average attribute value of the nearest neighbor is used & in height, the most commonly occuring attribute value is taken.

iii) Ignore the missing value during analysis.

2. Given the following binary classification problem:

Instance	A_1	A_2	Target Class
1	T	T	+
2	T	T	+
3	T	F	−
4	F	F	+
5	F	T	−
6	F	T	−
7	F	F	−
8	T	F	+
9	F	T	−
10	T	F	−

Calculate separately the info gain when splitting is done on A_1 and on A_2. Which attribute would the decision tree induction algo choose? Calculate separately the gain in the Gini index when splitting is done on A_1 and A_2. Which attribute would the decision tree induction algo choose? Is it possible that info gain and the gain in Gini index favour different attributes? Explain your answer.

i) Contingency tables after splitting on attributes A_1 and A_2

$A_1 = T$	$A_1 = F$
3+	1+
2−	4−

$A_2 = T$	$A_2 = F$
2+	2+
3−	3−

Gini Index before splitting

Node	Count
+	4
−	6

$$\text{Gini} = 1 - \sum_{i=1}^{n} (p_i; t)^2$$

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

$$= 1 - (0.4)^2 - (0.6)^2$$

$$= 0.4898$$

* on splitting A_1

$$\text{Gini } A_1 = T = 1 - (3/5)^2 - (2/5)^2$$

$$= 1 - (0.6)^2 - (0.4)^2$$

$$= 0.4898$$

Gini A=F = $1 - (1/5)^2 - (4/5)^2$

$= 1 - 0.04 - 0.64$

$= 0.32$

Gain = Gini $- \frac{5}{10}$ Gini A=T $- \frac{5}{10}$ Gini A=F

$= 0.4898 - (0.5)(0.4898) - (0.5)(0.32)$

$= 0.08$

* on splitting A_2

Gini A_2=T $= 1 - (\frac{3}{5})^2 - (\frac{2}{5})^2$

$= 0.48$

Gini A=F $= 1 - (\frac{2}{5})^2 - (\frac{3}{5})^2$

$= 0.48$

Gain = Gini $- \frac{5}{10}$ Gini A_2=T $- \frac{5}{10}$ Gini A_2=F

$= 0.48 - (0.5)(0.48) - (0.5)(0.48)$

$= 0.48 - 0.24 - 0.24$

$= 0$

ii) A_1 will be choosen as Gain A_1 > Gain A_2

iii) After Information gain :-

Entropy before splitting

Entropy $= -\frac{4}{10}\log_2\frac{4}{10} - \frac{6}{10}\log_2\frac{6}{10}$

$= -(0.4)(\log_2 0.4) - (0.6)(\log_2 0.6)$

$= -(0.4)(-1.322) - (0.6)(-0.737)$

$= 0.5288 + 0.4422$

$= 0.971$

* on splitting A_1

Entropy $A_{11} = -P_1\log_2 P_1 - P_2\log_2 P_2$

$= -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}$

$= -(0.6)(-0.737) - (0.4)(-1.322)$

$= 0.971$

Entropy $A_{12} = -\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5}$

$= -(0.2)(-2.322) - (0.8)(-0.322)$

$= 0.4644 + 0.2576$

$= 0.722$

Weighted Entropy = P_1 Entropy A_{11} + P_2 Entropy A_{12}

$= \frac{5}{10}(0.971) + \frac{5}{10}(0.722)$

$= 0.4855 + 0.361$

$= 0.8465$

Info Gain = Entropy − Weighted Entropy

$= 0.971 - 0.8465$

$= 0.1245$

* On splitting A_2

Entropy $A_{21} = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}$

$= -(0.4)(-1.322) - (0.6)(-0.737)$

$= 0.971$

Entropy $A_{22} = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}$

$= 0.971$

Weighted Entropy $= \frac{5}{10}(0.971) + \frac{5}{10}(0.971)$

$= 0.971$

Info Gain = Entropy − Weighted Entropy

$= 0.971 - 0.971$

$= 0$

iv) A_1 will be chosen as Info Gain A_1 > Info Gain A_2

v) Yes, they can favour diff. attributes. Gini index operate on categorical target variables in terms of 'success' and or 'failure' and performs only binary split. Info gain computes the difference b/w entropy before and after the split and indicates the impurity in class elements. In addition, info gain favour small partitions with distinct values where as gini index favours large ones.

3. Consider the one dimension labeled data set given below:

x:	0.5	3.0	4.5	4.6	4.9	5.2	5.3	5.5	7.0	9.5
y:	−	−	+	+	−	−	+	+	−	−

Classify the data pnt $x = 5.0$ according to its 3- & 5- nearest neighbour using majority vote.

Suppose that there are a total of 60 data mining related documents in a library of 200 documents. Suppose that a search engine retrieves 20 documents after a user enters 'data mining' as a query, of which 5 are data mining related docs. What are the precision and recall?

i) $X = 5.0$

1) 3- nearest neighbour
2) 5- nearest neighbour

Distance : $|x_i - x|$

X_i	Y	$\lvert x_i - x \rvert$	3-NN	5-NN
0.5	-	4.5		
3.0	-	2.0		
4.5	+	0.5		
4.6	+	0.4		+
4.9	-	0.1	-	-
52	-	0.2	-	-
5.3	+	0.3	+	+
5.5	+	0.5		+
7.0	-	2.0		
9.5	-	4.5		

using voting scheme

1) 3NN : 2- , 1+
∴ Y = -

2) 5NN : 2- , 3+
∴ y = +

ii) Precision = $\frac{TP}{TP+FP}$

Recall = $\frac{TP}{TP+FN}$

Total Records = 60
Retrieved doc. = 20
Relevant doc. = 5 (TP)

FN = Total records - TP
= 60-5
= 55

FP = Retrieved doc - TP
= 20-5
= 15

∴ Recall = $\frac{5}{5+55} = \frac{5}{60} = 0.083$

Precision = $\frac{5}{5+15} = \frac{5}{20} = 0.25$

4. Consider the Market Basket dataset shown below:

Transaction Id	Items Bought
0001	{a, d, e}
0024	{a, b, c, e}
0012	{a, b, d, e}
0031	{a, c, d, e}
0015	{b, c, e}
0022	{b, d, e}
0029	{c, d}
0040	{a, b, c}
0033	{a, d, e}
0038	{a, b, e}

Compute support for itemsets {e}, {b,d}, {b,d,e}, {a,b,c,e} & {a,b,c,d,e}. Find all association Rules that can be generated from item set {b,d,e} along with the confidence of each rule. Is confidence a symmetric measure?

