Data Mining

SAQ's :-

1Q) Give an example of Data Matrix and Dissimilarity Matrix?

Ans Data Matrix :- are used to build the evolutionary trees.

For ex :- Datamatrix code - encoding the text "wikipedia" use the color to show "datagrees", "padding yellow" "error correction red", finder & Timing Agenda "an unused orange".

⇒ Data matrix symbols are made up modules arrange within a perimeter finder & timing pattern.

* Dissimilarity matrix :- is a matrix that will express the similarity pair to pair btw two sets.
It is square and symmetric in nature.

⇒ The diagonal members are defined as "0" meaning that the 0 is measure of dissimilarity of between an element and itself.

2Q) Define Data Mining and its functionalities.

Ans Data Mining :- The process of automatically discovering patterns, relationships, and insights from large datasets, using various techniques and tools.

Ex :- Imagine a supermarket analyzing its sales data to identify patterns and relationships between products.

FUNCTIONALITIES :-

⇒ Data can be associated with classes and concept

⇒ Association Analysis.

⇒ classification and prediction of data
⇒ clustering analysis.
⇒ evolution and derivation analysis.

3Q) Enumerate different Technologies Used in Data Mining?

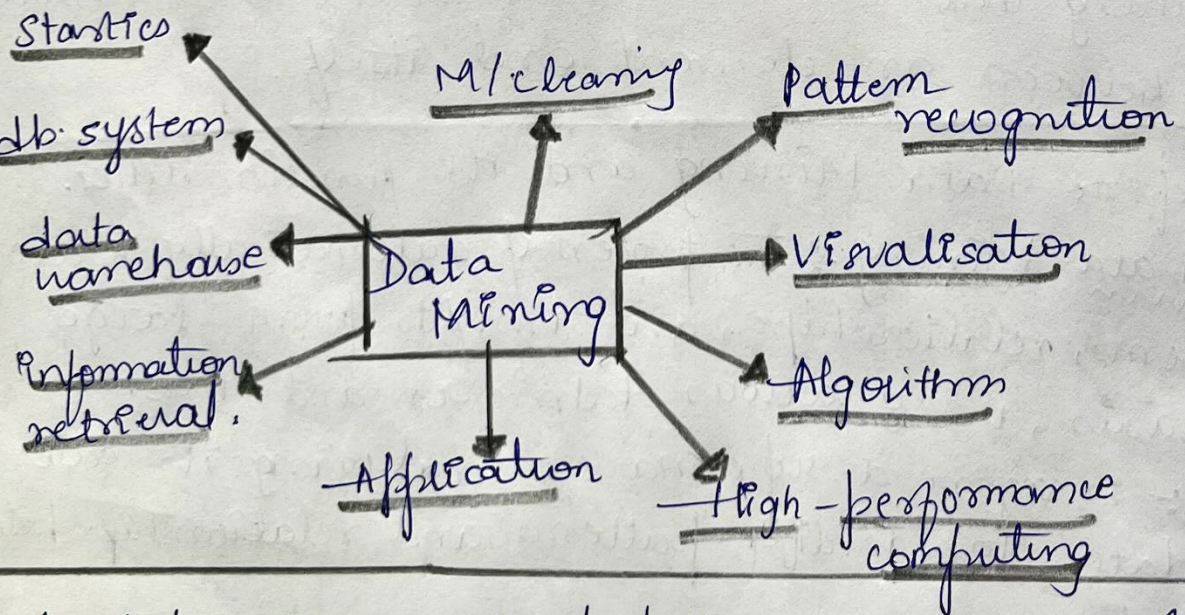
Ans Technology Used :-

1) ⇒ classification according to the kinds of database mined.
2) ⇒ classification according to the kinds of knowledge "
3) ⇒ classification according to the kind of technique utilize.

* Techniques A sophisticated data mining system will often adopt multiple data mining technique or workout an objective, integrated technique which combines the merits of a few individual approaches.

Techniques Used :-

- Statistics
- Machine learning
- Database system & Data warehouse
- Information Retrieval.



4Q) Calculate the mode of following data and its modality? 2, 4, 3, 8, 12, 14, 12, 13, 15, 20, 22.

Ans. MODE :- The most common / repeated (most) value in a dataset.

∴ Mode = 12. Modality :- unimodal.

