



शिक्षा मंत्रालय
MINISTRY OF
EDUCATION

INDIAN INSTITUTE OF TECHNOLOGY
JODHPUR



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING



P M R F

Prime Minister's Research Fellowship

Week 1 - Live Session

Data Mining

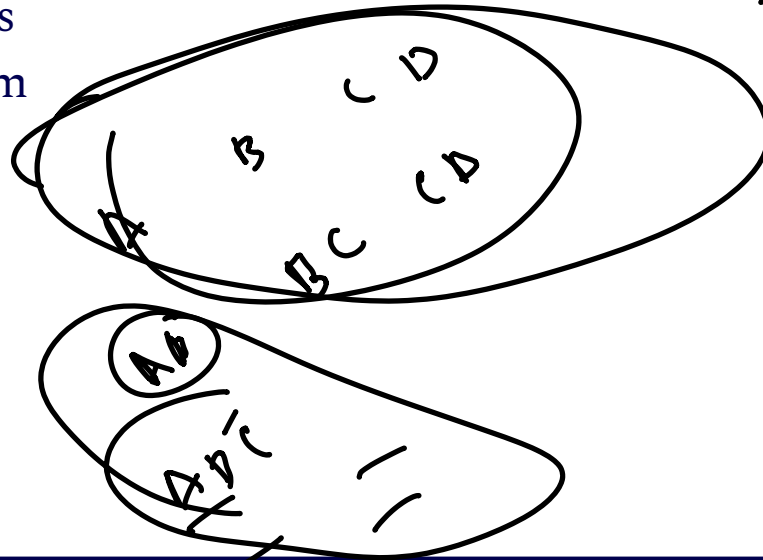
Swapnil S. Mane

mane.1@iitj.ac.in

PMRF Research Scholar

Summary Week 1

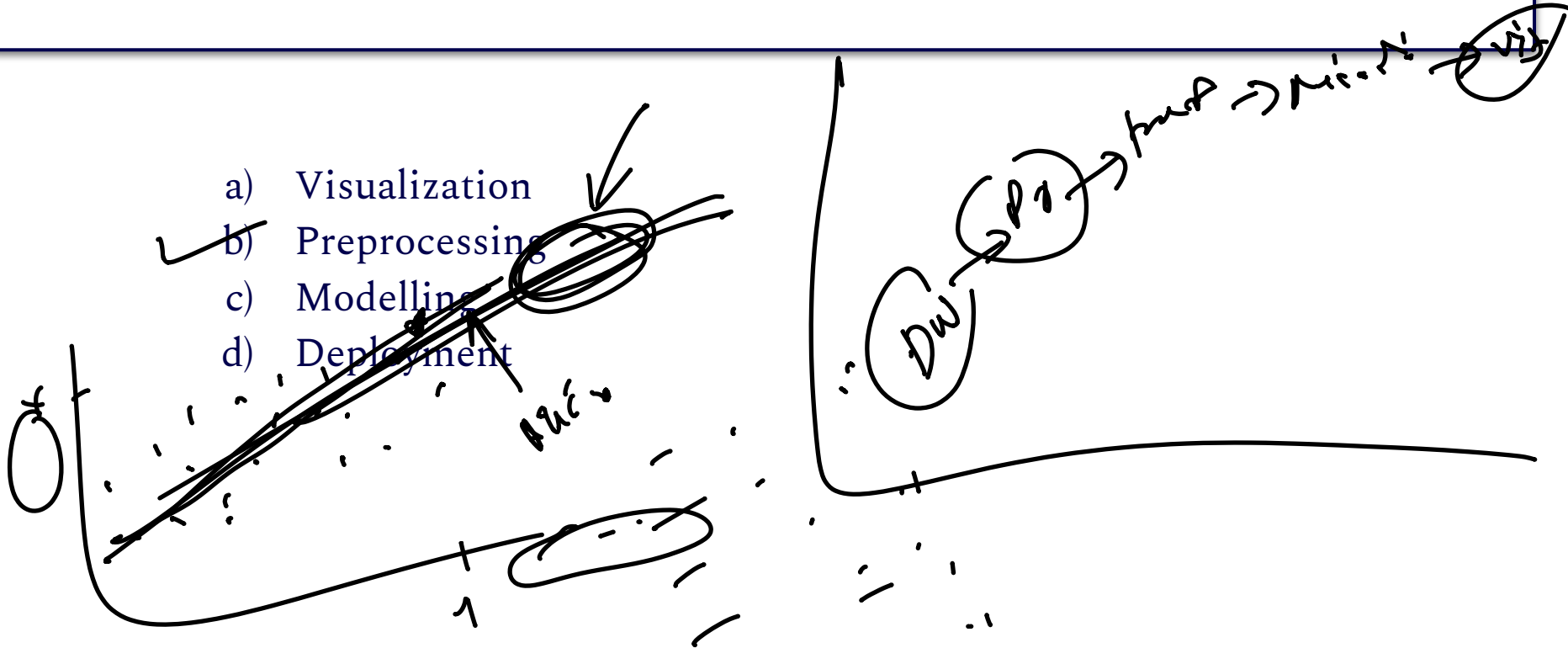
- Data Mining (Knowledge Discovery)
- KDD Process
- Data Preprocessing (Attr. types, data type, noise, etc.)
- Association Rules
- Apriori Algorithm



The diagram illustrates a transaction database table. It shows a grid with columns labeled $A_1, A_2, A_3, A_4, \dots$ and rows representing transactions. The first row contains the values 1, 1, 1, 1, and the second row contains the values 1, 1, 1, 1. The grid is drawn with hand-drawn lines, and the labels are also hand-drawn.

Q1. The earliest step in the data mining process is usually?

