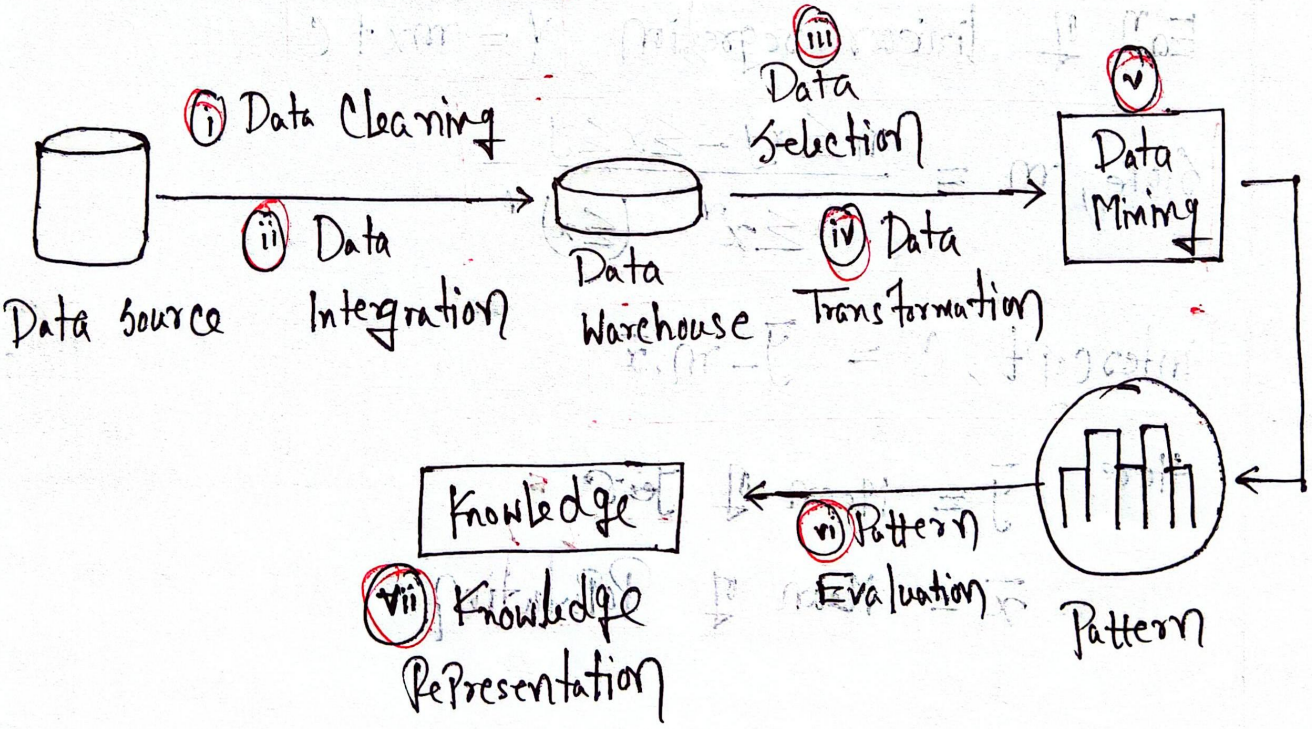
"Machine Learning & Data Mining"

Segment - (01)

Data mining: Data mining is a process of extracting useful information and insights from large data set.

KDD Process: (Knowledge Discovery from Data)



7-steps of Data mining

Adv. of KDD

(i) Decision support
(ii) Automation
(iii) Discovery of hidden Pattern

Dis-Adv. of KDD

(i) Data Quality
(ii) Computational Complexity
(iii) Data Security

Math Problem:

Eqn of linear regression, $Y = mx + C$

$$\text{Slope, } m = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$$

$$\text{intercept, } C = \bar{y} - m.\bar{x}$$

Here, $\bar{y}$ = Mean of Year

$\bar{x}$ = Mean of Population

SP'23 | (1)

x	y	xy	$(xy)^2$
1980	2.1	4158	3920400
1983	2.4	4759.2	3932289
1986	2.9	5759.4	3944196
1989	3.2	6364.8	3956121
1992	3.8	7569.6	3968064
1995	4.1	8179.5	3980025
1998	4.3	8591.4	3992004
2001	4.7	9404.7	4004001
15924	27.5	54786.6	31697100

Now,

$$\bar{x} = \frac{1980 + 1983 + \dots + 2001}{8} = 1990.5$$

$$\bar{y} = \frac{2.1 + 2.4 + \dots + 4.7}{8} = 3.44$$

$$m = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2} = \frac{8 * 54786.6 - 15924 * 27.5}{8 * 31697100 - (15924)^2}$$

$$= 0.127$$

$$C = 3.44 - 0.127 * 1990.5 = -249.35$$

(i) For, Population (2007) $= mx + C$

$= 0.127 * 2007 + (-249.35)$

$= 5.539$

(ii) Population (2010) $= 0.127 * 2010 + (-249.35)$

$= 5.92$ ✓

Types of Data Mining:

(i) Classification
(ii) Association Analysis
(iii) Clustering
(iv) Regression
(v) Anomaly Detection.