(i) $s(\{e\}) = \frac{8}{10} = 0.8$

$s(\{b,d\}) = \frac{2}{10} = 0.2$

$s(\{b,d,e\}) = \frac{2}{10} = 0.2$

$s(\{a,b,c,e\}) = \frac{1}{10} = 0.1$

$s(\{a,b,c,d,e\}) = 0$

ii) All associated rules ~~are~~ from {b,d,e} are:-

$R_1 : \{b,d\} \rightarrow \{e\}$
$R_2 : \{b\} \rightarrow \{d,e\}$
$R_3 : \{d\} \rightarrow \{b,e\}$
$R_4 : \{e\} \rightarrow \{b,d\}$
$R_5 : \{d,e\} \rightarrow \{b\}$
$R_6 : \{b,e\} \rightarrow \{d\}$

~~c(R1) = Done in 20~~ Done in 2019 6(ii)

$c(R_1) = \frac{2}{2} = 1$

$c(R_2) = \frac{2}{6} = 0.33$

$c(R_3) = \frac{2}{6} = 0.33$

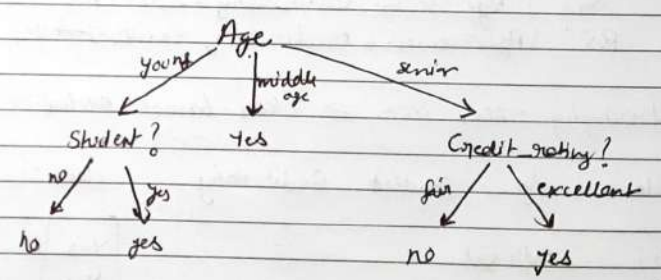
$c(R_4) = \frac{2}{8} = 0.25$

$c(R_5) = \frac{2}{5} = 0.4$

$c(R_6) = \frac{2}{5} = 0.4$

iii) ~~No~~ Done in 2018. 2.(a)(iii)

5. Generate ~~the genes~~ all the rules given the following decision tree.



Arrange the generated rule ~~accord~~ acc. to class based ordering and classify the following tuples:

Age = young, Studat = no, loan_approval = ?
Age = senior, credit_rating = excellent, loan_approval = ?

Consider a training set that catain 90 +ve eg. and 300 -ve eg. for each of the following rules:

R_1 : A → + (covers 30 +ve & 10-ve eg)

R2 : B → + (covers 90 +ve & 80-ve eg)

R3 : C → + (covers 4 +ve & 1-ve eg)

Determine which is the best and which is the worst candidate rule acc. to rule Accuracy.

i) Rules generated are :-

R_1 : (Age = middle_aged) → Yes
R2 : (Age = young, student != no) → No
R3 : (Age = young, student = yes) → yes
R4 : (Age = senior, credit_rating = fair) → No
R5 : (Age = senior, credit_rating = excellent) → Yes

Arranging them acc to class-based order :-

Rule	Age	Student	Credit Rating	class
R_1	middle aged	-	-	Yes
R3	young	Yes	-	Yes
R5	~~young~~ senior	-	Excellent	Yes
R2	young	No	-	No
R4	senior	-	fair	No

R_y : {R_1, R3, R5}
R_N : {R2, R4}

Good Write

ii) Age = young
student = No

so acc. to R3
loan_approval = NO

Age. = Senior
Credit Rating = excellent
so acc. to R5
loan_approval = Yes

iii) * R_1 :

Total Observation = 30+10 = 40

Accuracy = No. of +ve Observation / Total Observation

= 30/40

= 0.75

* R2

Total Records = 90+80 = 170

Accuracy = 90/170 = 0.52

* Rule R3 :

Total record = 4+1 = 5

Accuracy = 4/5 = 0.80

Accuracy R_2 < Accuracy R_1 < Accuracy R_3

∴ Best Rule is R_3
Worst Rule is R_2

6. Use K-means algo and Euclidean distance to cluster five data pnts (A_4-A_8) given below, into 3 clusters. The coordinates of the data pnts are:

$A_1(2,8)$, $A_2(2,5)$, $A_3(1,2)$, $A_4(5,8)$, $A_5(7,3)$, $A_6(6,4)$, $A_7(8,4)$, $A_8(4,7)$.

Use A_1, A_2 and A_3 as initial ~~cluster~~ centroids. For which situations K-means clustering will give good results and when will it fail to produce good results?

K=3

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

Intial means : $\mu_1 = (2,8)$, $\mu_2 = (2,5)$, $\mu_3 = (1,2)$

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

Iteration 1:

(x_i, y_i)	$A_1(2,8)$	$A_2(2,5)$	$A_3(1,2)$	cluster
A_1 2,8	0	3	7	A_1
A_2 2,5	3	0	4	A_2
A_3 1,2	7	4	0	A_3
A_4 5,8	3	6	10	A_1
A_5 7,3	10	7	7	A_3
A_6 6,4	8	5	7	A_2
A_7 8,4	10	7	9	A_2
A_8 4,7	3	4	8	A_1

cluster 1 = { ~~(2,8)~~ ~~(2,8)~~, $A_1(2,8)$, $A_4(5,8)$, $A_8(4,7)$ }
cluster 2 = { $A_2(2,5)$, $A_6(6,4)$, $A_7(8,4)$ }
cluster 3 = { $A_3(1,2)$, $A_5(7,3)$, }

New $\mu_1 = \left(\frac{2+5+4}{3}, \frac{8+8+7}{3}\right) = (3.66, 7.66)$

New $\mu_2 = \left(\frac{2+6+8}{3}, \frac{5+4+4}{3}\right) = (5.33, 4.33)$

New $\mu_3 = \left(\frac{1+7}{2}, \frac{2+3}{2}\right) = (4, 2.5)$

Iteration 2:

(x_i, y_i)	$A_1(3.66, 7.66)$	$A_2(5.33, 4.33)$	$A_3(4, 2.5)$	cluster
$A_1(2,8)$	2	7	7.5	A_1
$A_2(2,5)$	4.32	4	4.5	A_2
$A_3(1,2)$	8.32	6.66	3.5	A_3
$A_4(5,8)$	1.68	4	6.5	A_1
$A_5(7,3)$	8	2	3.5	A_2
$A_6(6,4)$	6	1	3.5	A_2
$A_7(8,4)$	8	3	5.5	A_2
$A_8(4,7)$	1	4	4.5	A_1

Cluster 1 = { $A_1(2,8)$, $A_4(5,8)$, $A_8(4,7)$ }
Cluster 2 = { $A_2(2,5)$, $A_5(7,3)$, $A_6(6,4)$, $A_7(8,4)$ }
Cluster 3 = { $A_3(1,2)$ }

New $\mu_1 = \left(\frac{2+5+4}{3}, \frac{8+8+7}{3}\right) = (3.66, 7.66)$

New $\mu_2 = \left(\frac{2+7+6+8}{4}, \frac{5+3+4+4}{4}\right) = (5.75, 4)$

New $\mu_3 = (1,2)$

Iteration 3:

(x_i, y_i)	$A_1(3.66, 7.66)$	$A_2(5.75, 4)$	$A_3(1,2)$	cluster
$A_1(2,8)$	2	7.75	7	A_1
$A_2(2,5)$	4.32	4.75	4	A_3
$A_3(1,2)$	8.32	6.75	0	A_3
$A_4(5,8)$	1.68	4.75	10	A_1
$A_5(7,3)$	8	2.25	7	A_2
$A_6(6,4)$	6	0.25	7	A_2
$A_7(8,4)$	8	2.25	9	A_2
$A_8(4,7)$	1	4.75	8	A_1

Cluster 1 = { $A_1(2,8)$, $A_4(5,8)$, $A_8(4,7)$ }
Cluster 2 = { $A_5(7,3)$, $A_6(6,4)$, $A_7(8,4)$ }
Cluster 3 = { $A_2(2,5)$, $A_3(1,2)$ }

