Date
Page

Question Bank.

DMBI

Q:15 Q:1 Using the data for age given as 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 answer the following: Use z-score and decimal scaling normalization to transform the value 35 for age. The standard deviation of age is 12.94 years.

Soln: Given: $\sigma = 12.94$; $v = 35$

(i) Z-score Normalization

$$v' = \frac{v - \mu}{\sigma}$$

μ = Mean

σ = Standard Deviation

$$\mu = \frac{(13+15+16+16+19+20+20+21+22+22+25+25+25+25+30+33+33+35+35+35+35+36+40+45+46+52+70)}{27}$$

$$\mu = \frac{809}{27}$$

$$= 29.96.$$

$$\therefore v' = \frac{35 - 29.96}{12.94}$$

$$= \frac{5.04}{12.94}$$

$$\boxed{v' = 0.389}$$

(ii) Normalization by decimal scaling

$$v' = \frac{v}{10^j}$$

where j is the smallest integer such that $\max(|v'|) < 1$

$$v' = \frac{v}{10^j}$$

$$= \frac{35}{10^2}$$

$$= \frac{35}{100}$$

$$\boxed{v' = 0.35}$$

Q:26. Suppose a group of 12 sales price records has been sorted as follows:

5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215

Partition them into three bins by equal-frequency (equal-depth) partitioning.

(i) Equal frequency partitioning.

bin 1 - 5, 10, 11, 13

bin 2 - 15, 35, 50, 55

bin 3 - 72, 92, 204, 215

Q:27 Use K-Means algorithm to create 2 clusters for given set of values {2, 3, 8, 9, 12, 15, 18, 22}

Given K = 2

{2, 3, 8, 9, 12, 15, 18, 22}

Step 1: M1 = 3, M2 = 12

~~K1 = {2}~~

K1 = {2, 3}

M1 = 2 M2 = 3

K1 = {2}

K2 = {3, 8, 9, 12, 15, 18, 22}

Q:27. Q:3 Use K-Means algorithm to create 2 clusters for given set of values {2, 3, 8, 9, 12, 15, 18, 22}

Given K = 2

{2, 3, 8, 9, 12, 15, 18, 22}

Step 1: M1 = 2 M2 = 3

K1 = {2}

K2 = {3, 8, 9, 12, 15, 18, 22}

Step 2: New centroid points has to be calculated

M1 = ? M2 = ?

M1 = 2 $M2 = \frac{3+8+9+12+15+18+22}{7}$

$M2 = 12.4$

K1 = {2, 3}

K2 = {8, 9, 12, 15, 18, 22}

Step 3: Realign the centroid.

M1 = ? M2 = ?

$M1 = \frac{2+3}{2}$ $M2 = \frac{8+9+12+15+18+22}{6}$

$= 2.5$ $= 14.$

K1 = {2, 3, 8}

K2 = {9, 12, 15, 18, 22}

Step 4 : Reasign the centroid.

$M1 = ?$ $\quad M2 = ?$

$M1 = \frac{2+3+8}{3}$ $\quad M2 = \frac{9+12+15+18+22}{5}$

$= 4.3$ $\quad M2 = 15.2$

$K1 = \{2, 3, 8, 9\}$

$K2 = \{12, 15, 18, 22\}$

Step 5 : Reasign the centroid

$M1 = ?$ $\quad M2 = ?$

$M1 = \frac{2+3+8+9}{4}$ $\quad M2 = \frac{12+15+18+22}{4}$

$= 5.5$ $\quad M2 = 16.75$

$K1 = \{2, 3, 8, 9\}$

$K2 = \{12, 15, 18, 22\}$

Step 6 : Reasign the centroid

$M1 = 5.5$ $\quad M2 = 16.75$

∴ Final clusters are $K1 = \{2, 3, 8, 9\}$

$K2 = \{12, 15, 18, 22\}$

Q:42. Use K-Means algorithm to create 3 clusters for given set of values {2, 3, 8, 9, 12, 15, 18, 22}

Q:4

Given K = 3

{2, 3, 8, 9, 12, 15, 18, 22}

Step 1: M1 = 2 M2 = 3 M3 = 8

K1 = {2}

K2 = {3}

K3 = {8, 9, 12, 15, 18, 22}

Step 2: New centroid point has to be calculated.

M1 = ? M2 = ? M3 = ?