LAQ's $\frac{0}{0}$

Sol) Suppose we have the following 2-D dataset. Consider the data as 2-data points. Given a new data point, $x = (1.4, 1.6)$ as a query, rank the database point based on similarity with the query using i) Euclidean Distance ii) Manhatton Distance iii) Supremum Distance & iv) Cosine similarity

	A_1	A_2
x_1	1.5	1.7
x_2	2	1.9
x_3	1.6	1.8
x_4	1.2	1.5
x_5	1.5	1.0

Sol. Considering the following data :-

A (x_1)	B (y_1)	Euclidean distance	Manhattan Distance	Minkowski distance	Supremum Distance	cosine similarity
1.5	1.7	0.1414	0.2	0.1414	0.1	0.99999
2	1.9	0.6708	0.9	0.6708	0.6	0.99575
1.6	1.8	0.2828	0.4	0.2828	0.2	0.99997
1.2	1.5	0.2236	0.3	0.2236	0.2	0.99903
1.5	1.0	0.6083	0.7	0.6083	0.6	0.96536

Let P_1 be a point with X_1, Y_1

P_2 be a point with X_2, Y_2

$X = (1.4, 1.6)$ (x_2, y_2)

Let $P_1 (X_1, Y_1)$

$P_2 (X_2, Y_2)$.

a) Euclidian distance :- $\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$

(ED)

* $(x_2, y_2) = (1.4, 1.6)$ — Given.

ED of $(1.5, 1.7) = \sqrt{(1.4-1.5)^2+(1.6-1.7)^2} = 0.1414$

ED of $(2, 1.9) = \sqrt{(1.4-2)^2+(1.6-1.9)^2} = 0.6708$

ED of $(1.6, 1.8) = \sqrt{(1.4-1.6)^2+(1.6-1.8)^2} = 0.2828$

ED of $(1.2, 1.5) = \sqrt{(1.4-1.2)^2+(1.6-1.5)^2} = 0.2236$

ED of $(1.5, 1.0) = \sqrt{(1.4-1.5)^2+(1.6-1.0)^2} = 0.6083$

b) Manhatton distance :-

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

MD of $(1.5, 1.7) = 0.2$

MD of $(2, 1.9) = 0.9$

MD of $(1.6, 1.8) = 0.4$

MD of $(1.2, 1.5) = 0.7$

MD of $(1.5, 1.0) = 0.7$

c) Minkowski distance :

$\Rightarrow \sqrt[h]{(x_2-x_1)^h + (y_2-y_1)^h}$ $\quad h \rightarrow$ height

(get cancel) $\Longrightarrow \sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$ $\quad x = (1.4, 1.6)$

$h = 2$.

$\Downarrow$

equals to euclidian distance column.

d) Supremum distance

$\Rightarrow \max(|x_2-x_1|, |(y_2-y_1|)$

SD of $(1.5, 1.7) = \max(|1.4-1.5|, |(1.6-1.7|) = 0.1$

SD of $(2, 1.9) = \max(|1.4-2.0|, |(1.6-1.9|) = 0.6$

SD of $(1.6, 1.8) = \max(|1.4-1.6|, |(1.6-1.8)|) = 0.2$

SD of $(2, 1.5) = \max(|1.4-1.2|, |(1.6-1.5)|) = 0.2$

SD of $(1.5, 1.0) = \max(|1.4-1.5)|, |(1.6-1.0)|) = 0.6$

e) cosine similarity :-

$$\frac{A \cdot B}{\|A\| \cdot \|B\|} = \frac{\Sigma A_i B_i}{\Sigma A_i^2 \Sigma B_i^2}$$

$(x_2, y_2) = (1.4, 1.6)$ //

$(\because \Sigma A_i \quad \Sigma B_i \quad)$.