New $\mu_1 = \left(\frac{2+5+4}{3}, \frac{8+8+7}{3}\right) = (3.66, 7.66)$

New $\mu_2 = \left(\frac{7+6+8}{3}, \frac{3+4+4}{3}\right) = (7, 3.66)$

New $\mu_3 = \left(\frac{2+1}{2}, \frac{5+2}{2}\right) = (1.5, 3.5)$

Iteration 4:

xiyi	$A_1(3.66, 7.66)$	$A_2(7, 3.66)$	$A_3(1.5, 3.5)$	Cluster
A_1 (2, 8)	2	9.34	5	A_1
A_2 (2, 5)	4.32	6.34	2	A_3
A_3 (1, 2)	8.32	7.66	2	A_3
A_4 (5, 8)	1.68	6.34	8	A_1
A_5 (7, 3)	8	0.66	6	A_2
A_6 (6, 4)	6	1.34	5	A_2
A_7 (8, 4)	8	1.34	7	A_2
A_8 (4, 7)	1	6.34	6	A_1

cluster 1 = { $A_1(2,8)$, $A_4(5,8)$, $A_8(4,7)$ }
Cluster 2 = { $A_5(7,3)$, $A_6(6,4)$, $A_7(8,4)$ }
Cluster 3 = { $A_2(2,5)$, $A_3(1,2)$ }

New $\mu_1 = \left(\frac{2+5+4}{3}, \frac{8+8+7}{3}\right) = (3.66, 7.66)$

New $\mu_2 = \left(\frac{7+6+8}{3}, \frac{3+4+4}{3}\right) = (7, 3.66)$

New $\mu_3 = \left(\frac{2+1}{2}, \frac{5+2}{2}\right) = (1.5, 3.5)$

As values for centroids from iteration 3 and 4 are same.
Algo stops here

Good Write

clusters formed :-
cluster 1 = { A_1, A_4, A_8 }
cluster 2 = { A_5, A_6, A_7 }
cluster 3 = { A_2, A_3 }

K-means will give good results when the clusters have a kind of spherical shape. It will fail to produce so when the data contains outliers and clusters are of different sizes, density and non globular shape.

Good Write

2021

1. Consider a sample dataset of patients visiting a clinic for consultation:

Patient ID	Patient Name	Blood Pressure	Blood Pres. Date	Chest Pain	Age	Exercise	Heart condition
101	Sarita	120	14-10-1995	1	20	High	Good
102	Maddy	110	16-1-2018	0	40	High	Good
103	Rohit	140	18-10-2018	3	?	Low	Bad
104	Gauhar	172	4-6-2018	3	39	Medium	Bad
105	Himani	150	8-6-2018	2	35	Low	Bad
106	Shubham	110	4-7-2018	1	40	Medium	Good
107	Suresh	120	5-3-2018	0	26	High	Good
108	Ammol	110	5-5-2018	1	27	Medium	Good

- Identify the type of each attribute and give justification.
- Perform min-max normalisation on "Blood Pressure" attribute.
- Identify the data quality issues for each of the following fields: Can these issues be resolved? If yes, how?
  - Blood Pressure date for patient- 101
  - The BP meter is miscalibrated and add 1 Hg 1 mmHg to each reading.
  - Age of patient 103.

- Attribute : Type

PID : Nominal
Patient Name : Nominal
Blood Pressure : Ratio
Blood Pressure Dat : Interval
Chest Pain : Ordinal
Age : Ratio
Exercise : Ordinal
Heart Condition : Ordinal

Nominal attributes have values that are just different names. ie. they provide only enough info to distinguish one object from another.

Ordinal attributes provide enough info to order objects.

Interval attributes, the difference b/w values is meaningful.

Ratio attributes, both difference and ratio are meaningful.

- $x_{min} = 110$
  $x_{max} = 172$

$X = \{ \underset{x_1}{120}, \underset{x_2}{110}, \underset{x_3}{140}, \underset{x_4}{172}, \underset{x_5}{150}, \underset{x_6}{110}, \underset{x_7}{120}, \underset{x_8}{110} \}$

$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$

$x_1' = \frac{120-110}{172-110} = \frac{10}{62} = 0.16$

$x_3' = \frac{140-110}{172-110} = \frac{30}{62} = 0.48$

$x_2' = \frac{110-110}{172-110} = 0$

$x_4' = \frac{172-110}{172-110} = 1$

$x_5' = \frac{150-110}{172-110} = \frac{40}{62} = 0.64$

$x_6' = \frac{110-110}{172-110} = 0$

$x_7' = \frac{120-110}{172-110} = 0.16$

$x_8' = \frac{110-110}{172-110} = 0$

$X' = \{0.16, 0, 0.48, 1, 0.64, 0, 0.16, 0\}$

i) Blood Pressure date for patient 101 is an outlier as it value is unusual with respect to the typical values of the attribute.
This can be resolved by changing the outlier to replace them by percentile-trimmed minimum and maximum.

ii) It is a measurement error which refers to any problem resulting from the measurement proces. Here, we know the amt of error : the true value can be recovered either analytically or numerically. This is called memo-of-moments or functional approach.

iii) It is a missing value. There are several strategies for resolving it. One such way is to estimate the value. As age is a continuous attribute, the average attribute value of the nearest neighbour is used.

2. A binary classification problem with class labels: 'Yes' and 'No' that denotes the access to the elevator, has the following set of attributes and attribute values:

Status = { faculty, Student }
floor = { first, Second, Third, fourth, fifth }
Health = { healthy, unhealthy }

Consider the following set of records for the above classification problem.

Transaction #	Status	floor	Health	Accessible
1	faculty	first	Healthy	Yes
2	Student	first	Healthy	No
3	Student	Third	Healthy	No
4	faculty	Second	Unhealthy	Yes
5	Student	fifth	Healthy	Yes
6	faculty	first	Healthy	No
7	Student	fifth	Unhealthy	No

A rule-based classifier produces the following rule set:

R_1 : Status = faculty, Floor = Second → Accessible = Yes
R_2 : Status = Student, Floor = Second → Accessible = No
R_3 : Floor = First → Accessible = No
R_4 : Health = Unhealthy → Accessible = Yes
R_5 : Status = Student, Floor = Fifth, Health = Healthy → Accessible = Yes

- Are the rules in the above rule set mutually exclusive? Justify.
- Is the rule set exhaustive? Justify.
- Is ordering needed for this set of rules? Justify
- Do you need a default class for the rule set? Justify
- Compute the coverage and accuracy of rules R_1 and R_5. Which one do you think is a better rule? Why?

- The rules in a rule set R are mutually exclusive if no two rules in R are triggered by the same record.

Record	Triggered Rule
1	None
2	R3
3	None
4	R_1, R_4
5	R_5
6	R3
7	None

Record 4 is triggered by triggering R_1 & R_4 Hence not mutually exclusive.

Good Write

- A rule set R has exhaustive coverage if there is a rule for each combination of attribute values. As record 1, 3, 7 are not triggered triggering any rule, this rule set is not exhaustive.

- Yes, as classifying a new test record can be quite expensive as the attribute of the test record may be compared against the precondition of every rule in the rule set. Also record 4 is triggered by more than 1 rule ∴ ordering is needed to make it mutually exclusive.

- Yes, as there are not exhaustive ie a record may not trigger any rule (eg: Record 1). In such a case, a default rule $r_d : () \rightarrow y_d$ must be added to cover the remaining cases.

- ✱ R_1:

Total records triggered = 1

$$\text{Accuracy} = \frac{\text{No. of true eg}}{\text{Total obs}} = \frac{|A \cap y)}{|A|} = \frac{1}{1} = 1$$

$$\text{Coverage } (r_1) = \frac{|A|}{|D|} = \frac{1}{7} = 0.14$$

- ✱ R_5:

$$\text{Accuracy} = \frac{|A \cap y)}{|A|} = \frac{1}{1} = 1$$

$$\text{Coverage} = \frac{|A|}{|D|} = \frac{1}{7} = 0.14$$

Good Write

As coverage and accuracy are same for both rules R_1 and R_5, neither is better than the other.