M1 = 2 M2 = 3 $M3 = \frac{8+9+12+15+18+22}{6}$

$= 14$

K1 = {2}

K2 = ~~{3}~~ {3, 8}

K3 = {9, 12, 15, 18, 22}

Step 3: Reassign the centroid

$M_1 = 2$ $M2 = \frac{3+8}{2}$ $M3 = \frac{9+12+15+18+22}{5}$

$= 5.5$ $= 15.2$

K1 = {2}

K2 = {3, 8, 9}

K3 = {12, 15, 18, 22}

Step 4: Reassign the centroid

$M1 = 2$ $\quad M2 = \frac{3+8+9}{3}$ $\quad M3 = \frac{12+15+18+22}{4}$

$= 6.66$ $\quad = 16.75$

$K1 = \{2\}$

$K2 = \{3, 8, 9\}$

$K3 = \{12, 15, 18, 22\}$

Step 5: Reassign the centroid

$M1 = 2$ $\quad M2 = 6.66$ $\quad M3 = 16.75$

∴ Final clusters are $K1 = \{2\}$

$K2 = \{3, 8, 9\}$

$K3 = \{12, 15, 18, 22\}$

Q:43. Suppose that the data for analysis including the attribute age. The age values for data tuples are (in increasing order):

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.

a) What is mean, Median, Mode, modality and Midrange of the data.

b) Show a boxplot of the data and state five number summary

c) Does outlier exits for given distribution?

a] Mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$= \frac{1}{27}(809)$$

$$\bar{x} = 29.96$$

Median $= \frac{n+1}{2}$

$= \frac{27+1}{2}$

$= 14^{th}$ value

Median $= 25$

Mode

13, 15, 19, 21, 30, 36, 40, 45, 46, 52, 70 occur once in data

16, 20, 22, 23 occur twice in data.

25, 35 occur four time in data.

Modality → Bimodal

Midrange $= \frac{\text{Min Value} + \text{max Value}}{2}$

$= \frac{13+70}{2}$

$= \frac{83}{2}$

$= 41.5$

b] $Q1 = \frac{20+20}{2}$

$Q1 = 20$

$Q3 = \frac{35+35}{2}$

$Q3 = 35$

b] $Q1 = 20$

$Q2 = 25$

$Q3 = 35$

five number summary

$= (13, 20, 25, 35, 70)$

$IQR = Q3 - Q1$

$= 35 - 20$

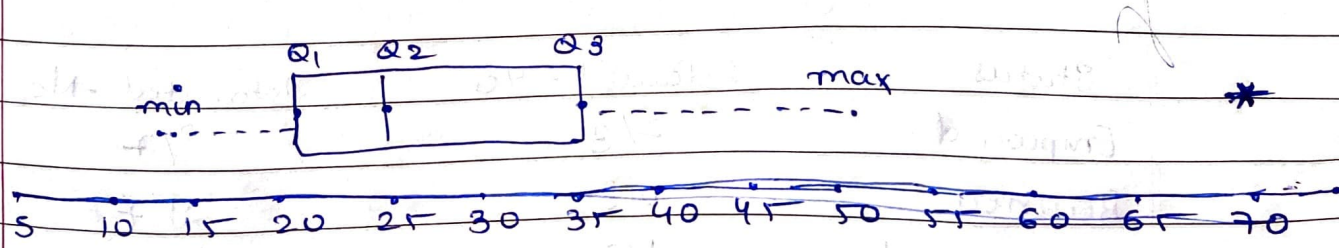
$= 15$

$Q1 - (1.5 \times IQR) = Q1 - (1.5 \times 15)$

$= -2.5$

$Q3 + (1.5 \times IQR) = 35 - (1.5 \times 15)$

$= 57.5$



c] Yes, 70

Q:6. Illustrate Naive Bayesian classification technique for the given data set. Show how can classify a new tuple, with (Homeowner = yes ; Status = Employed, Income = Average)

ID	Homeowner	Status	Income	Defaulted
1	Yes	E	H	No.
2	No	B	A	No.
3	No	E	L	No.
4	Yes	B	H	No.
5	No	UE	A	Y
6	No	B	L	N.
7	Yes	UE	H	No:
8	No	E	A	Y
9	No	B	L	N
10	No.	E	A	Y

$P(Yes) = \frac{3}{10}$

$P(No) = \frac{7}{10}$

Homeowner	Defaulted - Yes	Defaulted - No
Yes	0/3	3/7
No	3/3	4/7

Status	Defaulted - Yes	Defaulted - No
Employed	2/3	2/7
Buiness	0/3	4/7
unemployed	1/3	1/7