i) $$\frac{(1.4*1.5) + (1.6 * 1.7)}{\sqrt{(1.5)^2+(1.7)^2} * \sqrt{(1.4)^2 + (1.6)^2}} = 0.999999$$

ii) $$\frac{(1.4 * 2.0) + (1.6 * 1.9)}{\sqrt{(2.0)^2+(1.9)^2} * \sqrt{(1.4)^2+(1.6)^2}} = 0.99575.$$

iii) $$\frac{(1.4 * 1.6) + (1.6 * 1.8)}{\sqrt{(1.6)^2+(1.8)^2} * \sqrt{(1.4)^2+(1.6)^2}} = 0.99997$$

iv) $$\frac{(1.4 * 2.0) + (1.6 * 1.5)}{\sqrt{(2.0)^2 + (1.5)^2} * \sqrt{(1.4)^2+(1.6)^2}} = 0.99903$$

v) $$\frac{(1.4 * 1.5) + (1.6 * 1.0)}{\sqrt{(1.5)^2+(1.0)^2} \sqrt{(1.4)^2+(1.6)^2}} = 0.96536$$

(6) 9) Elucidate about Decision tree Induction Algorithm with an example?

Ans DECISION TREE INDUCTION :-

- Flowchart like tree structure
- supports in taking decisions as it's classify the data.
- It defines the rules visually in form of tree.

Types of nodes

1) Root node - Main question
2) Branch node - Intermediate processing node
3) Leaf node - Answer

* Attribute Selection Measures :-

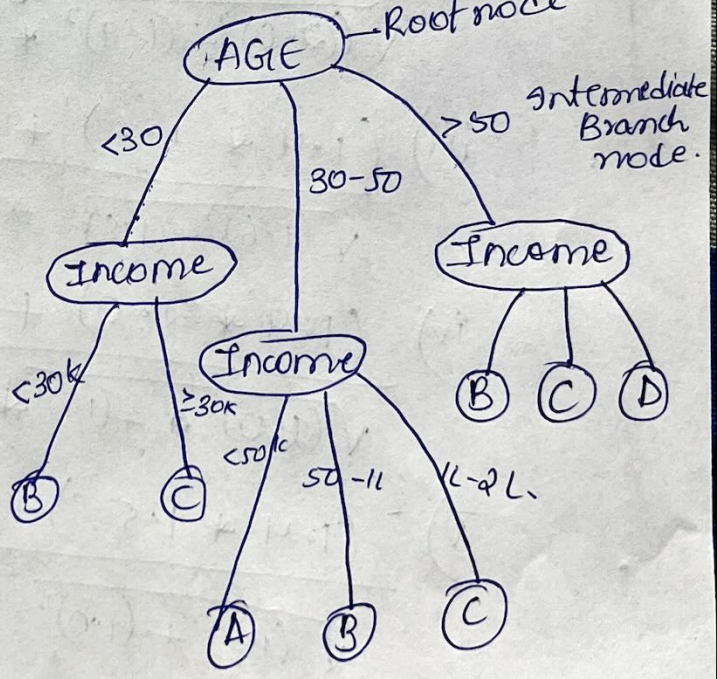
i) Information Gain :-
How much it does the answer to the specific question provide.
(info? prob that it may go wrong)

ii) Entropy :-
Measures the amount of uncertainity in the info

As IG ↑ Entropy ↓.

Ex :- Credit Score Rating

A - Avg
B - Bad
C - Good
D - excellent
} classification



* Rules that can be defined are :-

→ if age <30, income <30k
the credit score = Bad.

→ if age <30, income ≥ 30k
then credit score = Good.

Advantages of Decision Tree :-

→ Easy to understand and map nicely to a set of production rules.

→ Succesfully can be applied to a real problems.
→ No prior assumptions about the nature of the data.
→ Decision Tree are able to build the models with datasets containing numerical as well as categorical data
→ DT generation are of 2 types :-

i) Tree construction :-
- At start all the training examples are at the root
- Partitioning examples recursively based on selected attributes.

ii) Tree pruning :-
- Identify the root node branches that reflect noise and out layers.
- The use of this DT is it can clarify the unknown samples and attribute values can be tested from the sample against the DT.

---

7 Q) Discuss how to improve efficiency of Apriori Algorithm?

Ans ⇒ Improving the efficiency of Apriori :-
There are several variations for improving the efficiency of Apriori Algorithm.
→ Hash-based technique
→ Transaction reduction
→ Partitioning
→ sampling
→ Dynamic - itemset counting

1) Hash-based technique :-

- hashing of itemset into corresponding buckets.

→ tech is used to reduced the size of candidate k.

