2018

1 (a) KDD - Knowledge discovery in database, It is a process of taking information out of an unknown database & comprises of following steps:-

1) Data selection :- We select the data from multiple sources & combine them if needed
2) Data preprocessing : It involves cleaning of data like handling missing values, duplicate data etc
3) Data transformation : We transform the data according to our needs
4) Data mining : Finding hidden patterns from the data which cannot be seen
5) Evaluating patterns.

Data mining is just a step involved in KDD

(b) (i) eye color - Nominal
(ii) grades - Ordinal
(iii) dates - interval
(iv) age - ratio

5 1-1

Camlin | Page
Date / /

(c) $0 \leq$ Gini index ≤ 1

** $1 - \left[\left(\frac{1}{6}\right)^2 + \left(\frac{5}{6}\right)^2\right]$

$= 1 - \left[\frac{1}{36} + \frac{25}{36}\right]$

$= \frac{10}{36} = \boxed{\frac{5}{18}}$

(d) ** Graph data structure

(e) p1 (0,2) p2 (2,0) p3 (3,1) p4 (5,1)

$p1p2 = \sqrt{(2)^2 + (2)^2} = 2\sqrt{2}$

$p3p4 = \sqrt{(2)^2 + (0)^2} = 2$

(f) ** Categorical attribute

(g)

Interval	Ratio
1) Operations like +/− can be performed	1) Operations like ×/÷ can also be performed
2) Values can be specified within a given range	2) Doesn't make sense when value is zero
eg - temp	eg - length, height, age

(h) Eager learner :- Here, a model is prepared first on the basis of the training set and then this model is tested on the test set. eg - decision tree, rule based classifier

Lazy learner - Rote based classifiers are lazy learners because they don't build a model rather they learn the whole training set & if records that matches train set comes, they are able to classify them. But if any record outside training set comes, they could not classify them as no info is available
eg - KNN

(i) Apriori principle : If an itemset is frequent, then all its subsets must also be frequent. On contrary, if an itemset is infrequent, all its supersets are also infrequent.

As we know that support count of an itemset never exceeds that of its subset, hence the given statement is true according to Apriori principle

(j) $\{18, 21, 22, 25\}$

0 – 1

new min = 0

new_max = 1

min = 18

max = 25

$x' = \frac{x - min}{max - min} (new\_max - new\_min) + new\_min$

(i) $x' = \frac{18 - 18}{25 - 18} (\quad) = 0$

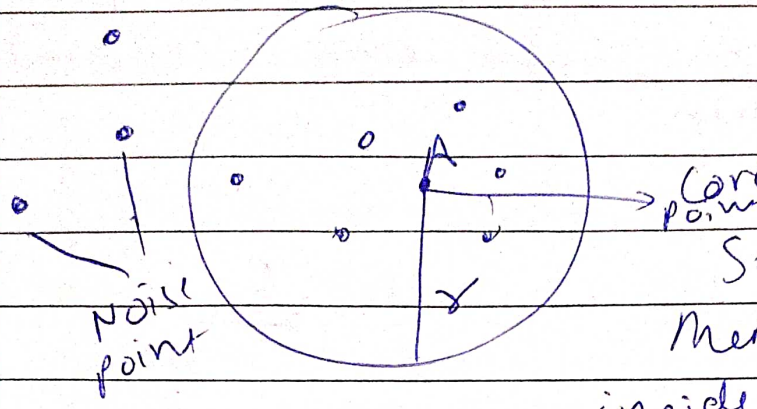
(ii) $x' = \frac{21 - 18}{25 - 18} (1) = \frac{3}{7}$

(iii) $x' = \frac{22 - 18}{25 - 18} = \frac{4}{7}$

(iv) $x' = \frac{25 - 18}{25 - 18} = 1$

(k) (i) Core point: When we take any point as the centre & draw a circle of radius r, if the no. of points inside the circle including that point exceeds the minimum value, we can term that point as core point.

(ii) Noise point: All the points which lie completely outside the circle we draw of radius r are termed as noise points



Given:
r, min = 7

So, we can see that there are 7 points inside the circle. Hence A is a core point

(i) Mutually exclusive:- If all the records in the database are triggered atmost one rule, then they are mutually exclusive rules. So, for mutual exclusiveness, two rules should not be triggered by one record.
If rules are not mutually exclusive, then we cannot determine to which class we should classify the record.
Such problem can be resolved by using two methods

1) Ordered method: Here, we give priority to every rule on basis of its accuracy & coverage and classify the record to the rule with higher priority

2) Unordered method: Here, we see the frequency of every class in dataset and classify the record to the class which is most frequent.