Income	Defaulted - Yes	Defaulted - No
High	0/3	3/7
Average	3/3	1/7
Low	0/3	3/7

$P(X' | Yes) = P(\text{Homeowner} = \text{Yes} | Yes) \times P(\text{Employed} | Yes) \times P(\text{Average} | Yes) \times P(Yes)$

$= 0 \times 2/3 \times 3/3 \times \frac{3}{10}$

$= 0$

$P(X' | No) = P(\text{Homeowner} = \text{No} | No) \times P(\text{Employed} | No) \times P(\text{Average} | No) \times P(No)$

$= \frac{4}{7} \times \frac{2}{7} \times \frac{1}{7} \times \frac{7}{10}$

$= 0.016$

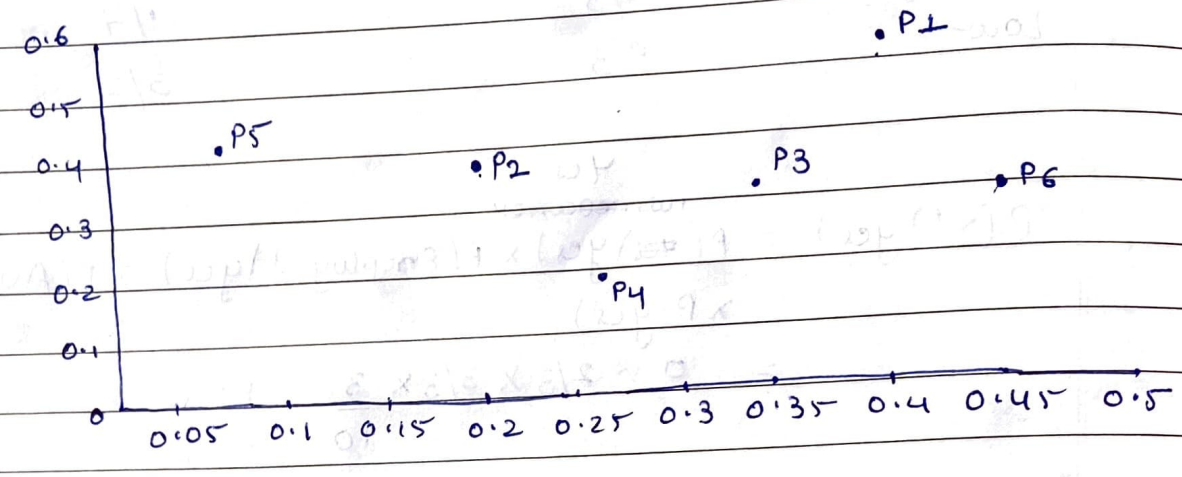
$= 0 < 0.016$

Therefore the Naive Bayesian classifier predicts Defaulted = "No" for sample X.

Q:7. Assume that the database D is given by the table below. Follow single link technique to find clusters in D. Use Euclidean distance measure.

Points	x	y
P1	0.40	0.53
P2	0.22	0.38
P3	0.35	0.32
P4	0.26	0.19
P5	0.08	0.41
P6	0.45	0.30

Step 1- Plot the given dataset in graph format for better understanding.



Step 2- Create a Distance matrix by using the Euclidean distance formula.