C_k for $k > 1$

eg :- create Hash table H_2 using hash function

$h(x,y) = (\text{order of } x) \times 10 + (\text{order of } y)$

th

bucket address	0	1	2	3	4	5	6
bucket count	2	2	4	2	2	4	4
bucket content	I_1, I_4 I_3, I_5	$I_1, 5$	⋮	⋮	⋮	⋮	⋮

2) Transaction Reduction :-

- reducing the no. of transactions scanned in future iterations.

- $A T^x$ that does not contain any frequent itemset cannot contain any freq $(k+1)$ itemsets

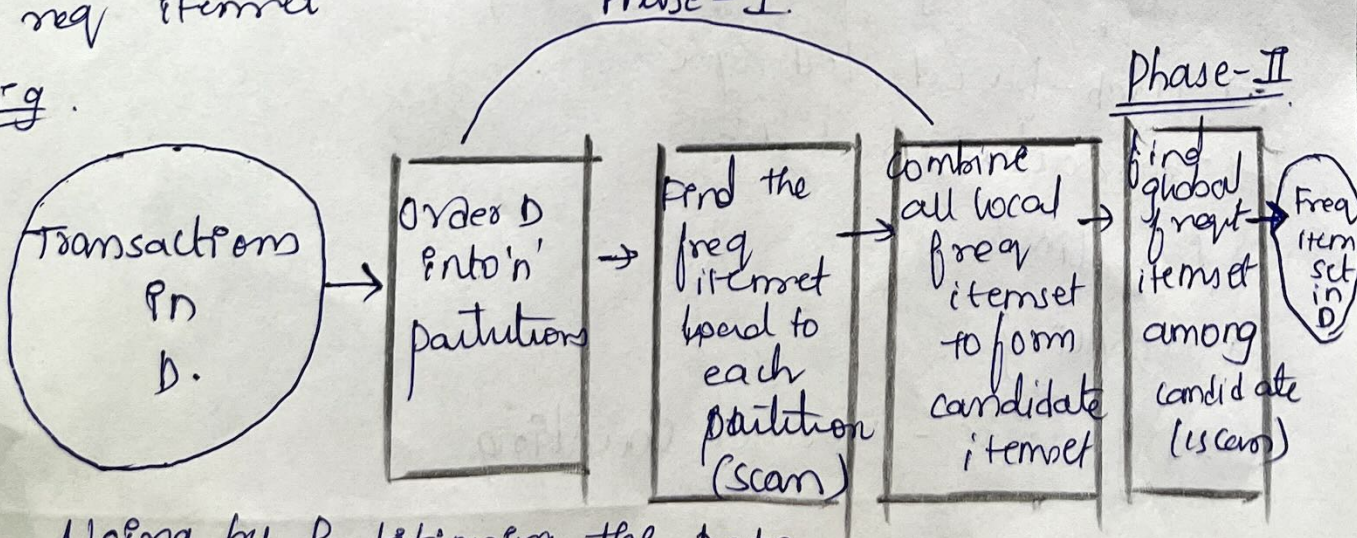
- Therefore such T^x can be removed $(j < k)$.

3) Partioning :-

Partition the data to find candidate itemsets.

- This technique requires 2 db to be scan to mine req itemset

fig.



Using by Partitioning the Data.

Phase I :-

Using Algor ÷ D ← 'n' → have non overlapy partitions → min_sup × no. of T^x in that partition

→ load freq itemset → itemset freq in that partition may or may not be freq wrt the entire db 'D'.

∴ Any itemset that is potentially freq wrt to D must occur as a freq itemset in atleast one of the partitions.

⇒ phase - II :-

this is the second scan of D is conducted in which actual support of each candidate is assessed to determined the global freq itemset.

- Partition size and no. of partitions are set so that each partition can fit into main memory & ∴ be read only once in each phase.

4) Sampling :- mining on a subset of given data.

5) Dynamic itemset counting :-

- adding candidate itemset at different points during a scan
- partitioning into blocks marking by a starting points

→ So this, new candidate itemsets can be added at any start point, unlike in Apriori, which determines new candidate itemset only immediately before each complete db scan.

⇒ This technique uses count so for as the lowerbound of actual count. and if the count passes min_support then it is added to freq itemset collection & can be used to generate longer candidates. This lead to fewer db scan than with Apriori for finding the freq itemsets.

8Q) explain how to mine closed frequent itemsets?