3. Consider the training examples shown in the following table for a classification problem.

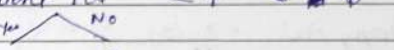
Student ID	Admission Category	Admission List	Gender	Predicted class	Actual class
1	Sports	first	M	C_0	C_0
2	Arts	Second	F	C_0	C_1
3	Arts	Third	F	C_0	C_0
4	Sports	fourth	M	C_0	C_1
5	Academics	first	F	C_0	C_0
6	Arts	Second	F	C_1	C_1
7	Arts	Third	M	C_1	C_1
8	Sports	fourth	M	C_1	C_0
9	Sports	Third	F	C_1	C_1
10	Arts	first	F	C_1	C_0
11	Sports	Third	M	C_1	C_1
12	Sports	Second	F	C_1	C_1
13	Sports	first	M	C_1	C_0

- Compute the Info Gain for the Student Id attribute.
- Compute the Gini Index for the Admission Category Attribute and Admission List attribute.
- Which is a better attribute for split based on the Gini Index: Admission Category or Admission List ? Why?
- Create a confusion matrix for the above data set and compute false true rate, accuracy, recall and precision.

- On splitting student Id

~~Std~~

- Student Id. is a continuous attribute so we choose ~~the condition~~ Binary split test condition i.e.



A_1	A_2
$C_0 = 3$	$C_0 = 3$
$C_1 = 3$	$C_1 = 4$

* Entropy before Splitting

Node	Count
C_0	6
C_1	7

$$\text{Entropy} = -\frac{6}{13}\log_2\frac{6}{13} - \frac{7}{13}\log_2\frac{7}{13}$$

$$= -\left(\frac{6}{13}\right)\left(\log_2 0.461\right) - \frac{7}{13}\left(\log_2 0.538\right)$$

$$= -\frac{6}{13} \times (-1.1172) - \frac{7}{13} \times (-0.8943)$$

$$= 0.515 + 0.481$$

$$= 0.996$$

~~Entropy A_1 $= -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6}$~~

~~$= -2 \times \left(\frac{1}{2}\log_2 0.5\right)$~~

Entropy A_{ns} = $-\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6}$

$= -2 \times (0.5 \log_2 0.5)$

$= -2 \times (0.5 \times -1)$

$= 1$

Entropy A_s = $-\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7}$

$= -(0.571)(\log_2 0.571) - 4$

$= -(0.428)(\log_2 0.428) - (0.571)(\log_2 0.571)$

$= -(0.428)(-1.22) - (0.571)(-0.808)$

$= 0.522 + 0.461$

$= 0.983$

Weighted Entropy = P_1 Entropy A_1 + P_2 Entropy A_2

$= \frac{6}{13} \times 1 + \frac{7}{13} \times 0.983$

$= 0.461 + 0.528$

$= 0.989$

Info gain = Entropy - Weighted Entropy

$= 0.996 - 0.989$

$= 0.007$

- Gini Index for Admission category.

Node	Count	
Sports	Arts	Academic
$C_0 = 3$	$C_0 = 2$	$C_0 = 1$
$C_1 = 4$	$C_1 = 3$	$C_1 = 0$

Gini$_s$ = $1 - \left(\frac{3}{7}\right)^2 - \left(\frac{4}{7}\right)^2$

$= 1 - \frac{9}{49} - \frac{16}{49}$

$= \frac{49-25}{49} = \frac{24}{49} = 0.489$

Gini$_a$ = $1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2$

$= 1 - \frac{4}{25} - \frac{9}{25}$

$= \frac{25-4-9}{25} = \frac{25-13}{25} = \frac{12}{25} = 0.48$

Gini$_{Ac.}$ = $1 - \left(\frac{1}{1}\right)^2 - 0) = 0.$

Gini Index for Academic List

First	Second	Third	fourth
$C_0 = 4$	$C_0 = 0$	$C_0 = 1$	$C_0 = 1$
$C_1 = 0$	$C_1 = 3$	$C_1 = 3$	$C_1 = 1$

Gini 1st $= 1 - \left(\frac{4}{4}\right)^2 - 0$

$= 0$

Gini 2nd $= 1 - 0 - \left(\frac{3}{3}\right)^2$

$= 0$

Gini 3rd $= 1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2$

$= 1 - \frac{1}{16} - \frac{9}{16}$

$= \frac{16 - 10}{16} = \frac{6}{16} = 0.375$

Gini 4th $= 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2$

$= 1 - \frac{1}{4} - \frac{1}{4}$

$= \frac{4-2}{4} = \frac{2}{4} = 0.5$

Good Write

- Gain on splitting Academic Category

~~Gain =~~

Gini before splitting,

$$\text{Gini} = 1 - \left(\frac{6}{13}\right)^2 - \left(\frac{7}{13}\right)^2$$

$$= 1 - \frac{36}{169} - \frac{49}{169}$$

$$= \frac{169 - 115}{169}$$

$$= \frac{54}{169} = 0.319$$

~~Gain = 0.319 - 0.489 - 0.48 = -0.65~~ ~~0.533~~ ~~0.384~~

~~Gain on splitting Academic list~~ ~~0.307~~ ~~0.153~~

~~Gain = 0.319 - 0.375 - 0.5 =~~

$$\text{Gain} = 0.319 - \frac{7}{13} \times 0.489 - \frac{5}{13} \times 0.48$$

$$= 0.319 - 0.263 - 0.184$$

$$= -0.128$$

$$\text{Gain} = 0.319 - \frac{4}{13} \times 0.375 - \frac{2}{13} \times 0.5$$

$$= 0.319 - 0.144 - 0.076$$

$$= 0.099$$

Good Write

Since Gain of Academic list > Gain of Academic category

Academic list is Choosen

Predicted Class		Actual Class C0	Actual Class C1
	C0	3	2
	C1	3	5

False Positive Rate = $\alpha = \frac{FP}{TN+FP}$

$= \frac{3}{3+3} = \frac{3}{6} = 0.5$

Accuracy = $\frac{TP+TN}{TP+FN+FP+TN}$

$= \frac{3+5}{13} = \frac{8}{13} = 0.615$

Recall = TP Rate = $\frac{TP}{TP+FN}$

$= \frac{5}{5+2} = \frac{5}{7} = 0.714$

Precision = the Predictive value = $\frac{TP}{TP+FP}$

$= \frac{5}{5+3} = \frac{5}{8} = 0.625$

Good Write

4. Consider the following set of pnts:

{44, 28, 48, 26, 32, 14, 52, 50}

Assuming K=2, and initial cluster centers for k-means algo are 54 and 38, compute the sum of squared errors (SSE) and cluster assignment for each iteration.