SECTION-B

(2) a) (i) $\{e\} = 8$

$\{b, d\} = 2$

$\{b, d, e\} = 2$

$\{a, b, d, e\} = 1$

(ii) $\{b, d\} \rightarrow \{e\}$

$$\text{Confidence} = \frac{\sigma(b \cup d \cup e)}{\sigma(b \cup d)} = \frac{2}{2} = 100\%$$

$\{e\} \rightarrow \{b, d\}$

$$\text{Confidence} = \frac{\sigma(e \cup b \cup d)}{\sigma(e)} = \frac{2}{8} = 25\%$$

(iii) No, confidence is not a symmetric measure as confidence for both the parts above is diff

Date / /

(b) ** metric

(3) (a) Aggregation :- Aggregation is a process where we combine multiple attributes into one for better analysis.
Example. When we are having elections in country, to calculate voter turnout by state, we sum up all the individual votes obtained in all the zones of the state and there is one total vote number for every state (total no. of votes)

Uses :-
1) Reduces the dimension of the dataset
2) Avoid analysis on unnecessary info. which is not needed
3) We can visualize the data better as we have reduced no. of attributes

(b)

Noise	Outliers
(i) It is just the unwanted data which occurs in a dataset	(i) The data points with characteristics diff from most of the datapoints in dataset

(i) No, noise is never interesting or desirable as it is just the unwanted data which needs to be removed

(ii) Yes, outliers can be desirable for example in fraud detection where there is a datapoint way different than most of the datapoints

(iii) No, noise objects can appear as normal data as well

(iv) No, outliers are not always noise because they can be useful sometimes & removing them will be ineffective

(4) (a) Confusion matrix

		Predicted class	
		Class = 1	Class = 0
Actual class	Class = 1	3 TP	3 FN
	Class = 0	1 FP	3 TN

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{3+3}{3+3+3+1} = \frac{6}{10} = 0.6$$

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

Date / /

b) k-fold cross validation

When the dataset is divided into k equal partitions, one of them is used as test set & other k-1 as training set.

Whole data can be used as test & training set.

Hold out method: The dataset is divided into two partitions- test set & training set. The proportion in which we divide can be 75-25, 50-50 etc. But, it can ignore some of the dataset for testing.

(5) (a) Partition based clustering

Makes use of centroid model in which data points are clustered on the basis of their distance from centroid

eg- K means

Hierarchical based clustering

Nesting clustering (A cluster is clustered into another cluster)

represented by dendograms

** ADV & DISADV

(b) Simple random sampling Randomly select the sample from dataset

(c) (i) Birds { R1 is triggered }

(ii) Mammals { R3 is triggered }

6 (a) ABE

A → B	AB → E	Conf = 2/3 = 66%
B → A	AE → B	Conf = 2/2 = 100%
A → E	BE → A	Conf = 2/4 = 50%
E → A	E → AB	Conf = 2/4 = 50%
B → E	B → AE	Conf = 2/5 = 40%
E → B	A → BE	Conf = 2/4 = 50%

(b) Strong association rules

AB → E

AE → B

BE → A

E → AB

A → BE

Camlin Page
Date / /

7) a) 2, 4, 10, 12, 3, 20, 30, 11, 25

$k = 3$

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

Points	Distance from μ_1	Distance from μ_2	Distance from μ_3	Cluster
2	0	2	4	C_1
4	2	0	2	C_2
10	8	6	4	C_3
12	10	8	6	C_3
3	1	1	3	C_1
20	18	16	14	C_3
30	28	26	24	C_3
11	9	7	5	C_3
25	23	21	19	C_3

$C_1 = \{2, 3\}$

$C_2 = \{4\}$

$C_3 = \{10, 12, 20, 30, 11, 25\}$

$\mu_1' = 2.5$

$\mu_2' = 4$

$\mu_3' = 18$

Points	$D(\mu_1)$	$D(\mu_2)$	$D(\mu_3)$	Clust
2	0.5	2	16	C_1
4	1.5	0	14	C_2
10	7.5	6	8	C_2
12	9.5	8	6	C_3
3	0.5	1	15	C_1
20	17.5	16	2	C_3
30	27.5	26	12	C_3
11	8.5	7	7	C_2
25	22.5	21	7	C_3

$C_1 = \{2, 3\}$

$C_2 = \{4, 10, 11\}$

$C_3 = \{12, 20, 30, 25\}$

$\mu_1' = 2.5 \approx 2$

$\mu_2' \approx 8$

$\mu_3' \approx 22$

Camlin Page
Date / /

	2	8	22	
Points	$D(\mu_1)$	$D(\mu_2)$	$D(\mu_3)$	Cluster
2	0	6	20	C_1
4	2	4	18	C_1
10	8	2	12	C_2
12	10	4	10	C_2
3	1	5	19	C_1
20	18	12	2	C_3
30	28	22	8	C_3
11	9	3	11	C_2
25	23	17	3	C_3

$C_1 = \{2, 4, 3\}$

$C_2 = \{10, 12, 11\}$

$C_3 = \{20, 30, 25\}$

$\mu_1' = 3$

$\mu_2' = 11$

$\mu_3' = 25$

	3	11	25	
Points	$D(\mu_1)$	$D(\mu_2)$	$D(\mu_3$	Cluster
2	1	9	23	C_1
4	1	7	21	C_1
10	7	1	15	C_2
12	9	1	13	C_2
3	0	8	22	C_1
20	17	9	5	C_3
30	27	19	5	C_3
11	8	0	14	C_2
25	22	14	0	C_3

$C_1 = \{2, 4, 3\}$

$C_2 = \{10, 12, 11\}$

$C_3 = \{20, 30, 25\}$

$\mu_1' = 3$

$\mu_2' = 11$

$\mu_3' = 25$

Now, as values are stagnant, we stop the process.