Popular ML Algorithm:

(i) Linear Regression (supervised Learning)
(ii) Artificial Nural Network
(iii) Dicision Tree (classificatin

Segment – 02

Similarity & Dissimilarity:

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

(ii) Manhattan Distance, $d = |x_1-x_2| + |y_1-y_2|$

(iii) Minkowski " , $d = \{(|x_1-x_2|)^p + (|y_1-y_2|)^p\}^{1/p}$

(iv) Supremum " , $d = \max(|x_1-x_2|, |y_1-y_2|)$

(v) Cosine Distance, $d = 1 - \frac{X \cdot Y}{\|X\| \cdot \|Y\|}$ [X, Y = Vector]

dissimilarity

e.g: For, $x = [2, 3, 5]$ and $Y = [1, 4, 6]$

$$d = 1 - \frac{(2*1)+(3*4)+(5*6)}{\sqrt{2^2+3^2+5^2} * \sqrt{1^2+4^2+6^2}} = 0.128$$

Similarity between binary Vector:

(i) Simple Matching Coefficient (SMC) [Symmetric]

$$SMC = \frac{\text{number of matching attribute}}{\text{number of attributes}}$$

$$= \frac{f_{11} + f_{00}}{f_{00} + f_{01} + f_{10} + f_{11}}$$

(ii) Jaccard Co-efficient (J) [Asymmetric binary Attribute]

$$J = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

e.g: P = 1 0 0 0 0 0 0 0 0 0

q = 0 0 0 0 0 0 1 0 0 1

$f_{01} = 2$

$f_{10} = 1$

$f_{00} = 7$

$f_{11} = 0$

(i) $SMC = \frac{f_{11} + f_{00}}{f_{00} + f_{01} + f_{10} + f_{11}} = \frac{0+7}{2+1+7+0}$

(ii) $J = \frac{f_{11}}{f_{01} + f_{10} + f_{11}} = \frac{0}{2+1+0} = 0$

Types of Data / Attributes with example :

(i) Quantitative / Numerical Data

- (i) Discrete Data → Marks of student, Age.
- (ii) Continuous Data → Weight, Height, Temperature.
- (iii) Numeric → Digits

(ii) Qualitative Data

- (i) Nominal Data → Gender, Color, State of residence.
- (ii) Ordinal Data → Ranking, ~~Bool Value~~ etc.
- (iii) Binary Data

Proximity measurement for Binary attribute :

Example :

	Col-1	Col-2	Col-3
object i	0	0	1
	0	1	1
K	0	0	0

a →	1	1
b →	1	0
c →	0	1
d →	0	0

Here, For (i, j) ⇒ a = 1
b = 0
c = 1

Disimilarity for (i,j) :

(i) Distance for symmetric binary,

$$d(i,j) = \frac{b+c}{a+b+c+d} = \frac{0+1}{1+0+1+1}$$

$$= \frac{1}{3} = 0.33$$

(ii) Distance for asymmetric binary,

$$d(i,j) = \frac{b+c}{a+b+c} = \frac{0+1}{1+0+1} = \frac{1}{2} = 0.5$$

Similarity for (i,j):

(i) Jaccard (Asymmetric Binary)

$$J(i,j) = \frac{a}{a+b+c} = \frac{1}{1+0+1} = 0.5$$

(ii) Coherence Co-efficient (symmetric binary)

$$C(i,j) = \frac{a+d}{a+b+c+d} = \frac{2}{3} = 0.67$$

Thus final value for (i,k) and (J,k) ——

———— ∘ ————

Proximity Measures of Nominal Attributes :

$$d(i,j) = \frac{p - m}{p}$$

p → Numbers of attribute
m → Number of matches

$$sim(i,j) = 1 - d(i,j)$$

$$= 1 - \frac{p-m}{p} = \frac{m}{p}$$

SP'22 | 1©

~~Write~~ Using Cosine measurement for finding similarity and disimilarity —

For (Doc-1 and Doc-2),

$$d(D_1, D_2) = \frac{D_1 * D_2}{\|D_1\| * \|D_2\|}$$

$$D_1 * D_2 = 5*3 + 3*2 + 2*1 + 2*1$$

$$= 15 + 6 + 2 + 2$$

$$= 25$$

$$\|D_1\| * \|D_2\| = \sqrt{5^2+3^2+2^2+2^2} * \sqrt{3^2+2^2+1+1+1+1}$$

$$= 6.481 * 4.123$$

$$= 26.722$$

$$\cos(D_1, D_2) = \frac{25}{26.72} = 0.94$$

So, $\text{Sim}(D_1, D_2) = 1 - 0.94 = 0.06$

Similarly Calculate for (D_1, D_3) (D_1, D_4) (D_2, D_3) (D_2, D_4) (D_3, D_4) ——

