PAGE No.:
DATE : / /

Q.1) Binning:- Morgan Kaufman (page no. 90).

A) Data = [4, 8, 15, 21, 21, 24, 25, 28, 34].

n = 9

∴ Total equal bins formed = 3.

Bin 1 = [4, 8, 15]
Bin 2 = [21, 21, 24]
Bin 3 = [25, 28, 34].

1) Smoothing by bin mean.

∴ Bin 1 :- mean = $\frac{4+8+15}{3}$ = 9.

Bin 1 = [9, 9, 9].

Bin 2 :- mean = $\frac{21+21+24}{3}$ = 22

Bin 2 = [22, 22, 22].

Bin 3 :- mean = $\frac{25+28+34}{3}$ = 29

Bin 3 = [29, 29, 29].

2) Smoothing by Bin median.

median = (for odd) = $\left(\frac{n+1}{2}\right)^{th}$ position

∴ Bin 1 = [8, 8, 8]
Bin 2 = [21, 21, 21]
Bin 3 = [28, 28, 28].

PAGE No.:
DATE : / /

3) Smoothing by bin boundary:-
Replace by nearest boundary value, i.e less difference.

Bin 1 = [4, 4, 15]
Bin 2 = [21, 21, 24]
Bin 3 = [25, 25, 34].

B] Exercise pg. 121, problem - 3.3.
Data = 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.
Bin depth = 3.
∴ Bin 1 = 13, 15, 16
Bin 2 = 16, 19, 20
Bin 3 = 20, 21, 22
Bin 4 = 22, 25, 25
Bin 5 = 25, 25, 30
Bin 6 = 33, 33, 35
Bin 7 = 35, 35, 35
Bin 8 = 36, 40, 45
Bin 9 = 46, 52, 70

1) Smoothing by bin mean. = $\frac{\Sigma x_i}{n}$.
~~Bin~~
Bin 1 = [14.66, 14.66, 14.66]
Bin 2 = [18.33, 18.33, 18.33]
Bin 3 = [21, 21, 21]
Bin 4 = [24, 24, 24]
Bin 5 = [26.66, 26.66, 26.66]
Bin 6 = [33.66, 33.6, 33.6]
Bin 7 = [35, 35, 35]
Bin 8 = [40.3, 40.3, 40.3]
Bin 9 = [56, 56, 56].

PAGE No.:
DATE : / /

2) Smoothing by bin median.

Bin1 = [15 15 15]
Bin2 = [19 19 19]
Bin3 = [21 21 21]
Bin4 = [25, 25, 25]
Bin5 = [25, 25, 25]
Bin6 = [33, 33, 33]
Bin7 = [35, 35, 35]
Bin8 = [40, 40, 40]
Bin9 = [52, 52, 52].

3) Smoothing by bin boundary :-

Bin1 = [13, 16, 16]
Bin2 = [16, 20, 20]
Bin3 = [20, 20, 22]
Bin4 = [22, 25, 25]
Bin5 = [25, 25, 30]
Bin6 = [33, 33, 35]
Bin7 = [35, 35, 35]
Bin8 = ~~[40, 40, 45]~~ [36, 36, 45]
Bin9 = [46, 46, 70]

Q.2) Normalization. (page no. ~~112~~ 121, Q. 3.6).

A) Data = 200, 300, 400, 600, 1000. range = [0, 1].

Min-max Normalization :-

$$v' = \frac{(v - min(A))}{(max(A) - min(A))} \times (newmax(A) - newmin(A)) + newmin(A).$$

| v | Formula. | v' |
|---|---|---|
| 200 | $\frac{(200-200)}{1000-200} \times (1-0) + 0$ | 0 |
| 300 | $\frac{300-200}{1000-200} \times (1-0) + 0$ | 0.125 |
| 400 | $\frac{400-200}{1000-200} \times (1-0) + 0$ | 0.25 |
| 600 | $\frac{600-200}{1000-200} \times (1-0) + 0$ | 0.5 |
| 1000 | $\frac{1000-200}{1000-200} \times (1-0) + 0.$ | 1 |

PAGE No.:
DATE : / /

∴ min-max Normalized data = 0, 0.125, 0.25, 0.5, 1

2) Z-Score Normalize.

$$v' = \frac{(v - \bar{A})}{\sigma A}$$

$\bar{A} = (200 + 300 + 400 + 600 + 1000) / 5$

$= 500$.

| v | $(v - \bar{A})^2$ |
|---|---|
| 200 | 90000 |
| 300 | 40000 |
| 400 | 10000 |
| 600 | 10000 |
| 1000 | 250000 |

$$\sigma A = \sqrt{\frac{\Sigma (v - \bar{A})^2}{n}} = \sqrt{\frac{400000}{5}}$$

$= \sqrt{80000}$

$= 89.442$

| v | Formula | v' |
|---|---|---|
| 200 | $\frac{200 - 500}{89.442}$ | -3.354 |
| 300 | $\frac{300 - 500}{89.442}$ | -2.236 |
| 400 | $\frac{400 - 500}{89.442}$ | -1.118 |
| 600 | $\frac{600 - 500}{89.442}$ | 1.118 |
| 1000 | $\frac{1000 - 500}{89.442}$ | 5.590 |

PAGE No.:
DATE : / /

B) Data = 23, 27, 39, 41, 47, 49, 50, 52, 54, 56, 57, 58, 60, 61. Range = [0, 1].

- min max Normalization :-

$$v' = \frac{(v - min(A))}{(max(A) - min(A))} \times (newmax(A) - newmin(A)) + newmin(A).$$

| v | Formula | v' |
|---|---|---|
| 23 | ((23-23)/61-23) x (1-0) + 0 | 0 |
| 27 | ((27-23)/61-23) x (1-0) + 0 | 0.1052 |
| 39 | ((39-23)/61-23) x (1-0) + 0 | 0.421 |
| 41 | ((41-23)/61-23) x (1-0) + 0 | 0.4736 |
| 47 | ((47-23)/61-23) x (1-0) + 0 | 0.6315 |
| 49 | ((49-23)/61-23) x (1-0) + 0 | 0.6884 |
| 50 | ((50-23)/61-23) x (1-0) + 0 | 0.7105 |
| 52 | ((52-23)/61-23) x (1-0) + 0 | 0.7631 |
| 54 | ((54-23)/61-23) x (1-0) + 0 | 0.8157 |
| 56 | ((56-23)/61-23) x (1-0) + 0 | 0.8684 |
| 57 | ((57-23)/61-23) x (1-0) + 0 | 0.8947 |
| 58 | ((58-23)/61-23) x (1-0) + 0 | 0.921 |
| 60 | ((60-23)/61-23) x (1-0) + 0 | 0.9736 |
| 61 | ((61-23)/61-23) x (1-0) + 0. | 1 |

- Z Score Normalization.

$$v' = \frac{(v - \bar{A})}{\delta A}$$

$$\bar{A} = \cancel{221} \quad \frac{\Sigma A}{n} = \frac{674}{14} = 48.142$$

$$\delta A = 11.375.$$

PAGE No.:
DATE : / /

| V | Formula. | V' |
|---|---|---|
| 23 | (23 − 48.142) / 11.375. | −2.21 |
| 27 | (27 − 48.142) / 11.375 | −1.858 |
| 39 | (39 − 48.142) / 11.375 | −0.803 |
| 41 | (41 − 48.142) / 11.375 | −0.627 |
| 47 | (47 − 48.142) / 11.375 | −0.1 |
| 49 | (49 − 48.142) / 11.375 | 0.0754 |
| 50 | (50 − 48.142) / 11.375 | 0.1633 |
| 52 | (52 − 48.142) / 11.375 | 0.3391 |
| 54 | (54 − 48.142) / 11.375 | 0.5149 |
| 56 | (56 − 48.142) / 11.375 | 0.6908 |
| 57 | (57 − 48.142) / 11.375 | 0.7787 |
| 58 | (58 − 48.142) / 11.375 | 0.8666 |
| 60 | (60 − 48.142) / 11.375 | 1.0424 |
| 61 | (61 − 48.142) / 11.375. | 1.1303. |

Q.3) Box plot :-

A) page no - 49. Example. 2.10.

Data = 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110.
(in thousands)

→ 1) minimum = 30

2) maximum = 110.