Iteration 1:

	$\mu_1 = 54$	$\mu_2 = 38$	cluster
x_i	$\|x_i - 54\|$	$\|x_i - 38\|$	
44	39	6	cluster 2
28	23	10	cluster 2
48	43	10	cluster 2
26	21	12	cluster 2
32	27	6	cluster 2
14	9	24	cluster 1
52	47	14	cluster 2
50	45	12	cluster 2

New $\mu_1 = 32$

New $\mu_2 = \frac{44+28+48+26+32+52+50}{7}$

$= \frac{280}{7} = 40$

cluster 1 = {14}

cluster 2 = {44, 28, 48, 26, 32, 52, 50}

Good Write

$SSE = 6^2 + 10^2 + 10^2 + 12^2 + 6^2 + 0^2 + 14^2 + 12^2 = 837$

Iteration 2:

x_i	$\mu_1 = 32$ $\lvert x_i - 32 \rvert$	$\mu_2 = 40$ $\lvert x_i - 40 \rvert$	cluster
44	12	4	cluster 2
28	4	12	cluster 1
48	16	8	cluster 2
26	6	14	cluster 1
32	0	8	cluster 1
14	18	26	cluster 1
52	20	12	cluster 2
50	18	10	cluster 2

$$\text{New } \mu_1 = \frac{28+26+32+14}{4} = \frac{100}{4} = 25$$

$$\text{New } \mu_2 = \frac{44+48+52+50}{4} = \frac{194}{4} = 48.5$$

$$SSE = 4^2 + 4^2 + 8^2 + 6^2 + 18^2 + 18^2 + 10^2 = 700$$

Iteration 3:

x_i	$\mu_1 = 25$ $\lvert x_i - 25 \rvert$	$\mu_2 = 48.5$ $\lvert x_i - 48.5 \rvert$	cluster
44	19	4.5	2
28	3	20.5	1
48	23	0.5	2
26	1	22.5	1
32	7	16.5	1
14	11	34.5	1
52	27	3.5	2
50	25	1.5	2

Good Write

$$\text{New } \mu_1 = \frac{28+26+32+14}{4} = 25$$

$$\text{New } \mu_2 = \frac{44+48+52+50}{4} = 48.5$$

As values of centroids in Iteration 3 & 2 are same, Algo ends here.

Clusters Obtained

cluster 1 = {28, 26, 32, 14}

cluster 2 = {44, 48, 52, 50}

$$SSE = 4.5^2 + 3^2 + 0.5^2 + 1^2 + 7^2 + 11^2 + 3.5^2 + 1.5^2$$

$$= 214.25$$

5. Consider the dataset given below

Age	Income	Employed	Credit_rating	Buys car
young	high	yes	fair	yes
young	high	no	good	no
middle	high	no	fair	yes
old	medium	no	fair	yes
old	medium	no	fair	yes
old	low	yes	good	no
old	medium	no	good	no
middle	high	yes	fair	yes

Good Write

Compute all class conditional and class prior Probability. Use Naive Bayes classifier to predict the class of the following tuple:

$X = (age = young, income = medium, employed = yes, credit\_rating = good)$

For each class y_j, the class conditional probability of attribute x_i is -

$$P(X_i = x_i \mid Y = y_j) = \frac{1}{\sqrt{2\pi}\,\sigma_{ij}} \exp^{-\left(\frac{(x_i - \mu_{ij})^2}{2\sigma_{ij}}\right)}$$

X_1 = Age, X_2 = Income, X_3 = employed, X_4 = credit_rating

Calculating Probability :-

$P(C_1) = P(Buys\ car = Yes) = \frac{5}{8} = 0.625$

$P(C_2) = P(Buys\ car = No) = \frac{3}{8} = 0.375$

Calculating class conditional Probability :-

$P(Age = young \mid Buys\ car = No) = \frac{1}{3} = 0.33$

$P(Age = middle \mid Buys\ car = No) = 0$

$P(Age = old \mid Buys\ car = No) = \frac{2}{3} = 0.66$

Good Write

$P(Age = young \mid Buys\ car = Yes) = \frac{1}{5} = 0.2$

$P(Age = middle \mid Buys\ car = Yes) = \frac{2}{5} = 0.4$

$P(Age = old \mid Buys\ car = Yes) = \frac{2}{5} = 0.4$

$P(Income = high \mid Buys\ car = No) = \frac{1}{3} = 0.33$

$P(Income = medium \mid Buys\ car = No) = \frac{1}{3} = 0.33$

$P(Income = low \mid Buys\ car = No) = \frac{1}{3} = 0.33$

$P(Income = high \mid Buys\ car = Yes) = \frac{3}{5} = 0.6$

$P(Income = \overset{medium}{\cancel{high}} \mid Buys\ car = Yes) = \frac{2}{5} = 0.4$

$P(Income = low \mid Buys\ car = Yes) = 0$

$P(Employed = Yes \mid Buys\ car = Yes) = \frac{2}{5} = 0.4$

$P(Employed = No \mid Buys\ car = No) = \frac{2}{3} = 0.66$

$P(Employed = Yes \mid Buys\ car = No) = \frac{1}{3} = 0.33$

$P(Employed = No \mid Buys\ car = Yes) = \frac{3}{5} = 0.6$

$P(Credit\ Rating = fair \mid Buys\ car = No) = 0$

$P(Credit\ Rating = fair \mid Buys\ car = Yes) = \frac{5}{5} = 1$

$P(Credit\ Rating = good \mid Buys\ car = No) = \frac{3}{3} = 1$

Good Write $P(Credit\ Rating = good \mid Buys\ car = Yes) = 0$

~~Calculating~~ class conditional probabilities:-

~~P(x/Buys car = Yes) = P(age = old / Buys car = Yes) ×~~

P(

Using Naive Bayes classifier :-

$P(x|No) = P(age = young | Buys car = No) \times$

$P(income = medium | Buys car = No) \times$

$P(employed = yes | Buys car = No) \times$

$P(credit\ Rating = good | Buys car = No)$

$= \frac{2}{6} \times \frac{1}{3} \times \frac{1}{3} \times \frac{4}{5} = 0.028$

$P(x|Yes) = \frac{1}{5} \times \frac{2}{5} \times \frac{2}{5} \times \frac{1}{7}$

$= 0.0044.$

$P(x|No) > P(x|Yes)$

So class predicted = No

Good Write

6. Consider the market basket transaction shown in the following table. Use Apriori algo to answer the questions that follows.

TID	Item Bought
1	Oregano, chocolate, milk, cheese, french fries
2	milk, french fries, cheese, ketchup
3	chocolate, cheese, oregano, ketchup
4	chocolate, cheese, french fries
5	french fries, cheese, oregano, chocolate
6	chocolate, ketchup
7	oregano, french fries, ketchup
8	oregano, french fries, chocolate
9	ketchup, oregano, milk
10	french fries, chocolate

- Assuming the minimum support threshold is fixed at 40%, list the set of frequent 1-itemset (L_1) and with their respective supports.
- List the itemsets in the set of candidate 2-itemsets (C_2) and calculate their supports.
- Generate all association rules from the itemsets in L_2 and also compute the confidence of these rules.

- Support for 1-itemsets

$s(Oregano) = \frac{6}{10} \times 100 = 60\%$

$s(chocolate) = \frac{6}{10} \times 100 = 60\%$

Good Write

$s(\text{milk}) = \frac{3}{10} \times 100 = 30\%$

$s(\text{cheese}) = \frac{5}{10} \times 100 = 50\%$

$s(\text{french fries}) = \frac{7}{10} \times 100 = 70\%$

$s(\text{ketchup}) = \frac{5}{10} \times 100 = 50\%$

minsup = 40%.

∴ milk will be pruned.

∴ frequent 1-itemset :-

$L_1 = \{\{\text{oregano}\}, \{\text{chocolate}\}, \{\text{cheese}\}, \{\text{french fries}\}, \{\text{ketchup}\}\}$

- Candidate 2 itemsets.

Minsup = 40%.