Distance $(P1, P2) = \sqrt{(0.4-0.22)^2 + (0.53-0.38)^2}$

$= \sqrt{0.0324 + 0.0225}$

$= \sqrt{0.0549}$

$= 0.23$

Distance $(P1, P3) = \sqrt{(0.4-0.35)^2 + (0.53-0.32)^2}$

$= \sqrt{0.0025 + 0.0441}$

$= 0.21$

Distance $(P_1, P4) = \sqrt{(0.4-0.26)^2 + (0.53-0.19)^2}$

$= 0.36$

Distance $(P_1, P_5) = \sqrt{(0.40-0.08)^2 + (0.53-0.41)^2}$

$= 0.34$

Distance $(P_1, P_6) = \sqrt{(0.4-0.45)^2 + (0.53-0.3)^2}$

$= 0.23$

Distance $(P_2, P_3) = \sqrt{(0.22-0.35)^2 + (0.38-0.32)^2}$

$= 0.15$

Distance $(P_2, P_4) = \sqrt{(0.22-0.26)^2 + (0.38-0.19)^2}$

$= 0.19$

Distance $(P_2, P_5) = \sqrt{(0.22-0.08)^2 + (0.38-0.41)^2}$

$= 0.14$

Distance $(P_2, P_6) = \sqrt{(0.22-0.45)^2 + (0.38-0.30)^2}$

$= 0.24$

Distance $(P_3, P_4) = \sqrt{(0.35-0.26)^2 + (0.32-0.19)^2}$

$= 0.15$

Distance $(P_3, P_5) = \sqrt{(0.35-0.08)^2 + (0.32-0.41)^2}$

$= 0.28$

Distance $(P_3, P_6) = \sqrt{(0.35-0.45)^2 + (0.32-0.30)^2}$

$= 0.10$

Distance $(P_4, P_5) = \sqrt{(0.26-0.08)^2 + (0.19-0.41)^2}$

$= 0.28$

Distance $(P_4, P_6) = \sqrt{(0.26-0.45)^2 + (0.19-0.30)^2}$

$= 0.21$

Distance $(P_5, P_6) = \sqrt{(0.08-0.45)^2 + (0.41-0.30)^2}$

$= 0.38$

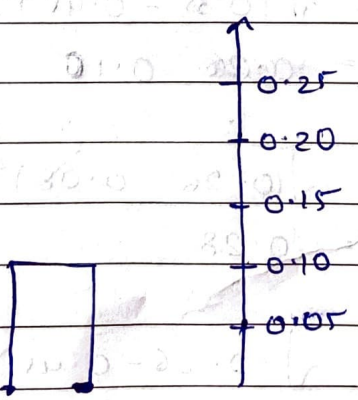
Based on the above computed value Distance Matrix can be formed as follows.

	P1	P2	P3	P4	P5	P6
P1	0					
P2	0.23	0				
P3	0.21	0.15	0			
P4	0.36	0.29	0.15	0		
P5	0.34	0.14	0.28	0.28	0	
P6	0.23	0.24	0.10	0.21	0.38	0

Step 3 : Find the minimum value element from the distance matrix

The minimum value element is (P3, P6) with a value is 0.10 - 1st cluster (P3, P6)

Step 4 - Now Recalculate or update the distance matrix for the cluster (P3, P6)



Step 4 - Now Recalculate or update the distance matrix for the cluster (P3, P6).

Min [dist (point 1, point 2)]

min [Dist ((P3, P6), P1)] = min (0.22, 0.23) = 0.22
min [Dist ((P3, P6), P2)] = min (0.14, 0.24) = 0.14
min [Dist ((P3, P6), P4)] = min (0, 0.21) = 0
min [Dist ((P3, P6), P5)] = min (0, 0.38) = 0

Therefore, the updated distance matrix looks as follows:

	P1	P2	P3, P6	P4	P5
P1	0				
P2	0.23	0			
P3, P6	0.21	0.15	0		
P4	0.36	0.19	0.15	0	
P5	0.34	0.14	0.28	0.28	0

Step 5 - Repeat steps 3 and 4

The min value element is (P2, P5) with a value is 0.14 - 2nd Cluster (P2, P5)

Recalculate or update the distance matrix for the cluster (P2, P6).

Min [dist (point 1, point 2)]

min [Dist ((P2, P5), P1)] = min (0.23, 0.34)] = 0.23
min [Dist ((P2, P5), P2)] = min (0, 0.14) = 0

Similarly, perform for (P3, P6) and P4.

Therefore, the updated Distance matrix looks as follows.

	P1	P2, P5	P3, P6	P4
P1	0			
P2, P5	0.23	0		
P3, P6	0.21	0.15	0	
P4	0.36	0.19	0.15	0

Step 6 - Repeat steps 3 & 4

The minimum value element is (P2, P5, P3, P6) with a value 0.15 - 3rd cluster (P2, P5, P3, P6)

Here 2 values are the same then the first element is choosen as the minimum value element.

Recalculate or update the distance matrix for the cluster (P2, P5, P3, P6)

Min [dist (point 1, point 2)]

Min [dist ((P2, P5, P3, P6), P1)] = min [dist (0.23, 0.21)]

= 0.21

& Min [dist ((P2, P5, P3, P6), P4)] = 0.