3) median value = $\frac{(n/2)^{th} + (n/2+1)^{th}}{2}$

$= \frac{(12/2)^{th} + (12/2+1)^{th}}{2}$

$= (6^{th} + 7^{th})/2$

$= 54$

PAGE No.:
DATE : / /

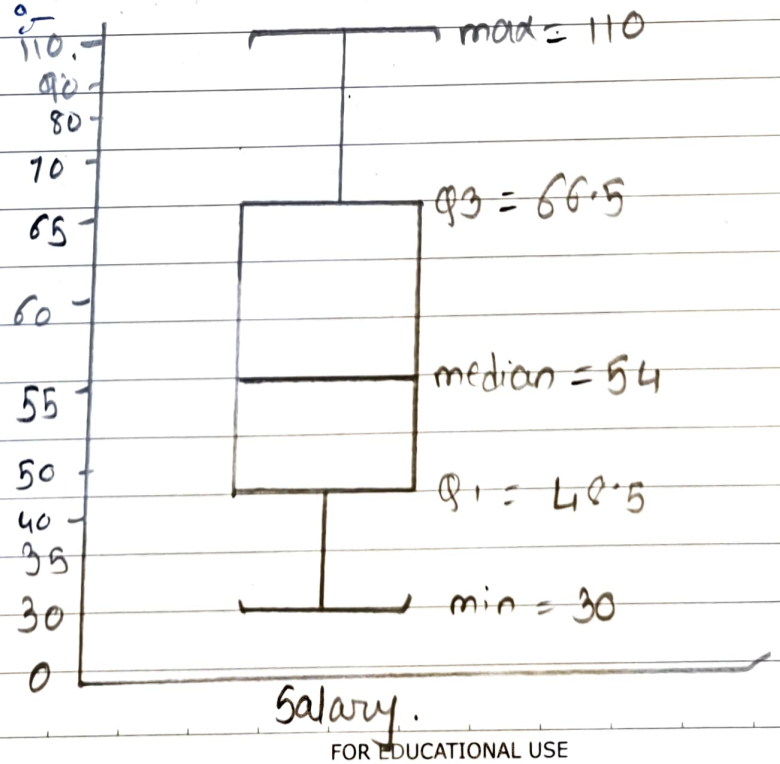
4) $Q_1$ = 1/4th value
= $(n/4^{th} + (n/4+1)^{th})/2$
= $(3^{rd} + 4^{th})/2$
= $(47+50)/2$
= 48.5.

5) $Q_3$ = 3/4th value.
= $[(3n/4)^{th} + (3n/4+1)^{th})/2$
= $(9^{th} + 10^{th})/2$
= $(63+70)/2$
= 66.5.

6) IQR = $Q_3 - Q_1$
= 66.5 − 48.5
= 18.

7) whiskers = 1.5 × IQR
= ± 27

Boxplot :-



PAGE No.:
DATE : / /

B) page no. 80 problem 2.2.

Data(Age) = 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

→ 1) minimum = 13

2) maximum = 70.

3) median value = $((n/2)^{th} + (n/2+1)^{th})/2$

$= ((27/2) + (27/2)+1)/2$

$= (13^{th} + 14^{th})/2$

$= (25+25)/2$

$= 25.$

4) $Q_1$ = 1/4th value.

$= ((n/4) + (n/4+1))/2$

$= (6^{th} + 7^{th})/2$

$= 20+20/2$

$= 20$

5) $Q_3$ = 3/4th value.

$= ((3n/4) + (3n/4+1))/2$

$= (20^{th} + 21^{th})/2$

$= (35+35)/2$

$= 35.$

6) IQR = $Q_3 - Q_1$

$= 35 - 20$

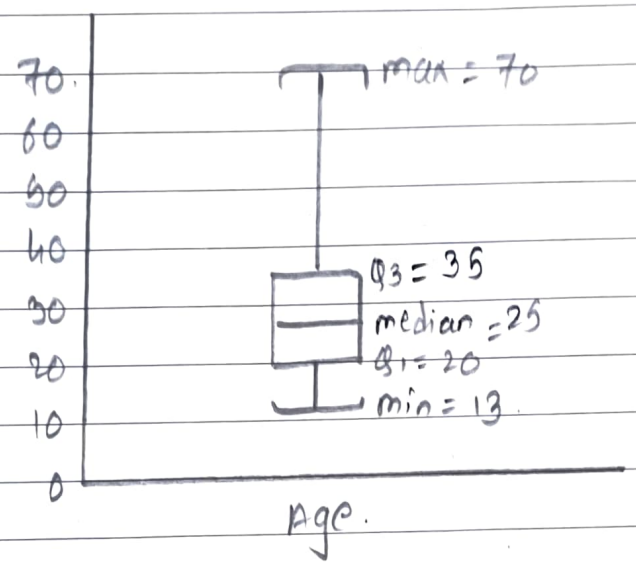
$= 15.$

7) whiskers = $1.5 \times IQR$

$= \pm 22.5.$

PAGE No.:
DATE : / /

Boxplot =




Q.4) twt. dwt :-

A) Class room example :-

| Region | Item 1 | Item 2 | Total. |
|---|---|---|---|
| $R_1$ | 80 | 240 | 320 |
| $R_2$ | 120 | 560 | 680 |
| Total | 200 | 800 | 1000. |

→ Solution.

| Region | Item 1 | | | Item 2 | | | Total. |
|---|---|---|---|---|---|---|---|
| | twt | dwt | Total | twt | dwt | Total | |
| $R_1$ | 25% | 40% | 80 | 75% | 30% | 240 | 320 |
| $R_2$ | 18% | 60% | 120 | 82% | 70% | 560 | 680 |
| Total. | | 100% | 200 | | 100% | 800 | 1000. |

B) Pg. 137 Table 4.2.

| Time | TV | PC | Phone | Security. |
|---|---|---|---|---|
| Q1 | 605 | 825 | 14 | 400 |
| Q2 | 680 | 952 | 31 | 512 |
| Q3 | 812 | 1023 | 30 | 501 |
| Q4 | 927 | 1038 | 38 | 580. |

PAGE No.:
DATE : / /

→ Solution.

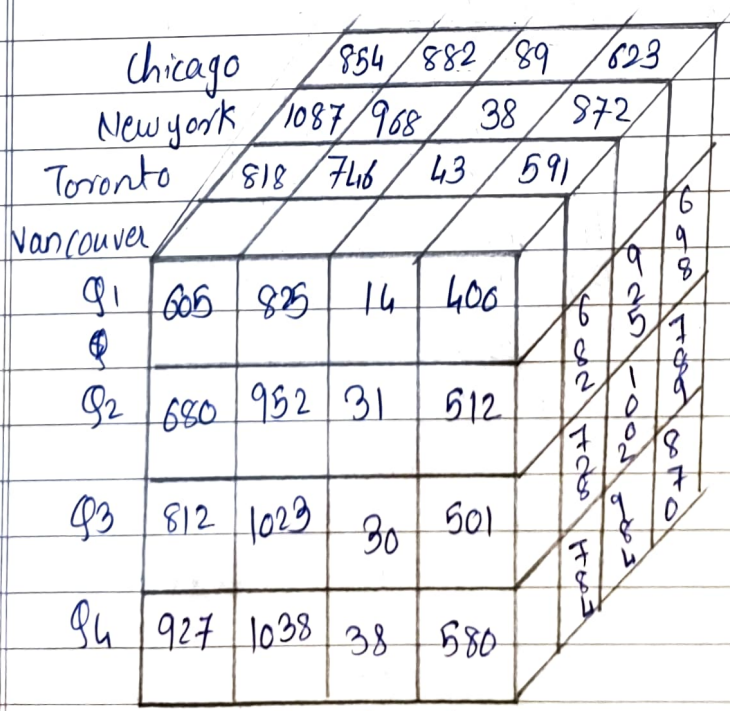
| Time | TV | | | PC | | | Phone | | | Security | | | Total. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Twt | dwt | Tot. | Twt | dwt. | Total | Twt | Dwt. | Tot | Twt. | dwt. | Total | |
| Q1 | 32.81 | 20.01 | 605 | 44.74 | 21.5 | 825 | 0.76 | 12.39 | 14 | 21.69 | 20.07 | 400 | 1844. |
| Q2 | 91.26 | 22.49 | 680 | 47.7 | 24.8 | 952 | 1.43 | 27.43 | 31 | 23.5 | 25.6 | 512 | 2175 |
| Q3 | 343 | 26.8 | 812 | 43.2 | 26.6 | 1023 | 1.27 | 26.5 | 30 | 21.1 | 25.1 | 501 | 2366 |
| Q4 | 35.8 | 30.6 | 927 | 40.1 | 2705 | 1038 | 1.47 | 33.6 | 38 | 22.45 | 29.10 | 580 | 2583 |

(Q.5) Generate OLAP cube from Schema date.