2 itemsets candidates :-

~~No. of candidate class = $2^k - 2 = 2^6 - 2 = 64 - 2 = 62$~~

	support
Oregano	60%
Chocolate	60%
Milk	30% pruned
Cheese	50%
french fries	70%
ketchup	50%

1. $s(\{\text{oregano, chocolate}\}) = \frac{4}{10} \times 100 = 40\%$

2. $s(\{\text{oregano, cheese}\}) = \frac{3}{10} \times 100 = 30\%$ - pruned

3. $s(\{\text{oregano, french fries}\}) = \frac{4}{10} \times 100 = 40\%$

4. $s(\{\text{oregano, ketchup}\}) = \frac{3}{10} \times 100 = 30\%$ - pruned

5. $s(\{\text{chocolate, cheese}\}) = \frac{4}{10} \times 100 = 40\%$

6. $s(\{\text{chocolate, french fries}\}) = \frac{5}{10} \times 100 = 50\%$

7. $s(\{\text{chocolate, ketchup}\}) = \frac{2}{10} \times 100 = 20\%$ - pruned

8. $s(\{\text{cheese, french fries}\}) = \frac{4}{10} \times 100 = 40\%$

9. $s(\{\text{cheese, ketchup}\}) = \frac{2}{10} \times 100 = 20\%$ - pruned

10. $s(\{\text{french fries, ketchup}\}) = \frac{2}{10} \times 100 = 20\%$ - pruned

∴ 2 itemset candidates are :-

$C_2 = \{\{\text{oregano, chocolate}\}, \{\text{oregano, french fries}\}, \{\text{chocolate, cheese}\}, \{\text{chocolate, french fries}\}, \{\text{cheese, french fries}\}\}$

- Association Rules :-

a) R_1 : {oregano} → {chocolate, cheese, french fries, ketchup}

$c(R_1) = 0$

R_2: {oregano, chocolate} → {cheese, frenchfries, ketchup}

$c(R_2)=0$

R_3: {oregano, chocolate, cheese} → {frenchfries, ketchup}

$c(R_3)=0$

R_4: {oregano, chocolate, cheese, frenchfries} → {ketchup}

$c(R_4)=0$

R_5: {oregano, cheese} → {chocolate, frenchfries, ketchup}

$c(R_5)=0$

R_6: {oregano, frenchfries} → {chocolate, cheese, ketchup}

$c(R_6)=0$

R_7: {oregano, ketchup} → {chocolate, cheese, frenchfries}

$c(R_7)=0$

R_8: {chocolate} → {oregano, cheese, ketchup, frenchfries}

$c(R_8)=0$

R_9: {chocolate, cheese} → {oregano, ketchup, frenchfries}

$c(R_9)=0$

R_{10}: {chocolate, ketchup} → {oregano, cheese, frenchfries}

$c(R_{10})=0$

R_{11}: {chocolate, frenchfries} → {oregano, cheese, ketchup}

$c(R_{11})=0$

R_{12}: {chocolate, frenchfries, oregano} → {cheese, ketchup}

$c(R_{12})=0$.

2022 QP 1174 A

1. (a) How are accuracy rate and error calculated for evaluation of a classification model?

~~Accuracy~~ We evaluate accuracy rate and error of a classification model using a confusion matrix. Confusion matrix is a summarized table used to assess the performance of a classification model.

~~Accuracy~~ - no. of correct predictions made
Accuracy Rate - Percentage of correct predictions made for a given data set.

Error = 1 - accuracy.

(b) Briefly describe the aggregation technique in data preprocessing?

Aggregation in is the process of finding, collecting, and presenting the data in a summarized format to perform statistical analysis of business schemes or analysis of human patterns.
Data aggregation is a need when a dataset as a whole is useless info and cannot be used for analysis. So, the datasets are summarized into useful aggregates to acquire ~~else~~ desirable results and also to enhance the user experience on the application itself. They provide aggregate measurements such as sum, count and ~~among~~ average.

(c) Normalize the age of four students, given by the values {18, 21, 22, 25}.

Done in 2018 Q 1.(j)

(d) Explain briefly the significance of dimensionality reduction.

- It reduces the time and storage space required.
- It becomes easier to visualize the data when reduced to very low dimensions such as 2D or 3D.
- It avoids the curse of dimensionality.
- It removes irrelevant features from the data. Because having irrelevant features in the data can decrease the accuracy of the models and make your model learn based on irrelevant features.

(e) What is an outlier in context of a dataset?

Outliers are either:-
i) data objects that, in some sense, have characteristics that are different from most of the other data objects in the data set, or
ii) values of an attribute that are unusual with respect to the typical values for that attribute.

(f) What kind of Association Rules do you think would be stronger and more interesting — the rules with high support and low confidence or the rules with low support and high confidence? Why?

Support is similar to Recall metric, a rule with high ~~m~~ support means it has high presence on the dataset.

Confidence is similar to Precision metric, a rule with high confidence means it has high precision whenever the rule appears.

While mining association rules, you ~~pre~~ prefer the ones with high confidence (precision) over ones with high support (recall). That is why usually the min support may be on the lower side, even lower than 50%, while min Confidence is usually set higher, above 50% for example.

(g) Define the use of sampling in data mining? Name two sampling methods.

Sampling is a commonly used approach for selecting a subset of the data objects to be analyzed. Data miners sample b/c it is too expensive or time consuming to process all the data. In some cases, using a sampling algorithm can reduce the data size to the point where a

better, but more expensive algorithm can be used.

Sampling Methods:-
i) Simple random sampling
ii) Progressive Sampling

(h) What are the three factors that affect the computational complexity of Apriori ~~Principle~~ algorithm?

The computational complexity of the Apriori algo can be affected by the following factors:-

i) Support Threshold - Lowering the support threshold often results in more items being declared as frequent. This has an adverse effect on the computational complexity of the algo b/c more ~~or~~ candidate itemsets must be generated and counted.

ii) Number of Items (Dimensionality) - As the no. of items ↑, more space ~~it~~ will be needed to store the support counts of items. If the no. of frequent items also grow with the dimensionality of the data, the computation and I/O costs will ↑ b/c of the larger no. of candidate itemsets generated by the algo.

iii) Number of Transactions - Since the Apriori algo makes repeated pass over the data set, its run time increases with a large no of transactions.

(i) Distinguish b/w the following :-

(i) Exclusive v/s fuzzy clustering

Exclusive Clustering	Fuzzy Clustering
• Exclusive clustering is the hard clustering in which data pnt exclusively belongs to one cluster.	• Fuzzy clustering is a clustering method where data pnts can belong to 1 or more than one groups (clusters).
• Eg :- K-means clustering	• Eg :- fuzzy c-means

(ii) Complete vs Partial clustering

Done in 2018 Q 7 (b)

(j) What do you understand by the term missing data in data mining ? Briefly describe two methods for dealing with missing data.

A missing value can signify a no. of different things. ~~Perpha~~ Perhaps the field was not applicable, the event did not happen, or the data was not available. It could be that the person who entered the data did not know the right value, or did not care if a field was not filled in.

Done in 2019. 1 (d)

(k) Define the term scalability and heterogeneity? What challenges do they pose while mining the data?

Scalability - The measure of a system's ability to increase or decrease in performance and cost in response to changes in application and system processing demands.
May also require the implementation of novel data structure to access individual records in an efficient manner.

Heterogeneity - refers to mining different sets of attributes in each participating node.
Pulling source data can be complicated of time-consuming if data sources have different formats, structures and types.
Leads to the influx of data from diverse source into a unified system which can ultimately lead to exponential growth in data volume.

(l) Describe precision and recall metrics used for classification.