	P1	P2, P5, P3, P6	P4
P1	0		
P2, P5, P3, P6	0.21	0	
P4	0.36	0.15	0

Step 7 -

Min [dist (point 1, point 2)]

Min [dist ((P2, P5, P3, P6, P4), P1)] = min [dist [0.21, 0.36]

= 0.21

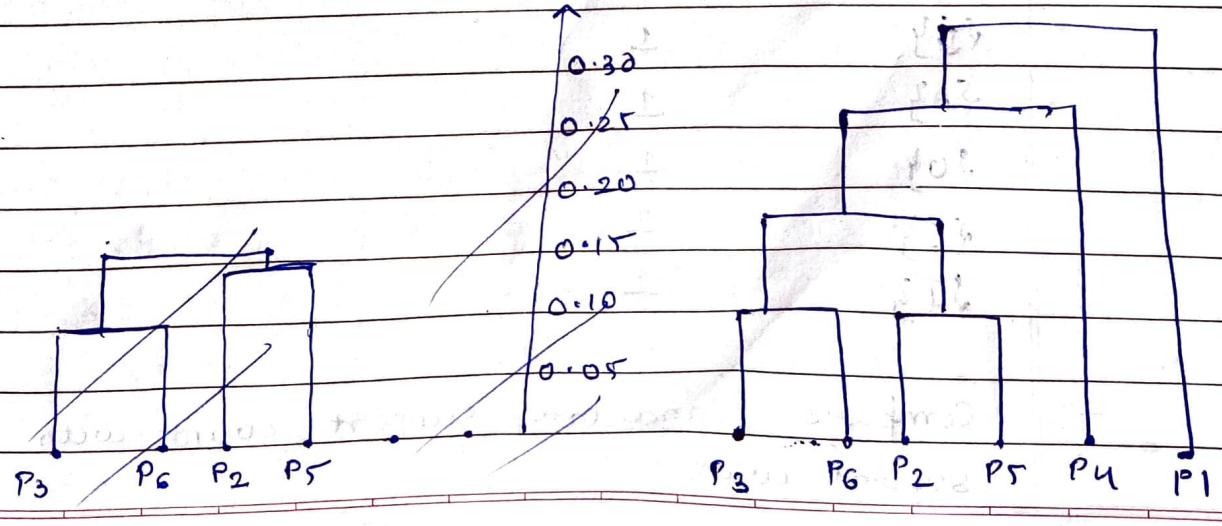
	P1	P2, P5, P3, P6, P4
P1	0	
P2, P5, P3, P6, P4	0.21	0.

Step 8 - Repeat the step 3 & 4.

The minimum value element is (P2, P5, P3, P6, P4, P1) with a value is 0.22 - 5^{th} cluster (P2, P5, P3, P6, P4, P1)

In this step, only 1 value is remaining so it is by default cluster.

Step 9 - Dendogram



Q:8. Consider the following database with minimum support count = 60%. find all the frequent itemset using apriori algorithm and also generate strong association rules if minimum confidance = 50%.

T-ID	Items bought
T1	M, O, N, K, E, Y
T2	D, O, N, K, E, Y
T3	M, A, K, E
T4	M, U, C, K, Y
T5	C, O, O, K, I, E

Min support count = $\frac{60 \times 5}{100}$

= 3

step 1 : Generating 1-itemset frequent pattern.

C1

Itemset	support
{M}	3
{O}	4
{N}	2
{K}	5
{E}	4
{Y}	3
{D}	1
{A}	1
{U}	1
{C}	2
{I}	1

Compare candidate support count with minimum support count.

Date
Page

L1

Itemset	Support Count
{M}	3
{O}	4
{K}	5
{E}	4
{Y}	3

Step 2 : Generate C2 - Itemset frequent Pattern.

C2.

Item set	Support Count
{M, O}	1
{M, K}	3
{M, E}	2
{M, Y}	2
{O, K}	3
{O, E}	3
{O, Y}	2
{K, E}	4
{K, Y}	3
{E, Y}	2

Compare Candidate support count with minimum support count

L2

Item set	Support Count
{M, K}	3
{O, K}	3
{O, E}	3
{K, E}	4
{K, Y}	3

Step 3 - Generating 3 - itemset Frequency Patte

C3

Itemset	Support Count
{MOK}	1
{OKE}	3
{KE4}	2

Compare candidate support count with minimu support count

L3

Itemset	Support Count
{OKE}	3

Q:9. Suppose a hospital tested the age and body fa data for 18 randomely selected adults with the followins result.

age	23	23	27	27	39	41	47	49
% fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2

age	52	54	54	56	57	58	58	60
% fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2

(a) Calculate the mean, median and standard deviation of age and % fat

(b) Draw the boxplot for age.

(a) Mean (age) $= \frac{1}{n}\sum_{i=1}^{n} x_i$

$= \frac{1}{18} \times 836$

$= 46.44$

Mean (% fat) $= \frac{1}{n}\sum_{i=1}^{n} x_i$

$= \frac{1}{18} \times 518$

$= 28.77$

Median (age) $= \frac{50+52}{2}$

$= 51$

Median (% fat) $= \frac{30.2+31.2}{2}$

$= 30.2$

% fat in increasing order.