A) Data = pg. 137 Table 4.3.

~~3D~~ Sales Data of All electronics according to time, item & location.

→ Solution :-



PAGE No.:
DATE : / /

Q.6) Calculate Entroply & Info Gain. of dataset.
A) Data = Table 1 from datasheet.
→ Given data =
Target class = Play. = 9/14
Contrasting class = No Play = 5/14.

A = Outlook.

| | Play | NoPlay. | |
|---|---|---|---|
| Sunny | 2 | 3 | 5/14 |
| Overcast | 4 | 0 | 4/14 |
| rain | 3 | 2 | 5/14 |

A = Temperature

| | Play | NoPlay | |
|---|---|---|---|
| hot | 2 | 2 | 4/14 |
| mild | 4 | 2 | 6/14 |
| Cold | 3 | 1 | 4/14 |

A = Humidity

| | Play | Noplay. | |
|---|---|---|---|
| High | 3 | 4 | 7/14 |
| Normal | 6 | 1 | 7/14 |

A = Wind

| | Play | Noplay. | |
|---|---|---|---|
| True | 3 | 3 | 6/14 |
| False | 6 | 2 | 8/14. |

→ Solution:-

$$\text{Entroply}(D) / \text{Info}(D) = -\sum_{i=1}^{n} p_i \log_2(p_i).$$

$$= -\left[\frac{9}{14}\log_2\left(\frac{9}{14}\right) + \frac{5}{14}\left[\log_2\left(\frac{5}{14}\right)\right]\right]$$

$$= 0.9403.$$

PAGE No.:
DATE : / /

$$Info(Outlook, D) = \sum_{j=1}^{V} \frac{|D_j|}{D} \times info(D_j).$$

$$= \frac{5}{14} \times \left[ \frac{-2}{5} \log_2\left(\frac{2}{5}\right) - \frac{3}{5} \log_2\left(\frac{3}{5}\right) \right] +$$
$$\frac{4}{14} \left[ \frac{-4}{4} \log_2\left(\frac{4}{4}\right) - \frac{0}{4} \log_2\left(\frac{0}{4}\right) \right] +$$
$$\frac{5}{14} \left[ \frac{-3}{5} \log_2\left(\frac{3}{5}\right) - \frac{2}{5} \log_2\left(\frac{2}{5}\right) \right]$$
$$= 0.694.$$

$$Info(temp, D) = \sum_{j=1}^{V} \frac{|D_j|}{D} \times info(D_j).$$

$$= \frac{4}{14} \left[ \frac{-2}{4} \log_2\left(\frac{2}{4}\right) + - \frac{2}{4} \log_2\left(\frac{2}{4}\right) \right] +$$
$$\frac{6}{14} \left[ \frac{-4}{6} \log_2\left(\frac{4}{6}\right) - \frac{2}{6} \log_2\left(\frac{2}{6}\right) \right] +$$
$$\frac{4}{14} \left[ \frac{-3}{4} \log_2\left(\frac{3}{4}\right) - \frac{1}{4} \log_2\left(\frac{1}{4}\right) \right]$$
$$= 0.9109.$$

$$Info(Humidity, D) = \sum_{j=1}^{V} \frac{|D_j|}{D} \times info(D_j)$$

$$= \frac{7}{14} \left[ \frac{-3}{7} \log_2\left(\frac{3}{7}\right) - \frac{4}{7} \log_2\left(\frac{4}{7}\right) \right] +$$
$$\frac{7}{14} \left[ \frac{-6}{7} \log_2\left(\frac{6}{7}\right) - \frac{1}{7} \log_2\left(\frac{1}{7}\right) \right]$$
$$= 0.7884$$

PAGE No.:
DATE : / /

$$Info(wind, D) = \sum_{j=1}^{v} \frac{|D_j|}{D} \times \log(D_j)$$

$$= \frac{6}{14}\left(\frac{-3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right)\right) + \frac{8}{14}\left(\frac{-6}{8}\log_2\left(\frac{6}{8}\right) - \frac{2}{8}\log_2\left(\frac{2}{8}\right)\right)$$

$$= 0.8921$$

∴ Information Gain = Info(D) − Info(A, D)

Gain(outlook) = 0.9403 − 0.694
= 0.2463.

Gain(Temp) = 0.9403 − 0.9109
= 0.0294

Gain(Humidity) = 0.9403 − 0.7884
= 0.1519

Gain(wind) = 0.9403 − 0.8921
= 0.0482.

As Gain of outlook is more it is selected for splitting criteria.

B) Data = Table 4 from datasheet.

Given data=

Target class. = Rain(yes) 4/10

Contrasting class = Rain(No) 6/10.

A = Temprature.

| | Rain(yes) | Rain(No) | |
|---|---|---|---|
| High | 2 | 1 | 3/10 |
| medium | 1 | 2 | 3/10 |
| low. | 1 | 3 | 4/10. |

PAGE No.:
DATE: / /

A = Humidity.

| | Pt. Rain(yes) | Rain(No) | |
|---|---|---|---|
| High | 2 | 2 | 4/10 |
| Low. | 2 | 4 | 6/10. |

A = Clouds.

| | Rain(yes) | Rain(No. | |
|---|---|---|---|
| ye | 4 | 1 | 5/10 |
| NO. | 0 | 5 | 5/10. |

→

$$Into(D) = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

$$= -\left[\frac{4}{10}\log_2\left(\frac{4}{10}\right) + \frac{6}{10}\log_2\left(\frac{6}{10}\right)\right]$$

$$= 0.9709.$$

$$Into(Temp) = \sum_{j=1}^{v} \frac{|D_j|}{D} \times into(D_j).$$

$$= \frac{3}{10}\left(-\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right) +$$

$$\frac{3}{10}\left(-\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right) +$$

$$\frac{4}{10}\left(-\frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{3}{4}\log_2\left(\frac{3}{4}\right)\right)$$

$$= 0.2754 + 0.2754 + 0.3245$$

$$= 0.8753.$$

PAGE No.:
DATE : / /

$$Info(Humidity) = \frac{4}{10}\left(-\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right) + \frac{6}{10}\left(-\frac{2}{6}\log_2\left(\frac{2}{6}\right) - \frac{4}{6}\log_2\left(\frac{4}{6}\right)\right)$$

$$= 0.4 + 0.5509$$

$$= 0.9509$$

$$Info(cloud) = \frac{5}{10}\left(-\frac{4}{5}\log_2\left(\frac{4}{5}\right) - \frac{1}{5}\log_2\left(\frac{1}{5}\right)\right) + \frac{5}{10}\left(-\frac{0}{5}\log_2\left(\frac{0}{5}\right) - \frac{5}{5}\log_2\left(\frac{5}{5}\right)\right)$$

$$= 0.3609 + 0$$

$$= 0.3609$$

$$\therefore Gain(Temp) = Info(D) - Info(Temp)$$

$$= 0.9709 - 0.8753$$

$$= 0.0956$$

$$Gain(Humidity) = 0.9709 - 0.9509.$$

$$= 0.02$$

$$Gain(Cloudy) = 0.9709 - 0.3609.$$

$$= 0.61.$$

As info Gain (Cloud) is high it is used as splitting criteria.

PAGE No.:
DATE : / /

Q.7) Find Frequent itemset & derive support & confidence of association rule.

A) pg. 274, Q. 6.6.

Given, min_sup = 60%. min_conf = 80%.

| TID | Items |
|---|---|
| $T_1$ | M, O, N, K, E, Y |
| $T_2$ | D, O, N, K, E, Y |
| $T_3$ | M, A, K, E |
| $T_4$ | M, U, C, K, Y |
| $T_5$ | C, O, O, K, I E |

→ For K = 1.

| Itemset | Frequency | Support | |
|---|---|---|---|
| E | 4 | 80% | ✓ |
| N | 2 | 40% | |
| K | 5 | 100% | ✓ |
| O | 3 | 60% | ✓ |
| M | 3 | 60% | ✓ |
| Y | 3 | 60% | ✓ |
| D | 1 | 20% | |
| A | 1 | 20% | |
| U | 1 | 20% | |
| C | 2 | 40% | |
| I | 1 | 20% | |

∴ Frequent Itemset = {E} {K} {O} {M} {Y}.