(i) Minimum value will be most similar

(ii) Maximum Value ⇒ Most dissimilar

For Numeric Attribute :

$$d(i,j) = \frac{|i^{th} \text{ Attribute value} - j^{th} \text{ attribute value}|}{\text{Max attribute value} - \text{Min attribute value}}$$

For Ordinal Attribute :

Obj	T
1	High
2	low
3	Medium
4	High

M_f = Total unique Attribute val

Step 1 : Give initial value of attributes

High = 1, medium = 2, Low = 3

Step 2 : Normalize the ranking between (0 to 1)

$$Z_i = \frac{\text{initial value of } j^{th} \text{ attribute} - 1}{M_f - 1}$$

$$\text{High} = \frac{1-1}{3-1} = \frac{0}{2} = 0$$

$$\text{Medium} = \frac{2-1}{3-1} = \frac{1}{2} = 0.5$$

$$\text{Low} = \frac{3-1}{3-1} = 1$$

Step 4 : $dis(2,1) = (0-0) = 1$

$d(3,1) = (.5-0) = 0.5$

$d(4,1) = (0-0) = 0$

0	
1	0
.5	.5	0
0	1	.5	0

SP'22 | 2(a)

For Test-1 (Nominal)

$$d(i,j) = \frac{\text{P-value}}{P}$$

$$d(2,1) = \frac{1-0}{1} = 1$$

$$d(3,1) = \frac{1-1}{1} = 0$$

$$d(3,2) = \frac{1-0}{1} = 1$$

0	
1	0
0	1	0

For Test-3 (Numeric)

$$d(2,1) = \frac{|220 - 440|}{440 - 121} = \frac{220}{319} = 0.69$$

$$d(3,1) = \frac{|121 - 440|}{319} = 1$$

$$d(3,2) = \frac{|121 - 220|}{319} = 0.31$$

$$\begin{bmatrix} 0 & & \\ .69 & 0 & \\ 1 & .31 & 0 \end{bmatrix}$$

For Test-4 (Binary)

$d(2,1) \Rightarrow a = 0, b = 1, c = 0$

$$d(2,1) = \frac{1 + 0}{0 + 1 + 0} = 1$$

$$d(3,1) = \frac{0 + 0}{} = 0$$

$$d(3,2) = \frac{0 + 1}{1} = 1$$

$$\begin{bmatrix} 0 & & \\ 1 & 0 & \\ 0 & 1 & 0 \end{bmatrix}$$

Mixed Attribute :

$$\begin{bmatrix} 0 & & \\ (1 + .69 + 1)/3 & 0 & \\ (0 + 1 + 0)/3 & (1 + .31 + 1)/3 & 0 \end{bmatrix} = \begin{bmatrix} 0 & & \\ .897 & 0 & \\ .33 & .77 & 0 \end{bmatrix}$$

Ans

Reasons for missing data:

(i) Data entry errors

(ii) Technical issues

(iii) Incomplete data Collection

(iv) Privacy Concern

(v) Outliers : Data Points that fall outside the expected range may be excluded as outliers, even though they are valid.

3- Missing Value Imputation technique:

(i) Mean/median : Replace missing value with avarage.

(ii) Mode : Use most frequent value in gap.

(iii) Zero imputation / Constant : Replace missing value with '0'.

Segment - (3)

Apriori Algorithm:

Example: Min Support – 50%

Confidence → 70%

Tid	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

Soln: **Step 1:**

Items	Support
1	2/4 = 50%
2	3/4 = 75%
3	3/4 = 75%
4	1/4 = 25%
5	3/4 = 75%

As 4 → 25% is below support 50%

So, 4 is eliminated.