Ans ⇒ closed frequent itemset Mining (CFIM) is a technique used to discover sets of items that frequently appear together in a dataset.

MAIN STEPS :-

i) Frequent itemset Generation :-

→ Use Apriori algorithm or similar to generate frequent itemsets (FI)

Ex :- {A, B} (support = 4), {B, C} (support = 3), {A, C} (support = 2).

ii) closed itemset Generation :-

→ Use PRUNE to generate closed itemsets (CI) from FI

Ex :- {A, B, C} (support = 2) is a closed itemset.

iii) Item Merging :-

→ combine itemsets with the same support into a single itemset

Ex :- {A, B} and {B, C} are merged into {A, B, C} (support = 3).

iv) Subitemset after skipping pruning :-

→ Remove itemsets that are subsets of other itemsets with the same support.

Eg :- {A, B} is removed because its a subset of {A, B, C} with the same support.

v) CDFS (closed Frequent Itemset Dataset) Mining :-

- Output :- closed frequent itemsets (CFI) with their support values.
- Eg :- {A, B, C} (support = 2) is a closed frequent itemset.

Eg :- Dataset :-

Transaction ID	Items.
01	A, B, C
02	A, B, D
03	B, C, F
04	A, C, F
05	B, C, G

⇒) By applying CFIM to this dataset, we can discover closed frequent itemsets like {A, B, C} (support = 2), which indicates that items A, B, and C often appear together in transactions.

(a) (9) suppose that the data for analysis includes the attributes age (are in increasing order) then

22, 25, 17, 19, 33, 64, 23, 17, 20, 18.

Find 1) Mean 2) Median 3) Mode / Modality
4) 5 number summary 5) Box plot
6) Scatter plot between two variables.

Sol

Given data :-

22, 25, 17, 19, 33, 64, 23, 17, 20, 18.

Step 1 :- Arrange them in their increasing order

17, 17, 18, 19, 20, 22, 23, 25, 33, 64.

(i) Mean = $\frac{\text{Sum of observations}}{\text{total no. of observations}}$

$\bar{x} = \Sigma x / n$.

$\Rightarrow \frac{17+17+18+19+20+22+23+25+33+64}{10}$

$= \frac{228}{10} = 22.8$

$\therefore$ Mean $(\bar{x}) = 22.8$ //

(ii) Median

17, 17, (18), 19, 20 | 22, 23, (25), 33, 64.

middle terms.

Median $= \frac{20+22}{2} = \frac{42}{2} = \boxed{21 = Q_2}$

(iv) min $\Rightarrow 17$ $Q_1 \Rightarrow 18$

$Q_2 \Rightarrow 21$ $Q_3 \Rightarrow 25$ $\longrightarrow$ 5-Number Summary

max $= 33$.

(iii) Mode :-

The most common / most repeated value in a dataset.

$\therefore$ 17 has occured 2 time

$\therefore$ Mode = 17

Modality :- Unimodal.

check outlier.

1) $Q3 + (1.5 * IQR)$

Higher outlier

$IQR = Q3 - Q1$

$= 25 - 18 = 7$

$\Rightarrow 25 + (1.5 * 7)$

$= 25 + 10.5$

$= 35.5$.

lower outlier

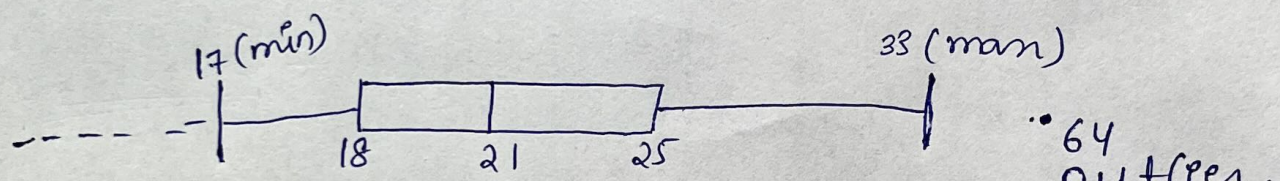
$Q1 - (1.5 * IQR)$

$\Rightarrow 18 - (1.5 * 7)$

$\Rightarrow 7.5$.

64 - outlier.

Higher outlier $\rightarrow$ 35.5.



Box plot.

vi) Scatter plot btw two variable.

~~Rollno~~ ~~marks~~

A B.