PAGE No.:
DATE : / /

For K = 2.

| Itemset | Frequency | Support | |
|---|---|---|---|
| {Y, E} | 2 | 40 % | |
| {O, Y} | 2 | 40 % | |
| {K, E} | 4 | 80 % | ✓ |
| {K, O} | 3 | 60 % | ✓ |
| {m, Y} | 2 | 40 % | |
| {K, m} | 3 | 60 % | ✓ |
| {K, Y} | 3 | 60 % | ✓ |
| {O, E} | 3 | 60 % | ✓ |
| {O, m} | 1 | 20 % | |
| {m, E} | 2 | 40 % | |

For K = 3.

| Itemset | Frequency | Support | |
|---|---|---|---|
| {E, K, O} | 3 | 60 | ✓ |
| {E, K, Y} | 2 | 40 | |
| {m, E, K} | 2 | 40 | |
| {m, K, Y} | 2 | 40 | |
| {m, K, O} | 1 | 20 | |
| {m, E, Y} | 1 | 20 | |
| {m, E, O} | 1 | 20 | |
| {Y, K, O} | 2 | 40 | |
| {m, O, Y} | 1 | 20 | |
| {Y, E, O} | 2 | 40 | |

∴ Frequent Itemset = ~~{m}~~ {O, K}, {O, E} {K, Y}, {K, m}, {K, E}., {O, K, E}.

PAGE No.:
DATE : / /

→ Association Rules & Confidence.

Confidence $(A \rightarrow B) = \frac{\text{Support}(A \cup B)}{\text{Support}(A)}$.

| Association Rule | Confidence. | |
|---|---|---|
| $O \rightarrow K$ | 100 % | ✓ |
| $K \rightarrow O$ | 60 % | |
| $O \rightarrow E$ | 100 % | ✓ |
| $E \rightarrow O$ | 75 % | |
| $K \rightarrow Y$ | 60 % | |
| $Y \rightarrow K$ | 100 % | ✓ |
| $K \rightarrow M$ | 60 % | |
| $M \rightarrow K$ | 100 % | ✓ |
| $K \rightarrow E$ | 80 % | ✓ |
| $E \rightarrow K$ | 100 % | ✓ |
| $\{O, K\} \rightarrow E$ | 100 % | ✓ |
| $\{O, E\} \rightarrow K$ | 100 % | ✓ |
| $\{K, E\} \rightarrow O$ | 75 % | |

∴ Association Rules formed :-

1) $O \rightarrow K$
2) $O \rightarrow E$
3) $Y \rightarrow K$
4) $M \rightarrow K$
5) $K \rightarrow E$
6) $E \rightarrow K$
7) $\{O, K\} \rightarrow E$
8) $\{O, E\} \rightarrow K$.

Rajdhani
DATE / /

$$\text{Info}(Y) = -\sum_{i=1}^{n} P_i \log_2(P_i)$$

$$= -\left(\frac{6}{10}\log_2\left(\frac{6}{10}\right) + \frac{4}{10}\log_2\left(\frac{4}{10}\right)\right) = 0.9710$$

$$\text{Info}(\text{Temperature}, Y) = \sum_{j=1}^{v} \frac{|Y_j|}{|Y|} \times \text{Info}(Y).$$

$$= \frac{4}{10}\times\left(\frac{-1}{4}\log_2\left(\frac{1}{4}\right) - \frac{3}{4}\log_2\left(\frac{3}{4}\right)\right) + \frac{3}{10}\left(\frac{-1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right) + \frac{3}{10}\left(-\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right)$$

$$= 0.8755$$

$$\text{Info}(\text{Humidity}, Y) = \frac{4}{10}\left(-\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right) + \frac{6}{10}\left(-\frac{2}{6}\log_2\left(\frac{2}{6}\right) - \frac{4}{6}\log_2\left(\frac{4}{6}\right)\right)$$

$$= 0.9510.$$

$$\text{Info}(\text{Clouds}, Y) = \frac{5}{10}\left(-\frac{4}{5}\log_2\left(\frac{4}{5}\right) - \frac{1}{5}\log_2\left(\frac{1}{5}\right)\right) + \frac{5}{10}\left(\log_2\left(\frac{5}{5}\right) - \log_2\left(\frac{0}{5}\right)\right)$$

$$= 0.3610.$$

∴ Gain of Temperature = 0.9710 − 0.8755
= 0.0955.

Gain of Humidity = 0.9710 − 0.9510
= 0.0200

∴ Gain of Clouds = 0.9710 − 0.3610
= 0.6100.

∴ Best attribute for classification = Clouds.

Rajdhani
DATE / /

Q10). Decision tree classification using Info Gain & Gini Index.

Dataset

| Day | Temperature | Humidity | Clouds | Rain. |
|---|---|---|---|---|
| 1 | High | High | Yes | Yes |
| 2 | High | High | No | No |
| 3 | Low | High | No | No |
| 4 | medium | Low | No | No |
| 5 | High | Low | Yes | ~~No~~ Yes. |
| 6 | Medium | Low | Yes | Yes |
| 7 | Low | Low | Yes | No |
| 8 | Low | Low | No | No |
| 9 | Low | High | Yes | Yes |
| 10 | Medium. | Low | No | No. |

N = 10.

Target Class = Yes = 4/10.

Contrasting Class = No = 6/10.

A = Temperature:

| | Yes | No. | |
|---|---|---|---|
| Low | 1 | 3 | 4/10. |
| Medium | 1 | 2 | 3/10. |
| High. | 2 | 1 | 3/10. |

A = Humidity.

| | Yes | No. | |
|---|---|---|---|
| High | 2 | 2. | 4/10. |
| Low. | 2 | 4. | 6/10. |

A = Clousds.

| | Yes | No | |
|---|---|---|---|
| Yes | 4 | 1 | 5/10. |
| No. | 0 | 5 | 5/10. |

Rajdhani
DATE / /

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2$$

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

$$= 0.48.$$

$$Gini_D(Temp) = \sum_{i=1}^{m} \frac{|D_i|}{|D|} Gini(D).$$

$$= \frac{4}{10}$$

∴ Gini of Temperature.

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

$$= 1 - (0.0625 - 0.5625).$$

$$= 0.375.$$

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

$$= 1 - 0.55$$

$$= 0.45$$

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

$$= 1 - 0.55$$

$$= 0.45.$$

$$\therefore Gini(Temperature) = \frac{4}{10} \times 0.375 + \frac{3}{10} \times 0.45 + \frac{3}{10} \times 0.45$$

$$= 0.4167.$$

Rajdhani
DATE / /

2) Gini (Humidity).

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

$$= 1 - 0.5 = 0.5.$$

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

$$= 1 - 0.55 = 0.45.$$

$$\therefore \text{Gini (Humidity)} = \frac{4}{10} \times 0.5 + \frac{6}{10} \times 0.45$$

$$= 0.4667.$$

3) Gini (Clouds).

$$\text{Gini (Yes)} = 1 - \left(\left(\frac{4}{5}\right)^2 - \left(\frac{1}{5}\right)^2\right)$$

$$= 1 - 0.68 = 0.32.$$

$$\text{Gini (No)} = 1 - \left(\left(\frac{0}{5}\right)^2 - \left(\frac{5}{5}\right)^2\right)$$

$$= 0.$$

$$\therefore \text{Gini (Clouds)} = 0.5 \times 0.32 + 0.5 \times 0$$

$$= 0.16.$$

∴ Best attribute for classification = Clouds.