Precision - The ability of a classification model to identify only the relevant data pnts. Mathematically, precision is the no. of true +ve divided by the no. of true +ve plus no. of false +ve.

Recall - the ability of a model to find all the relevant cases within a data set. Mathematically, we define recall as the no. of true +ve divided by the no. of

true +ves plus the no. of false +ves.

2. (a) Explain discretization and binarization in context of data preprocessing.

Discretization –

When we have a wide diversity of possible data ranges it becomes difficult to classify data which are needed to be in the form of categorical attributes. Thus it becomes necessary to transform a continuous attribute to a categorical attribute. This concept of transforming the attribute is known as discretization.

Binarization –

Binarization is process that is used to transform data features of any entity into binary nos. It is done to ~~clarify~~ classify algorithms more efficiently. To convert into binary, we can use transform data using binary threshold.

(b) Consider a categorical attribute Customer satisfaction {unsatisfactory, poor, neutral, good, very good}

i) Convert to 3 binary attributes.

Categorical attributes	Integer val.	x_0	x_1	x_2
unsatisfactory	0	0	0	0
poor	1	0	0	1
neutral	2	0	1	0
good	3	0	1	1
very good	4	1	0	0

ii) Convert to 5 asymmetric binary attributes.

Categorical attributes	Integer values	x_0	x_1	x_2	x_3	x_4
Unsatisfactory	0	1	0	0	0	0
poor	1	0	1	0	0	0
neutral	2	0	0	1	0	0
good	3	0	0	0	1	0
very good	4	0	0	0	0	1

(c) State Apriori Principle.

Done in 2018. Q1. (i)

3. For the given employee table, identify the type of each attribute (nominal, ordinal, interval-scaled, ratio-scaled), giving justification for your choice. For each attribute that has missing values, briefly state how will you handle missing values therein.

Emp_id – Nominal, since they only provide enough info to distinguish one object from another (=, ≠).

Gender – Nominal, since they only provide enough info to distinguish one obj from another. Here an obj can either be male / female.

Age – Ratio, since both difference ~~are~~ b/w values and ratios are meaningful.

Home_pincode – Nominal, since they only provide enough info to distinguish one obj from another.

Date of joining - Interval, The gap b/w two dates is meaningful and diff. gaps are comparable. We can compare 2 date to see how has joined first.

Design - Ordinal, Since we can put & arrange obj's in order.

Contact_No - Nominal, since contact No. are distinguishable & cannot be put into order.

Email_id - Nominal, Since they only have distinguishable name values and same do not provide enough info to put into order.

Emp_id	Gender	Age	Home_pincode	Date of joining	Desig.	contact No	Email id
1001	M	28	232372	16/4/10	Captain	9818267x	b@gma.com
1002	F	31	222321	21/3/11	Captain	98112[illegible]72	f@gma.com
1003	F	34	243431	23/4/08	Major	992[illegible]7	??
1004	M	??	238432	21/5/09	Captain	9[illegible]10	x@gma.com
1005	M	35	454656	13/4/07	Colonel	981[illegible]	d@gma.com
1006	??	36	465643	04/5/05	Colonel	7[illegible]	e@gma.com
1007	F	30	234123	09/7/12	Captain	88567[illegible]	??
1008	M	32	676678	18/7/16	Major	??	x@gma.com
1009	M	33	565768	24/6/11	Colonel	98[illegible]90	e@gma.com
1010	M	30	498976	05/9/12	Major	??	d@gma.com

- To handle missing values for gender column, we can assign the most commonly occuring ~~to~~ category to the missing value.
- To handle missing value for age column we can find the average of all values & assign it the avg to it.
- To handle missing value for Contact_No and Email_id column where is no way to estimate their value neither can we eliminate them so we just ignore them during analysis.

4.(a) Consider the following dataset where each data object has a class label along with five features associated with it.

	Class	Cap Shape	Bruises	Odour	Stalk Shape	Habitat
1.	Edible	flat	Yes	anise	Tapering	grasses
2.	Poisonous	Convex	Yes	pungent	enlargening	grasses
3.	Edible	Convex	Yes	almond	enalargering	grasses
4.	Edible	Convex	Yes	almond	Tapering	meadows
5.	Edible	flat	Yes	anise	enlargening	woods
6.	Edible	flat	NO	none	enlargening	urban
7.	Poisonous	conical	Yes	pungent	enlargening	urban
8.	Edible	flat	Yes	anise	enlargening	meadows
9.	Poisonous	Convex	Yes	pungent	enlargening	urban

Consider the following pair of rules:

- (Odour = pungent) and (habitat = urban) → (class = poisonous) R1
- (Bruises = yes) → (class = edible) R2

(i) Are the two rules mutually exclusive? Justify.

(ii) Calculate coverage and accuracy for each of the rules.

(i) Yes the rules are mutually exclusive since no record is triggering both R1 & R2.

(ii) ~~R1 R2~~

Coverage for a rule $r: A \rightarrow y$, for given dataset D

$$\text{Coverage}(r) = \frac{|A|}{|D|}$$

$$\text{Accuracy}(r) = \frac{|A \cap y|}{|A|}$$

* Rule R1:

$$\text{Coverage}(R_1) = \frac{2}{9}$$

$$\text{Accuracy}(R_1) = \frac{2}{2} = 1$$

* Rule R2:

$$\text{Coverage}(R_2) = \frac{8}{9}$$

$$\text{Accuracy}(R_2) = \frac{5}{9}$$

(b) Consider the one-dimensional labeled data set given below:

x:	0.5	3.0	4.5	4.6	4.9	5.2	5.3	5.5	7.0	9.5
Y:	-	-	+	+	-	-	+	+	-	-

Classify the data pnt $x = 4.0$ acc. to the 5 NN, using the majority voting scheme.

$$\text{Distance} = |x_i - x|$$

$x = 4.0$

x_i	$\lvert x_i - 4.0 \rvert$	Y_i	NN
0.5	3.5	-	
3.0	1	-	5NN
4.5	0.5	+	5NN
4.6	0.6	+	5NN
4.9	0.9	-	5NN
5.2	1.2	-	5NN
5.3	1.3	+	
5.5	1.5	+	
7.0	3	-	
9.5	5.5	-	

5 NN of $x = 4.0$ are :-

$\{3.0, 4.5, 4.6, 4.9, 5.2\}$

ie. 3 -ve and 2 +ve

so $x = 4.0$ will have class -ve.

5. (a) What are the three conditions needed to be satisfied by a distance measure, so that it can be established as a distance metric?

~~Done in 2019 2.(b)~~

Measures that needs to be satisfied are :-

1) Positivity

$d(x,y) \geq 0$ } $\forall x, y$

$d(x,y) = 0$ } $\forall x = y$

2) Symmetric

$d(x,y) = d(y,x)$, $\forall x, y$

3) Triangle Inequality

$d(x,z) \leq d(x,y) + d(y,z)$, $\forall x, y, z$

(b) Show whether Euclidean Distance, used for finding distance between two data objects $o_1(x_1, y_1)$ and $o_2(x_2, y_2)$, can be treated as a distance metric.