- a) Visualization
- ✓ b) Preprocessing
- c) Modelling
- d) Deployment



Q2. Which of the following is an example of a continuous attribute?

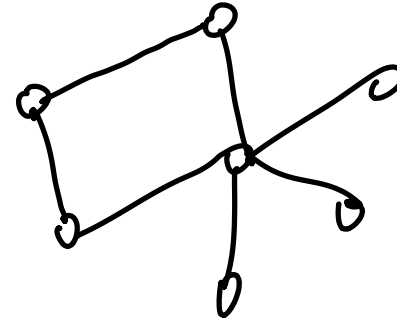
- ✓ a) Height of a person
- b) Name of a person
- c) Gender of a person
- d) None of the above

0.1
✓?
2.
2 - 5-

Q3. Friendship structure of users in a social networking site can be considered as an example of:

- a) Record data
- b) Ordered data
- c) Graph data
- d) None of the above

$G = \{V, E\}$
 $V = \text{Users}$
 $E = \text{Friendship}$



Q4. Name of a person, can be considered as an attribute of type?

- a) Nominal
- b) Ordinal
- c) Interval
- d) Ratio

P ₁	20 E	25 Jan	5-52x
P ₂	26 Jan	26 Jan	
P ₃	26 Jan	26 Jan	

Q5. A store sells 15 items. The maximum possible number of candidate 2-itemsets is:

- a) 120
- b) 105
- c) 150
- d) 2

$${}^{15}C_2 = \frac{15 \times 14}{2 \times 1} = \frac{210}{2} = 105$$

A store = 30 items

t ₁	m, b
t ₂	w, m, b
t ₃	

${}^{15}C_2 = \frac{15 \times 14}{2 \times 1} = 105$
 ${}^{15}C_2 = \frac{15 \times 14}{2 \times 1} = 105$
 ${}^{15}C_2 = \frac{15 \times 14}{2 \times 1} = 105$

Q6. If a record data matrix has a reduced number of rows after a transformation, the transformation has performed:

- a) Data Sampling
- b) Dimensionality Reduction
- c) Noise Cleaning
- d) Discretization

A hand-drawn diagram of a data matrix. The matrix has 4 rows and 4 columns. The columns are labeled A_1 , A_2 , A_3 , and \dots . The rows are labeled O_1 , O_2 , O_3 , and \dots . A bracket on the left side of the matrix is labeled "rows". A bracket on the top side of the matrix is labeled "Colr.". The matrix is drawn with a grid of lines.

	A_1	A_2	A_3	\dots
O_1				
O_2				
O_3				
\dots				

Answer Q7-Q10 based on the following table:

1 {a,d,e,b,c} MB
 2 {a,b,c,d,e}
 3 {b,c,d,e}
 4 {a,b,c,d}
 5 {a,b,d,e}

Customer ID	Transaction ID	Items Bought
1	1	{a,d,e}
1	2	{a,b,c,e}
2	3	{a,b,d,e}
2	4	{a,c,d,e}
3	5	{b,c,e}
3	6	{b,d,e}
4	7	{c,d}
4	8	{a,b,c}
5	9	{a,d,e}
5	10	{a,b,e}

5

e b d b d e
 } 1 1 1
 } 1 1 1
 } 1 1 1
 } 0 1 0
 } 1 1 1

Q7. Taking transaction ID as a market basket, support for each itemset $\{e\}$, $\{b,d\}$, and $\{b,d,e\}$ is:

- ☒ a) 0.8, 0.2, 0.2
- b) 0.3, 0.3, 0.4
- c) 0.25, 0.25, 0.5
- d) 1,0,0

$$\begin{aligned}\{e\} &= \frac{\sigma(e)}{T} = \frac{8}{10} = 0.8 \\ \{b,d\} &= \frac{\sigma(b,d)}{T} = \frac{2}{10} = 0.2 \\ \{b,d,e\} &= \frac{\sigma(b,d,e)}{T} = 0.2\end{aligned}$$

Q8. Based on the results in (7), the confidence of association rules $\{b,d\} \rightarrow \{e\}$ and $\{e\} \rightarrow \{b,d\}$ are:

a) 0.5, 0.5

☒ b) 1, 0.25

c) 0.25, 1

d) 0.75, 0.25

$$\tau_1 = \frac{\sigma(b, d, e)}{\sigma(b, d)} = \frac{2}{2} = 1$$

$$\tau_2 = \frac{\sigma(b, d, e)}{\sigma(e)} = \frac{2}{8} = 0.25$$

Q9. Repeat (7) by taking customer ID as market basket. An item is treated as 1 if it appears in at least one transaction done by the customer, 0 otherwise. Support of itemsets {e}, {b,d}, {b,d,e} are:

- a) 0.3, 0.5, 0.2
- b) 0.8, 1, 0.2 ✓
- c) 1, 0.2, 0.8
- ✓ d) 0.8, 1, 0.8 —

$$\{e\} = \frac{\sigma(\{e\})}{CT} = \frac{4}{5} = 0.8$$

$$\{b,d\} = \frac{\sigma(\{b,d\})}{CT} = \frac{5}{5} = 1$$

$$\{b,d,e\} = \frac{\sigma(\{b,d,e\})}{CT} = \frac{4}{5} = 0.8$$

Q10. Based on the results in (9), the confidence of association rules $\{b,d\} \rightarrow \{e\}$ and $\{e\} \rightarrow \{b,d\}$ are:

$$C_R = \frac{\sigma(\{b,d,e\})}{\sigma(\{b,d\})} = \frac{4}{5} = 0.8$$

- ✓ a) 0.8, 1
- b) 1, 0.8
- c) 0.25, 1
- d) 1, 0.25

$$C_F = \frac{\sigma(\{b,d,e\})}{\sigma(\{e\})} = \frac{4}{4} = 1$$