Rajdhani

DATE / /

Q.) Book (Page 122)

→ Pearsons correlation formula is given as:

$$r = \frac{n(\Sigma xy) - \Sigma x \Sigma y}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

| age | fat (%) |
|---|---|
| 23 | 9.5 |
| 23 | 26.5 |
| 27 | 7.8 |
| 27 | 17.8 |
| 39 | 31.4 |
| 41 | 25.9 |
| 47 | 27.4 |
| 49 | 27.2 |
| 50 | 31.2 |
| 52 | 34.2 |
| 54 | 42.5 |
| 54 | 28.8 |
| 56 | 33.4 |
| 57 | 30.2 |
| 58 | 34.1 |
| 58 | 32.9 |
| 60 | 41.2 |
| 61 | 35.7 |

Rajdhani

DATE / /

Thus,

| age (x) | % fat (y) | xy | $x^2$ | $y^2$ |
|---|---|---|---|---|
| 23 | 9.5 | 218.5 | 529 | 84.64 |
| 23 | 26.5 | 609.5 | 529 | 702.25 |
| 27 | 7.8 | 210.6 | 729 | 60.84 |
| 27 | 17.8 | 480.6 | 729 | 316.84 |
| 39 | 31.4 | 1224.6 | 1521 | 9859.6 |
| 41 | 25.9 | 1061.9 | 1621 | 670.81 |
| 47 | 27.4 | 1287.8 | 2209 | 750.76 |
| 49 | 27.2 | 1332.8 | 2401 | 739.84 |
| 50 | 31.2 | 1560 | 2500 | 973.44 |
| 52 | 34.6 | 1799.2 | 2704 | 1197.16 |
| 54 | 42.5 | 2295 | 2916 | 1806.25 |
| 54 | 28.8 | 1555.2 | 3136 | 829.44 |
| 56 | 33.4 | 1870.4 | 3136 | 1115.56 |
| 57 | 30.2 | 1721.4 | 3249 | 912.04 |
| 58 | 34.1 | 1977.8 | 3634 | 1162.81 |
| 58 | 32.9 | 1908.2 | 3634 | 1082.41 |
| 60 | 41.2 | 2472 | 3600 | 1697.44 |
| 61 | 35.7 | 2177.7 | 3721 | 1274.49 |
| Total 836 | 518.1 | 25763.2 | 41558 | 25237.06 |

$$r = \frac{18 \times (25763.2) - (836 \times 518.1)}{\sqrt{[18 \times 41558 - (836)^2] \times [18 \times 25237.06 - (518.1)^2}}$$

$$= \frac{3606}{221.69 \times 95569.79} = 0.320$$

Thus $r = 0.320$ represents an weak positive correlation between age and fat(%). Since it is positive as one variable increases other also increases. but the relationship is not strong or reliable.

ii) Web (Scribbr)

Consider data:

| $x$ | $y$ | $x^2$ | $y^2$ |
|---|---|---|---|
| 3.63 | 50.31 | 13.18 | 2819.6 |
| 3.02 | 49.7 | 9.12 | 2470.1 |
| 3.82 | 48.4 | 14.59 | 2342.6 |
| 3.42 | 54.2 | 11.7 | 2937.6 |
| 3.59 | 54.9 | 12.89 | 3014 |
| 2.87 | 43.7 | 8.24 | 1907.7 |
| 3.03 | 47.2 | 9.18 | 2227.8 |
| 3.46 | 45.2 | 11.97 | 2043 |
| 3.36 | 54.4 | 11.29 | 2959.4 |
| 3.3 | 50.4 | 10.89 | 2540.2 |

Similarly $\Sigma x^2 = 113.05$ $\Sigma x = 33.5$

$\Sigma y^2 = 25264$ $\Sigma y = 501.2$

$\Sigma xy = 1684.2$

$$r = \frac{10 \times 1684.2 - (33.5)(501.2)}{\sqrt{(10 \times (113.05) - (33.5)^2)(10 \times (252640 - 251201.4))}}$$

$$= \frac{51.8}{\sqrt{11868.45}} \approx 0.47$$

Rajdhani
DATE / /

8) Draw FP tree and find conditional pattern

i) Book (page 274)

| Tid | items |
|---|---|
| 1 | M, O, N, K, E, Y |
| 2 | D, O, N, K, E, Y |
| 3 | M, A, K, E |
| 4 | M, U, C, K, Y |
| 5 | C, O, O, K, I, E |

⇒

| item | frequency |
|---|---|
| M | 3 |
| O | 3 |
| N | 2 |
| K | 5 |
| E | 4 |
| Y | 3 |
| D | 1 |
| A | 1 |
| U | 1 |
| C | 2 |
| I | 1 |

Now consider minimum support as 3

Thus,

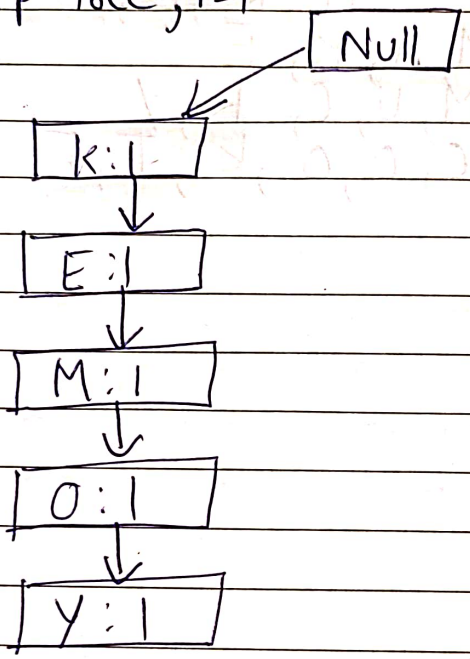
| item | freq. |
|---|---|
| K | 5 |
| E | 4 |
| M | 3 |
| O | 3 |
| N | 3 |
| Y | |

Rajdhani
DATE / /

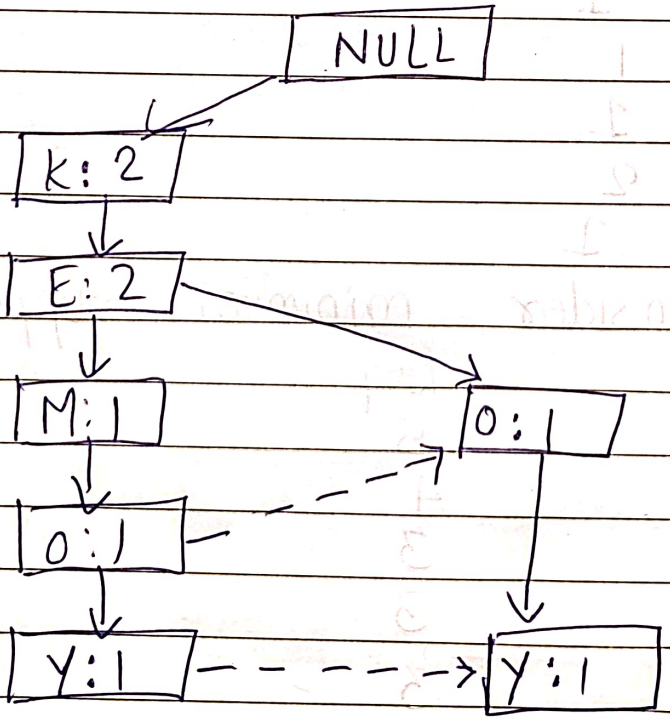
Thus ordered item set.