7.8, 9.5, 17.8, 25.9, 26.5, 27.2, 27.4, 28.8, 30.2, 31.2, 31.4, 32.9, 33.4, 34.1, 34.6, 35.7, 41.2, 42.5

Standard Deviation of age.

Age	$(x_i - \mu)$	$(x_i - \mu)^2$
23	-23.44	549.4336
23	-23.44	549.4336
27	-19.44	377.9136
27	-19.44	377.9136
39	-7.44	55.3536
41	-5.44	29.5936
47	0.56	0.3136
49	2.56	6.5536
50	3.56	12.6736
52	5.56	30.9136
54	7.56	57.1536
54	7.56	57.1536
56	9.56	91.3936
57	10.56	111.5136
58	11.56	133.6336

Date
Page
28.22

58	11.56	133.6336
60	13.56	183.8736
61	14.56	211.9936

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

$$= \frac{2970.448}{18}$$

$$\sigma^2 = 165.02$$

Standard Deviation

$$\sigma = \sqrt{165.02}$$

$$\sigma_{age} = 12.84$$

Standard Deviation of % fat

% fat	$(x_i - \mu)$	$(x_i - \mu)^2$
7.8	-21	441
9.5	-19.3	372.49
17.8	11	121
25.9	-2.9	8.41
26.5	2.3	5.29
27.2	-1.6	2.56
27.4	-1.4	1.96
28.8	0	0
30.2	1.4	1.96
31.2	2.4	5.76
31.4	2.6	6.76

32.9	4.1	16.81
33.4	4.6	21.16
34.1	5.3	28.02
34.6	5.8	33.64
35.7	6.9	47.61
41.2	12.4	153.76
42.5	13.2	187.69

$$\sigma^2 = \frac{1455.95}{18}$$

$$= 80.886$$

$$\sigma \text{ \% fat} = \sqrt{80.886}$$

$$= 8.99 \approx 9$$

(b) Draw boxplot for age.

Q1 = 39

Q2 = 51

Q3 = 57

min = 23

max = 61

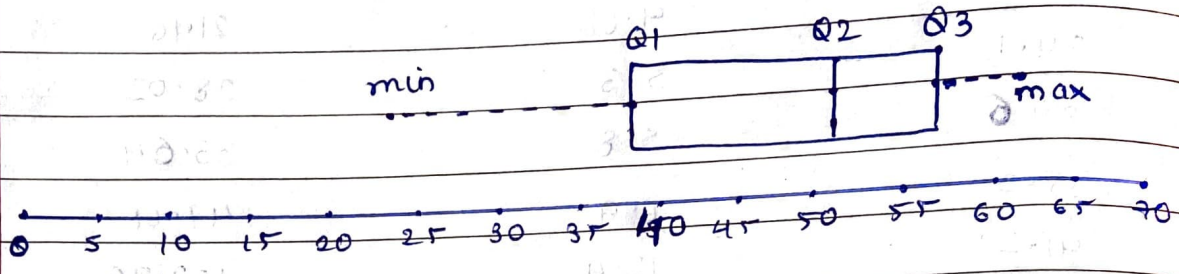
IQR = Q3 − Q1

= 57 − 39

= 18

Q1 − (1.5 × IQR) = 39 − (1.5 × 18) = 12

Q3 + (1.5 × IQR) = 57 + (1.5 × 18) = 84



Q:10. Consider the fouwing database with minimum support count = 50%. Find all the frequent itemset using apriori algorithm and also generate strong association rules if minimum confidence = 50%.