Step 2: (1, 2, 3, 5)

Items	Support
(1,2)	1/4 = 25% ✗
(1,3)	2/4 = 50%
(1,5)	1/4 = 25% ✗
(2,3)	2/4 = 50%
(2,5)	3/4 = 75%
(3,5)	2/4 = 50%

Step 3:

Items	Support
✗ (1,2,3)	1/4 = 25%
✗ (1,2,5)	1/4 = 25%
(2,3,5)	2/4 = 50% ←
✗ (1,3,5)	1/4 = 25%

Step 4: (2,3,5)

Rules	Support	Confidence
(2^3) → 5	2	100%
(3^5) → 2	2	100%
(2^5) → 3	2	66% ×
3 → (2^5)	2	66% ×
2 → (3^5)	2	66% ×
5 → (2^3)	2	66% ×

Confidence = s(A∪B) / s(A)

For, (2^3) → 5 Confidence = s{(2^3) ∪ 5} / s(2^3)

A B

= 2/2 = 100%

As the first 2 is above Confidence level

So, (2^3) → 5 & (3^5) → 2 final association rules.

(i) Min-Max Normalization:

$$\bar{x} = \frac{x - Min}{Max - Min}$$

(ii) Z-score Normalization:

$$Z = \frac{x - Mean}{std}$$

x = Original Value
Min = Min value of Column
Max = Max " "
Mean = Mean of Column
std = Standard deviation of Column.

Aut'22 | 3.5

humidity	
53	x_0
53	x_1
64	x_2
50	x_3
51	x_4

Soln: Min-Max

$$\bar{x}_0 = \frac{53 - 50}{64 - 50} = 0.21$$

$$\bar{x}_1 = \frac{53 - 50}{64 - 50} = 0.21$$

$$\bar{x}_2 = \frac{64 - 50}{64 - 50} = 1$$

Dress Code / in Islam

Humidity
0.21
0.21
1
0
0.07

$$\bar{x}_3 = \frac{50 - 50}{64 - 50} = 0$$

$$\bar{x}_4 = \frac{51 - 50}{14} = 0.07$$

Z-score:

$$\text{mean} = 53 + 53 + 64 + 50 + 51 / 5 = 54.2 = \mu$$

$$\text{std} = \frac{1}{n}\left(\sqrt{|x_0 - \mu|^2 + (|x_1 - \mu|)^2 + |x_2 - \mu|^2 + \dots + |x_n - \mu|^2}\right)$$

$$= 3.94$$

$$Z_0 = \frac{53 - 54.2}{3.94} = -0.3$$

$$Z_1 = -0.3$$

$$Z_2 = \frac{64 - 54.2}{3.94} = 2.48$$

Five Number Summary:

(i) Min (ii) Q_1 (iii) Median (iv) Q_3 (v) Max

Exercise:

Dataset ⇒ 1, 3, 5, 6, 9, 10, 12, 15, 17, 20

(i) **Five Number Summary:**

Min = 1

$$Q_1 = \frac{3+5}{2} = 4$$

$$\text{Median} = \frac{9+10}{2} = 9.5$$

$$Q_3 = \frac{15+17}{2} = 16$$

Max = 20

(ii) IQR Method :

$$IQR = Q_3 - Q_1$$

$$= 16 - 4 = 12$$

$$\text{Lower bound} = Q_1 - (1.5 * IQR)$$

$$= 4 - (1.5 * 12)$$

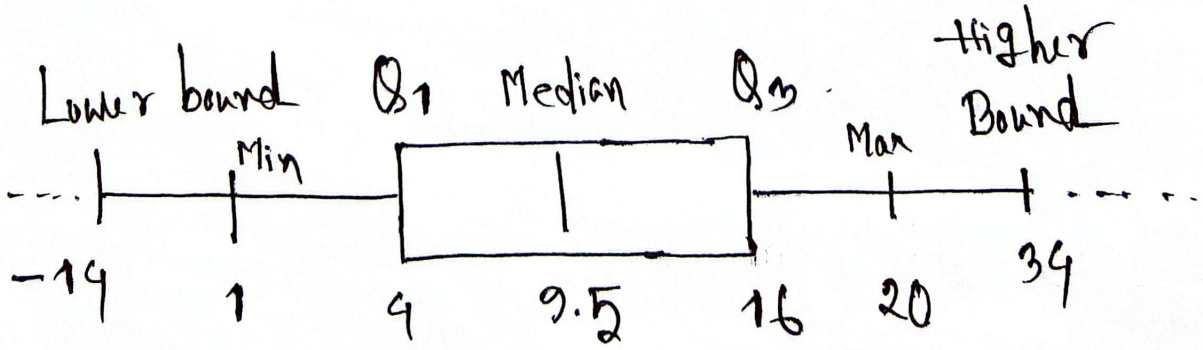
$$= -14$$

$$\text{Upper bound} = Q_3 + (1.5 * IQR)$$

$$= 16 + (1.5 * 12)$$

$$= 34$$

(iii) Box-Plot of Data



FAHIM REZA | C211046
fahim.bdctg51@gmail.com