| | Tid |
|---|---|
| K, E, M, O, Y | 1 |
| K, E, O, Y | 2 |
| K, E, M | 3 |
| K, M, Y | 4 |
| K, E, O | 5 |

Fp tree, T=1



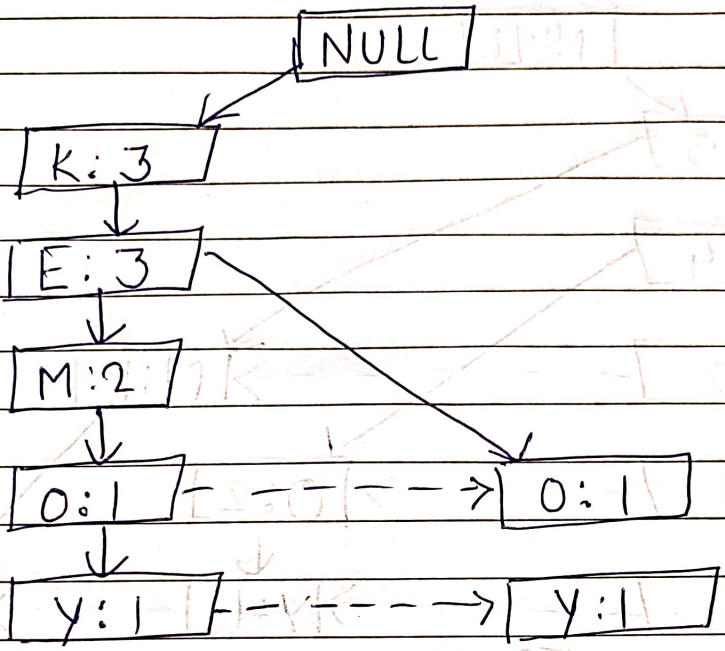

T=2



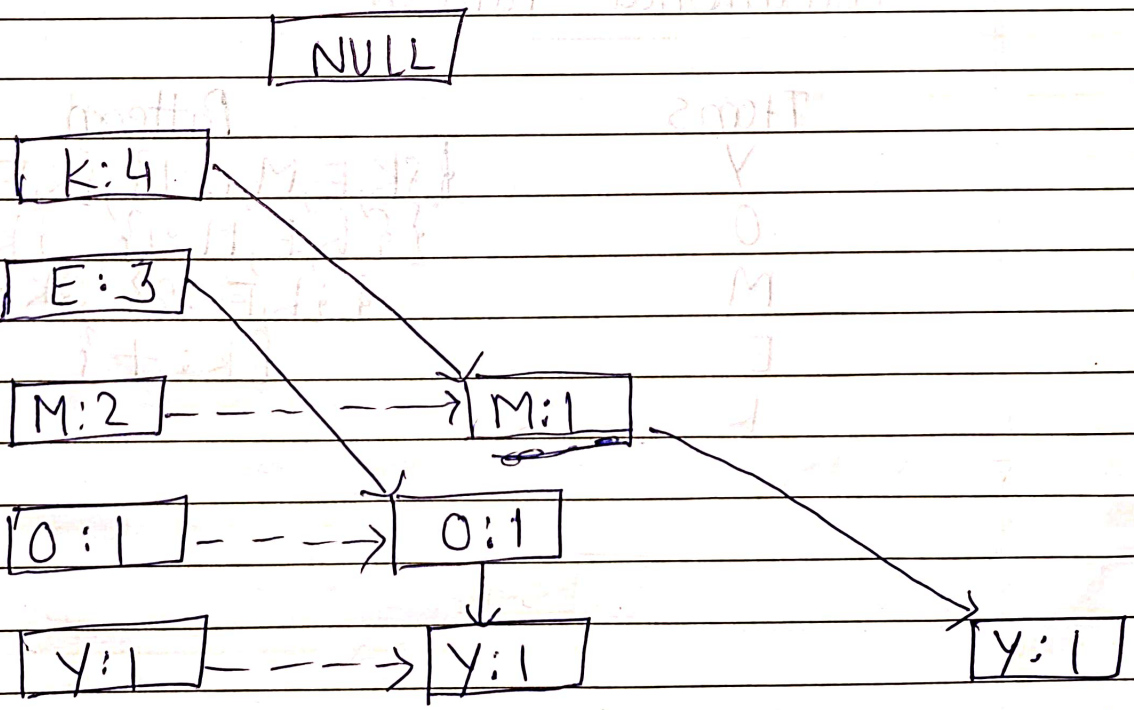
Rajdhani

DATE / /

T=3 {K, E, M}

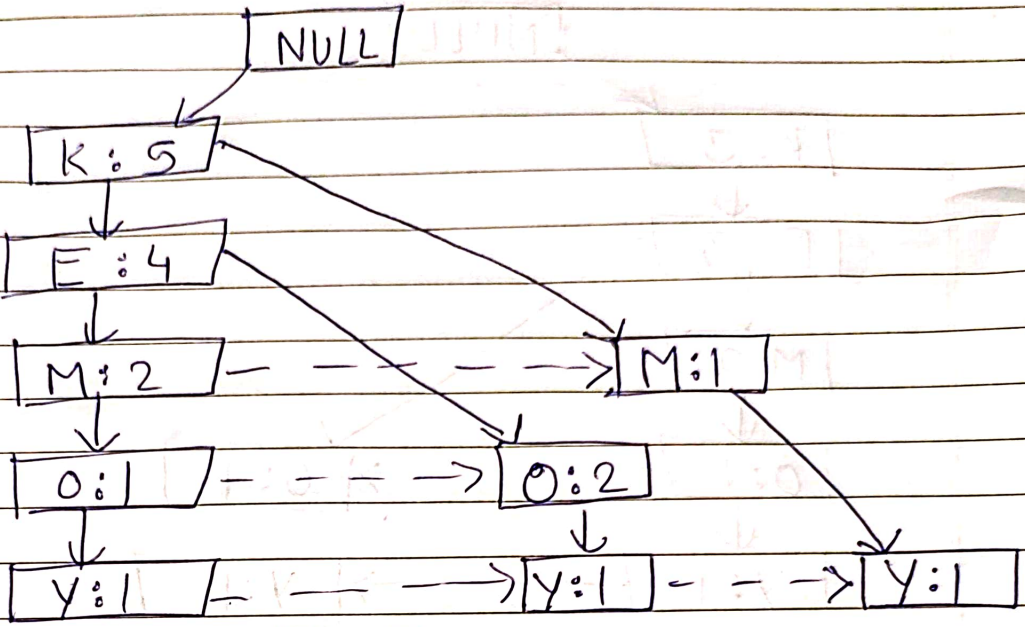



T4 {K, M, Y}



Rajdhani
DATE / /

TS {K, E, O}




FP Tree

Thus,
Conditional Pattern

| Items | Pattern |
|---|---|
| Y | {{K,E,M,O:1},{K,E,O:1},{K,M:1}} |
| O | {{K,E,M:1},{K,E:2}} |
| M | {{K,E:2},{K:1}} |
| E | {K:4} |
| K | |

Q.12 Find Linear regression coefficient on $(x, y)$ data (2-Dimensional Data)

Morgan-Kauffmann

① Predict salary with 10 years experience [356]

| $x$ years experience | $y$ salary (in ₹1000s) |
|---|---|
| 3 | 30 |
| 8 | 57 |
| 9 | 64 |
| 13 | 72 |
| 3 | 36 |
| 6 | 43 |
| 11 | 59 |
| 21 | 90 |
| 1 | 20 |
| 16 | 83 |

$$y = a + bx$$

$$a = \frac{\Sigma y \Sigma x^2 - \Sigma x \Sigma xy}{n\Sigma x^2 - (\Sigma x)^2}$$

$$b = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$$

$\Sigma x = 88 \quad \Sigma y = 554 \quad \Sigma x^2 = 1130 \quad \Sigma xy = 6134$

$$a = \frac{(554 \times 1130) - (88 \times 6134)}{10 \times 1130 - 88^2} = 24.25$$

$$b = \frac{(10 \times 6134) - (88 \times 554)}{10 \times 1130 - 88^2} = 3.54$$

∴ Regression line $y = 24.25 + (3.54)x$

for $x = 10$ $\quad y = 24.25 + 3.54 \times 10$

$= 59.65$

∴ Salary with 10 years experience is ₹ 59650

② Predict marks if one studies 10 hours

| x no. of hours | y marks obtained. |
|---|---|
| 5 | 60 |
| 4 | 45 |
| 6 | 30 |
| 8 | 80 |
| 2 | 20 |
| 11 | 95 |
| 13 | 87 |

$$y = a + bx$$

$$a = \frac{\Sigma y \Sigma x^2 - \Sigma x \Sigma xy}{n\Sigma x^2 - (\Sigma x)^2}$$

$$b = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$$

$\Sigma x = 49 \quad \Sigma x^2 = 435 \quad \Sigma y = 417 \quad \Sigma xy = 3516$

$$\therefore a = \frac{(417 \times 435) - 49 \times 3516}{(7 \times 435) - 49^2} = 14.15$$

$$b = \frac{(7 \times 3516) - (49 \times 417)}{(7 \times 435) - 49^2} = 6.49$$

$\therefore y = 14.15 + 6.49x.$

for $x = 10$ $\quad y = 14.15 + 6.49(10)$

$= 79.05$

∴ Marks with 10 hours of study is 79.05

Q.13 Perform kmeans clustering 3 steps on (x, y) data (2-Dimensional data) Morgan-Kaufmann 451

(1)

| | x | y |
|---|---|---|
| A | 2 | 10 |
| B | 2 | 5 |
| C | 8 | 4 |
| D | 5 | 8 |
| E | 7 | 5 |
| F | 6 | 4 |
| G | 1 | 2 |
| H | 4 | 9 |

(i) Select $k = 3$ i.e number of clusters to be formed.