TID	List of ItemIDs
1	A, B, D
2	A, D
3	A, C
4	B, D, E, F

$$\text{Min Support count} = \frac{50}{100} \times 4$$

$$= 2$$

Step 1 - Generating 1 - itemset frequency pattern

C1

itemset	sup
{A}	3
{B}	2
{D}	3
{C}	1
{E}	1
{F}	1

L1 →

itemset	sup
{A}	3
{B}	2
{D}	3

Step 2 - Generating 2 - itemset frequency pattern

C2

itemset	sup
{A, B}	1
{A, D}	2
{B, D}	2

$\xrightarrow{L2}$

itemset	sup
{A, D}	2
{B, D}	2

Step 3 - Generating 3 - itemset frequency pattern

C3

itemset	sup
{A, B, D}	1

Therefore frequent item sets are

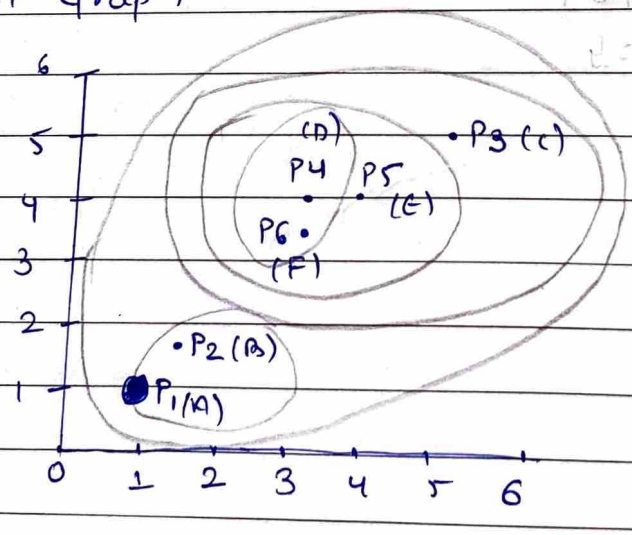
L1 = {A, D}

L2 = {B, D}

Q11. Suppose we have six objects (with name A, B, C, D, E and F) and each object have two measured features [x1 and x2] Apply single linkage clustering and Draw dendogram

object	Attribut x1	Attribute x2
A	1	1
B	1.5	1.5
C	5	5
D	3	4
E	4	4
F	3	3.5

Step 1 Plot Graph



Step2. Creating distance matrix by using the Euclidean distance formula.

Distance (P1, P2) =

Distance (A, B) = $\sqrt{(1-1.5)^2 + (1-1.5)^2}$

= 0.71

Date
Page

Distance (A, C) = $\sqrt{(1-5)^2 + (1-5)^2}$

= 5.66

Distance (A, D) = $\sqrt{(1-3)^2 + (1-4)^2}$

= 3.61

Distance (A, E) = $\sqrt{(1-4)^2 + (1-4)^2}$

= 4.24

Distance (A, F) = $\sqrt{(1-3)^2 + (1-3.5)^2}$

= 3.20

Distance (B, C) = $\sqrt{(1.5-5)^2 + (1.5-5)^2}$

= 4.95

Distance (B, D) = $\sqrt{(1.5-3)^2 + (1.5-4)^2}$

= 2.92

Distance (B, E) = $\sqrt{(1.5-4)^2 + (1.5-4)^2}$

= 3.54

Distance (B, F) = $\sqrt{(1.5-3)^2 + (1.5-3.5)^2}$

= 2.50

Distance (C, D) = $\sqrt{(5-3)^2 + (5-4)^2}$

= 2.24

Distance (C, E) = $\sqrt{(5-4)^2 + (5-4)^2}$

= 1.41

Distance (C, F) = $\sqrt{(5-3)^2 + (5-3.5)^2}$

= 2.50

Distance (D, E) = $\sqrt{(3-4)^2 + (4-4)^2}$

= 1.00

Distance (D, F) = $\sqrt{(3-3)^2 + (4-3.5)^2}$

= 0.5

Distance (E, F) = $\sqrt{(4-3)^2 + (4-3.5)^2}$

= 1.12

Based on the above computed value distance matrix can be formed as follows

	A	B	C	D	E	F
A	0					
B	0.71	0				
C	0.66	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

Step 3 → min value = 0.5 Cluster (D, F)

d(D, F) → A) = min (3.61, 3.20)

= 3.20

D(D, F) → B) = min (2.92, 2.50)

= 2.5

d(D, F) → C) = min (2.24, 2.5)

= 2.24

d(E) → D, F) = min (1, 1.12)

= 1

	A	B	C	D, F	E
A	0				
B	0.71	0			
C	5.66	4.95	0		
D, F	3.20	2.5	2.24	0	
E	4.24	3.54	1.41	1	0

Step 4 → min value = 0.71 cluster (A, B)

$d_{C \to (A,B)}$ $d((A,B) \to C) = \min(5.66, 4.95)$

$= 4.95$

$d_{(D,F) \to (A,B)}$ $d((A,B) \to (D,F)) = \min(3.20, 2.50)$

$= 2.50$

$d_{(E) \to (A,B)}$ $d((A,B) \to E) = \min(4.24, 3.54)$

$= 3.54$

	A, B	C	D, F	E
A, B	0			
C	4.95	0		
D, F	2.5	2.24	0	
E	3.54	1.41	1.0	0

Step 5 → min value = 1 cluster ((D, F), E)