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

let us consider 3 objects

$o_1(2,5)$, $o_2(4,4)$, $o_3(2,1)$

1. Positivity / Symmetric

$d(o_1, o_2) = \sqrt{(2-4)^2 + (5-4)^2}$

$= \sqrt{4+1} = \sqrt{5}$

$d_1(o_2, o_1) = \sqrt{2(4-2)^2 + (4-5)^2}$

$= \sqrt{2+1} = \sqrt{5}$

$\because d(o_1, o_2) = d(o_2, o_1)$

Symmetry satisfied

2. Positivity

$d_2(o_1, o_2) = \sqrt{(2-4)^2 + (5-4)^2} = \sqrt{5} \geq 0$

take ~~any~~ $o_2(2,5) = o_1$

$d(o_1, o_2) = \sqrt{(2-2)^2 + (5-5)^2} = \sqrt{0-0} = 0$

$\because d(o_1, o_2) \geq 0$, $\forall o_1, o_2$

$d(o_1, o_2) = 0$, $\forall o_1 = o_2$

Positivity satisfied.

3. Triangle Inequality

$d(o_1, o_3) = \sqrt{(2-2)^2 + (5-1)^2} = \sqrt{0+16} = 4$

$d(o_1, o_2) = \sqrt{5}$

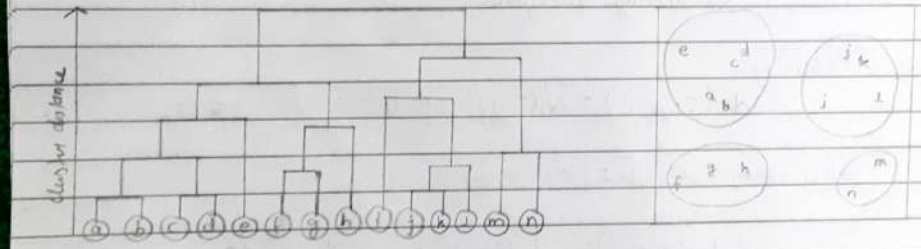
$d(O_2, O_3) = \sqrt{(4-2)^2 + (4-1)^2}$
$= \sqrt{4+9} = \sqrt{13}$

$d(O_1, O_3) \leq$
$d(O_1, O_2) + d(O_2, O_3) = \sqrt{5} + \sqrt{13} = 5.84.$

$\because d(O_1, O_3) < d(O_1, O_2) + d(O_2, O_3), \forall O_1, O_2, O_3$
Hence Triangle Inequality satisfied

Hence Euclidean Distance is a metric.

(c) With the help of a diagram, explain the usage of a dendogram.



A dendogram is a tree-like structure that explains the relationship b/w all the data pts in the system. However, like a regular family tree, a dendogram need not branch out at regular intervals from top to bottom as the vertical direction in it represents the distance b/w clusters in some metric.

6. Consider a transaction db D, consisting of nine transactions. Suppose the min sup is set at 45% & min conf is set at 70%, show clearly the steps for finding out frequent itemsets of all sizes using the Apriori algo. Also generate the strong association rules from frequent itemsets of size 3.

TID	List of Items
T1	A, B, C, F
T2	B, D
T3	B, C
T4	A, B, C
T5	A, C, F
T6	B, C, F
T7	A, D
T8	A, B, C, E, F
T9	A, B, C

$\text{supp} = \frac{\sigma}{|T|}$

$\text{con } X \rightarrow Y$
$\text{confidence} = \frac{\sigma(X \cup Y)}{\sigma(X)}$

Min Sup = 45%, Min Conf = 70%.

Itemsets (1-Itemsets)

Item	Count	Support	Confidence
A	6	66.6%	Frequent
B	7	77.7%	Frequent
C	7	77.7%	Frequent
Pruned: D	2	22.2%	Infrequent
Pruned: E	1	11.1%	Infrequent
Pruned: F	4	44.4%	Infrequent

2-Itemset

	Items	Count	Support	
Pruned	A, B	4	44.4%	Infrequent
	A, C	5	55.5%	Frequent
	B, C	6	66.6%	Frequent

3 itemset

Items	Count	Support	
Removed - A, B C	4	44.4%	Unfrequent

Frequent Itemsets of all sizes:-

1. A
2. B
3. C
4. A,C
5. B,C

there are no frequent itemsets of size 3 therefore we cannot find any strong association rules.

7. Consider a dataset of images of dogs and cats. Suppose there are 500 images of dogs and cats each. The classification model predicts 340 correct images of dogs and 410 correct images of cat. Perform the operation that follow!-

a) Draw the confusion matrix for this problem.

b) Compute the classifier ~~matrix for this problem~~ accuracy, error and sensitivity.

Total Images of dogs = 500
Total Images of cats = 500
Correct prediction of dogs = 340
Correct predictions of cats = 410

Good Write

a)

		Actual class	
		Dogs	Cats
Predicted class	Dogs	340	90
	Cats	160	410

b) Accuracy = $\frac{TP + TN}{Total}$

$= \frac{340 + 410}{1000}$

$= \frac{750}{1000} = 0.75$

Error = 1 - accuracy
= 1 - 0.75
= 0.25

Sensitivity (Dog) = $\frac{TP}{TP + FN}$

$= \frac{340}{340 + 160}$

$= 0.68$

Sensitivity (Cats) = $\frac{TP}{TP + FN}$

$= \frac{410}{410 + 90}$

$= 0.82$.

8. Given the following data pnts: 4, 9, 18, 13, 11, 2, 6, 25, k=3 & initial centroids $\mu_1 = 5$, $\mu_2 = 10$ & $\mu_3 = 15$. Show clearly the clusters & new cluster centers obtained after each iteration of K-means algo for 2 iterations.

Good Write

x_i	$\mu_1=5$ $\lvert x-5 \rvert$	$\mu_2=10$ $\lvert x_i-10 \rvert$	$\mu_3=15$ $\lvert x_i-15 \rvert$	cluster
4	1	6	11	1
9	4	1	6	2
18	13	8	3	3
13	8	3	2	3
11	6	1	4	2
2	3	8	13	1
6	1	4	9	1
25	20	15	10	3

cluster 1 = { 4, 2, 6 }

New $\mu_1 = \frac{4+2+6}{3} = 4$

cluster 2 = { 9, 11 }

New $\mu_2 = \frac{9+11}{2} = 10$

Cluster 3 = { 18, 13, 25 }

New $\mu_3 = \frac{18+13+25}{3} = 18.66$

x_i	$\mu_1=4$ $\lvert x_i-4 \rvert$	$\mu_2=10$ $\lvert x_i-10 \rvert$	$\mu_3=18.66$ $\lvert x_i-18.66 \rvert$	cluster
4	0	6	14.6	1
9	5	1	9.6	2
18	14	8	5.6	3
13	9	3	5.6	2
11	7	1	7.3	2
2	2	8	16.6	1
6	2	4	12.6	1
25	21	15	6.4	3

cluster 1 = { 4, 2, 6 }

New $\mu_1 = \frac{4+2+6}{3} = 4$

cluster 2 = { 9, 13, 11 }

New $\mu_2 = \frac{9+13+11}{3} = 11$

cluster 3 = { 25, 18 }

New $\mu_3 = \frac{25+18}{2} = 21.5$

DATE : __/__/____
PAGE : _____

2019

Q5.(b) let X denotes the categorical attributes having values {awful, poor, ok, good}. What is the representation of each value when X is converted to binary form using:
i) 2 bits ii) 4 bits?

i) categorical Values | Integer values | X_0 | X_1

awful	0	0	0
poor	1	0	1
ok	2	1	0
good	3	1	1

ii) categorical val. | Integer val. | X_0 | X_1 | X_2 | X_3

awful	0	1	0	0	0
poor	1	0	1	0	0
ok	2	0	0	1	0
good	3	0	0	0	1

Good Write