Iteration 1.

$C_1 = A(2,10)$ $C_2 = D(5,8)$ $C_3 = G(1,2)$

| | $C_1$ | $C_2$ | $C_3$ | Point belongs to |
|---|---|---|---|---|
| A | 0 | 3.60 | 8.06 | $C_1$ |
| B | 5 | 4.24 | 3.16 | $C_3$ |
| C | 8.48 | 3.16 | 7.28 | $C_2$ |
| D | 3.60 | 0 | 7.21 | $C_2$ |
| E | 7.07 | 3.61 | 6.71 | $C_2$ |
| F | 7.21 | 4.12 | 5.38 | $C_2$ |
| G | 8.06 | 7.21 | 0 | $C_3$ |
| H | 2.24 | 1.41 | 7.62 | $C_2$ |

$\therefore C_1(A(2,10))$ $C_2(C(8,4), D(5,8), E(7,5), F(6,4), H(4,9))$

$C_3(G(1,2))$

Recompute centroids

$C_1(2,10)$ $C_2 = \left(\frac{30}{5}, \frac{30}{5}\right) \equiv C_2(6,6)$ $C_3(1,2)$

Iteration 2

| | $C_1$ | $C_2$ | $C_3$ | Points belong to |
|---|---|---|---|---|
| A | 0 | 5.65 | 8.06 | $C_1$ |
| B | 5 | 4.12 | 3.16 | $C_3$ |
| C | 8.48 | 2.83 | 7.28 | $C_2$ |
| D | 3.60 | 2.24 | 7.21 | $C_2$ |
| E | 7.07 | 1.41 | 6.71 | $C_2$ |
| F | 7.21 | 2 | 5.38 | $C_2$ |
| G | 8.06 | 6.40 | 0 | $C_3$ |
| H | 2.24 | 3.60 | 7.62 | $C_1$ |

$\therefore C_1(A(2,10), H(4,9))$ $C_2(C(8,4), D(5,8), E(7,5), F(6,4))$

$C_3(B(2,5), G(1,2))$

Recompute centroids

$C_1(3, 9.5)$ $C_2(6.5, 5.25)$ $C_3(1.5, 3.5)$

Iteration 3

| | $C_1$ | $C_2$ | $C_3$ | Point belongs to |
|---|---|---|---|---|
| A | 1.12 | 6.54 | 6.52 | $C_1$ |
| B | 4.61 | 4.51 | 1.58 | $C_3$ |
| C | 7.43 | 1.95 | 6.52 | $C_2$ |
| D | 2.5 | 3.13 | 5.7 | $C_1$ |
| E | 6.02 | 0.56 | 5.7 | $C_2$ |
| F | 6.26 | 1.35 | 4.53 | $C_2$ |
| G | 7.76 | 6.39 | 1.58 | $C_3$ |
| H | 1.12 | 4.51 | 6.04 | $C_1$ |

$C_1(A(2,10), D(5,8), H(4,9))$ $C_2(C(8,4), E(7,5), F(6,4))$
$C_3(B(2,5), G(1,2))$

Recompute centroids

$C_1(3.67, 9)$ $C_2(7, 4.33)$ $C_3(1.5, 3.5)$

| | $C_1$ | $C_2$ | $C_3$ | Point belongs to |
|---|---|---|---|---|
| A | 1.94 | 7.56 | 6.52 | $C_1$ |
| B | 4.34 | 5.04 | 1.58 | $C_3$ |
| C | 6.61 | 1.05 | 6.52 | $C_2$ |
| D | 1.66 | 4.18 | 5.7 | $C_1$ |
| E | 5.20 | 0.67 | 5.7 | $C_2$ |
| F | 5.52 | 1.05 | 4.53 | $C_2$ |
| G | 7.49 | 6.44 | 1.58 | $C_3$ |
| H | 0.33 | 5.55 | 6.04 | $C_1$ |

$\therefore C_1(A(2,10), D(5,8), H(4,9))$ $C_2(C(8,4), E(7,5), F(6,4))$
$C_3(B(2,5), G(1,2))$

No changes in cluster points.

$\therefore$ Final Centroids $C_1(3.67, 9)$ $C_2(7, 4.33)$ $C_3(1.5, 3.5)$

A, D, H belongs to $C_1$

C, E, F belongs to $C_2$

B, G belongs to $C_3$

Rajdhani
DATE / /

DM Concept & Technique - Morgan Kaufmann (page - 353)

## Bayes

(Q1)

To predict whether a customer will buy a computer given its description.

$X = (age = Youth, income = medium, student = yes, credit\text{-}rating = fair)$

[Data given in table 8.1]

$$P(yes) = \frac{9}{14} = 0.643$$

$$P(No) = \frac{5}{14} = 0.357$$

Conditional probabilities

| Age | Yes | No |
|---|---|---|
| Youth | 2/9 | 3/5 |
| Middle-aged | 4/9 | 0/5 |
| Senior | 3/9 | 2/5 |

| Student | Yes | No |
|---|---|---|
| Yes | 6/9 | 1/5 |
| No | 3/9 | 4/5 |

| Income | Yes | No |
|---|---|---|
| high | 2/9 | 2/5 |
| medium | 4/9 | 2/5 |
| low | 3/9 | 1/5 |

| Credit-Rating | Yes | No |
|---|---|---|
| fair | 6/9 | 2/5 |
| excellent | 3/9 | 3/5 |

$$P(X, Yes) = P(Youth, yes) \cdot P(medium, yes) \cdot P(stud = yes, yes) \cdot P(fair, yes) \cdot P(yes)$$

$$= \frac{2}{9} \times \frac{4}{9} \times \frac{6}{9} \times \frac{6}{9} \times \frac{9}{14}$$

$$\boxed{P(X, yes) = 0.028}$$

Rajdhani
DATE / /

$P(x, No) = P(youth, No) \cdot P(medium, no) \cdot P(stud=Yes, no) \cdot P(fair, No) \cdot P(No)$

$= \frac{3}{5} \times \frac{2}{5} \times \frac{1}{5} \times \frac{2}{5} \times \frac{5}{14}$

$P(x, No) = 0.0068$

∴ as $P(x, Yes) > P(x, No)$ so model predicts that customer will buy Computer.

Q-2] Use Bayes Theorem for finding out whether it will rain or not for a tuple x (Table 4 Datasheet)

$X = (Temperature = Medium, Humidity = Low, Clouds = Yes)$

a) $P(Rain = Yes) = \frac{4}{10}$ $P(Rain = No) = \frac{6}{10}$

conditional probabilities

| Temperature | Yes | No |
|---|---|---|
| High | 2/4 | 1/6 |
| Medium | 1/4 | 2/6 |
| Low | 1/4 | 3/6 |

| Humidity | Yes | No |
|---|---|---|
| High | 2/4 | 2/6 |
| Low | 2/4 | 4/6 |

| Clouds | Yes | No |
|---|---|---|
| Yes | 4/4 | 1/6 |
| No | 0/4 | 5/6 |

$P(x, Yes) = P(Temp=Medium, Yes) \cdot P(Humid=low, Yes) \cdot P(Clouds=Yes, Yes) \cdot P(Yes)$

$= \frac{1}{4} \times \frac{2}{4} \times \frac{4}{4} \times \frac{4}{10} = 0.05$

Rajdhani
DATE / /

$P(x, No) = P(Temp = Medium, No) \cdot P(Humid = low, No)$
$P(clouds = Yes, No) \cdot P(No)$

$= \frac{2}{6} \times \frac{4}{6} \times \frac{1}{6} \times \frac{6}{10}$

$= 0.022$

∴ as $P(x, Yes) > P(x, No)$, it will rain.

Q-3] (Table 7 Datasheet)

Using Naive Bayes Classifier predict whether a person is Defaulter or not.

X = (Marritial Status = Married, ~~Int~~ Income = 99
Home loan refund = No)

$P(Defaulter, Yes) = \frac{3}{10}$ $P(Defaulter, No) = \frac{7}{10}$

Conditional Probabilities

| Marritial Status | Yes | No |
|---|---|---|
| Married | 0/3 | 4/7 |
| Single | 2/3 | 2/7 |
| Divorced | 1/3 | 1/7 |

| Home loan refund | Yes | No |
|---|---|---|
| Yes | 0/3 | 3/7 |
| No | 3/3 | 4/7 |

Income is numerical attribute. So we use gaussian probability.

Rajdhani

DATE / /

$$P(A,C) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(A-\mu)^2}{2\sigma^2}}$$

~~Income, Play~~

Income, Yes

95

85

90

~~Income, Not Play~~

Income, No

125

100

70

120

60

220

75

$\mu_{Yes} = 90$

$\sigma_{Yes} = 4.0825$

$\mu_{No} = 110$

$\sigma_{No} = 50.4975$

$$P(99, Yes) = \frac{1}{\sqrt{2\times3.14\times4.0825^2}} \times e^{-\frac{(99-90)^2}{2\times(4.0825)^2}}$$

$= \cancel{0.0174}\ 0.0086$

$$P(99, No) = \frac{1}{\sqrt{2\times3.14\times(50.4975)^2}} \times e^{-\frac{(99-110)^2}{2\times(50.4975)^2}}$$

$= 0.0077$

$P(X, Yes) = P(Married, Yes) \cdot P(99, Yes) \cdot P(Profession, Yes) \cdot P(Yes)$

$= 0 \times 0.0086 \times 1 \times \frac{3}{10}$

$= 0$

Rajdhani

DATE / /

$P(x, No) = P(Married, No) \cdot P(99, No) \cdot P(ref = No, No) \cdot P(No)$

$= \frac{4}{7} \times 0.0077 \times \frac{4}{7} \times \frac{7}{10}$

$= 0.0018$

∴ As $P(x, No) > P(x, Yes)$ so the model classifies the person as not defaulter.

Q.14 Hierarchical Clustering Single & Complete linkages

Morgan-Kaufmann [457]

① 

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| A | 0 | | | | | |
| B | 0.71 | 0 | | | | |
| C | 5.66 | 4.95 | 0 | | | |
| D | 3.61 | 2.92 | 2.24 | 0 | | |
| E | 4.24 | 3.54 | 1.41 | 1.00 | 0 | |
| F | 3.20 | 2.50 | 2.50 | [0.50] | 1.12 | 0 |

I] Single linkage

① d(DF) = 0.50 is minimum.

∴ Merge D & F as point P & rewrite distance matrix

d(PA) = min(DA, FA) = min(3.61, 3.20) = 3.20

d(PB) = min(DB, FB) = min(2.92, 2.50) = 2.50

d(PC) = min(DC, FC) = min(2.24, 2.50) = 2.24

d(PE) = min(DE, FE) = min(1.00, 1.12) = 1.00

| | A | B | C | P | E |
|---|---|---|---|---|---|
| A | 0 | | | | |
| B | [0.71] | 0 | | | |
| C | 5.66 | 4.95 | 0 | | |
| P | 3.20 | 2.50 | 2.24 | 0 | |
| E | 4.24 | 3.54 | 1.41 | 1.00 | 0 |

d(AB) = 0.71 is minimum

∴ Merge A & B as point Q & rewrite distance matrix

d(Q, C) = min(AC, BC) = min(5.66, 4.95) = 4.95

d(QP) = min(AP, BP) = min(3.20, 2.50) = 2.50

d(QE) = min(AE, BE) = min(4.24, 3.54) = 3.54

| | | | | |
|---|---|---|---|---|
| Q | 0 | | | |
| C | 4.95 | 0 | | |
| P | 2.50 | 2.24 | 0 | |
| E | 3.54 | 1.41 | 1.00 | 0 |
| | Q | C | P | E |

③ $d(P,E) = 1.00$ is minimum

∴ Merge P & E as point R & rewrite distance matrix.

$d(QR) = \min(QP, QE) = \min(2.50, 3.54) = 2.50$

$d(CR) = \min(CP, CE) = \min(2.24, 1.41) = 1.41$

∴

| | | | |
|---|---|---|---|
| Q | 0 | | |
| C | 4.95 | 0 | |
| R | 2.50 | 1.41 | 0 |
| | Q | C | R |

④ $d(CR) = 1.41$ is minimum

∴ merge C & R as point S & rewrite distance matrix

$d(SQ) = \min(CQ, QR) = \min(4.95, 2.50) = 2.50.$

∴

| | | |
|---|---|---|
| Q | 0 | |
| S | 2.50 | 0 |
| | Q | S |

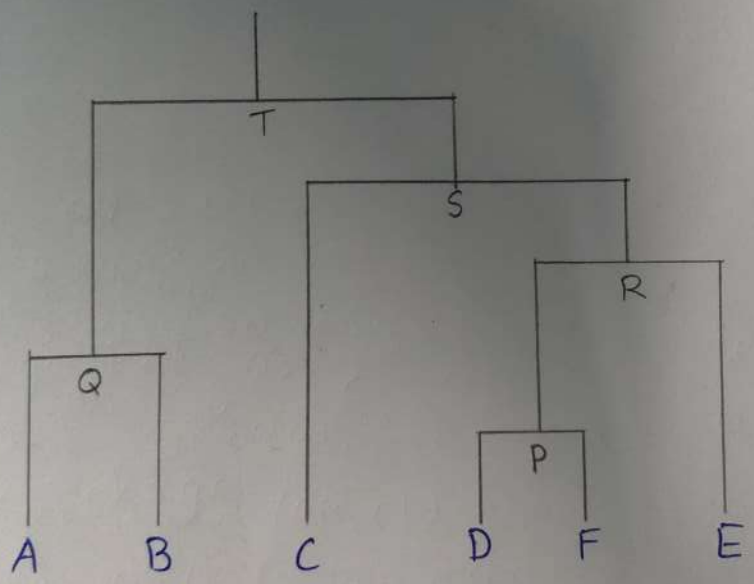
⑤ $d(QS) = 2.50$ is minimum

∴ merge Q & S as point T & rewrite distance matrix

| | |
|---|---|
| T | 0 |
| | T |

i.e one cluster final

Dendogram



II Complete linkage

① d(DF) = 0.50 is minimum.

∴ merge D & F as point P & rewrite distance matrix

d(PA) = max(DA, FD) = max(3.61, 3.20) = 3.61

d(PB) = max(DB, FB) = max(2.92, 2.50) = 2.92

d(PC) = max(DC, FC) = max(2.24, 2.50) = 2.50

d(PE) = max(DE, FE) = max(1.00, 1.12) = 1.12.

| | | | | | |
|---|---|---|---|---|---|
| A | 0 | | | | |
| B | 0.71 | 0 | | | |
| C | 5.66 | 4.95 | 0 | | |
| E | 4.24 | 3.54 | 1.41 | 0 | |
| P | 3.61 | 2.92 | 2.50 | 1.12 | 0 |
| | A | B | C | E | P |

② $d(AB) = 0.71$ is minimum

∴ merge A & B as point Q & rewrite distance matrix.

$d(QC) = \max(AC, BC) = \max(5.66, 4.95) = 5.66$

$d(QP) = \max(AP, BP) = \max(3.61, 2.92) = 3.61$

$d(QE) = \max(AE, BE) = \max(4.24, 3.54) = 4.24$

| | | | | |
|---|---|---|---|---|
| Q | 0 | | | |
| C | 5.66 | 0 | | |
| P | 3.61 | 2.50 | 0 | |
| E | 4.24 | 1.41 | 1.12 | 0 |
| | Q | C | P | E |

③ $d(PE) = 1.12$ is minimum

∴ merge P & E as point R & rewrite distance matrix

$d(RQ) = \max(PQ, EQ) = \max(3.61, 4.24) = 4.24$

$d(RC) = \max(PC, EC) = \max(2.50, 1.41) = 2.50$

| | | | |
|---|---|---|---|
| Q | 0 | | |
| C | 5.66 | 0 | |
| R | 4.24 | 2.50 | 0 |
| | Q | C | R. |

④ $d(RC) = 2.50$ is minimum

∴ merge R & C as point S & rewrite distance matrix

$d(SQ) = \max(RQ, CQ) = \max(4.24, 5.66) = 5.66.$

∴

| | | |
|---|---|---|
| Q | 0 | |
| S | 5.66 | 0 |
| | Q | S |

⑤ $d(QS) = 5.66$ is minimum.

∴ merge Q & S to form final cluster T.

T O T

Dendogram