$d \to (D,F)E \to AB = \min(2.5, 3.54)$

$= 2.5$

$d \to (D,F)E \to C = \min(2.24, 1.41)$

$= 1.41$

	A, B	C	(D, F)E
A, B	0		
C	4.95	0	
(D,F)E	2.5	1.41	0

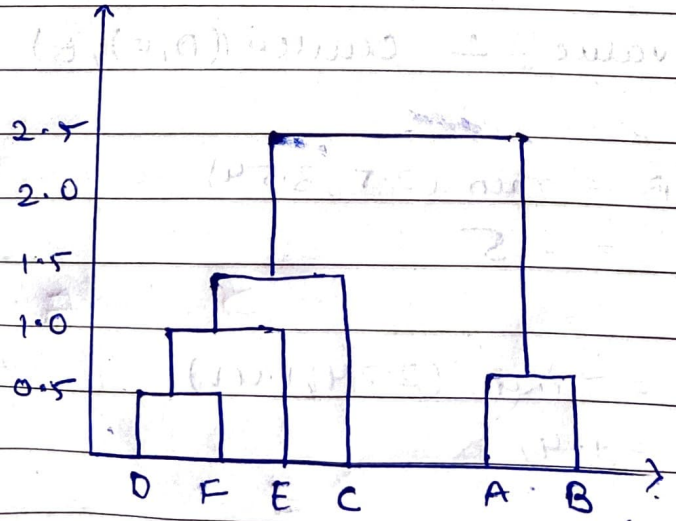
d → (D, F)E → AB = min (2.50, 3.54)
= 2.5

Step 6 → min value → 1.41 cluster (((D, F), E), C)

	A, B	(((D, F), E), C)
(A, B)	0	
(((D,F), E), C)	2.5	0

Step 7 → min value → 2.5 cluster ((((D, F), E), C, (A, B))

Dendogram



Q:12 Find all the frequent itemset and also generate association rule using Apriori algorithm for the given transaction. Assume minimum support be 50% and minimum confidence be 80%.

TID	items
100	A, C, D
200	B, C, E
300	A, B, C, E
400	B, E

Min Support count = $\frac{50}{100} \times 4$

$= 2$.

TID	items
100	A C D
200	B C E
300	A B C E
400	B E

Step 1 → Generating 1-itemset frequency pattern

C1

Items	Sup
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

→ (4)

items	Sup
{A}	2
{B}	3
{C}	3
{E}	3

C_2

itemset	sup
{A, B}	1
{A, C}	2
{A, E}	1
{B, C}	2
{B, E}	3
{C, E}	2

→ L_2

itemset	sup
{A, C}	2
{B, C}	2
{B, E}	3
{C, E}	2

C_3

itemset	sup
{A, B, C}	1
{B, C, E}	2

→ L_3

itemset	sup
{B, C, E}	2

Therefore frequent item sets are
$L = \{B, C, E\}$

Now, for strong association rule
Generate non-empty subset for L
$S = \{\{B\}, \{C\}, \{E\}, \{B, C\}, \{B, E\}, \{C, E\}\}$

Confidence $(X \rightarrow Y) = P(Y|X) = P(X \cup Y) / P(X)$

Association Rule	Support	Confidance	confidance %
{B} → {C, E}	2	2/3	66.66
{C} → {B, E}	2	2/3	66.66
{E} → {B, C}	2	2/3	66.66
{B, C} → {E}	2	2/2	100
{B, E} → {C}	2	2/3	66.66
{C, E} → {B}	2	2/2	100

Therefore strong association rules are

{B, C} → {E}

{C, E} → {B}

Q:13. Predict a class label for an unknown sample using Naive Bayesian classification on the following training dataset from all electronic customer database.

X' = {age = "<=30", income = "medium", Student = "yes", credit rating = fair}

Age	Income	student	credit_rating	buys_computer
<=30	High	N	Fair	No
<=30	H	N	Excellent	N
31...40	H	N	F	Yes
>40	Medium	N	F	Yes
>40	LOW	Y	F	Yes
>40	L	Y	E	N
31...40	L	Y	E	Yes
<=30	M	N	F	N
<=30	L	Y	F	Yes
>40	M	Y	F	Yes
<=30	M	Y	E	Yes
31...40	M	N	E	Yes
31...40	H	Y	F	Yes
>40	M	N	F	N

P(buy_computer = yes) = 9/14

P(buys_computer = No) = 5/14

age	buys-computer = yes	buys-computer = No
<=30	2/9	3/5
31..40	4/9	0/5
>40	3/9	2/5

Income	yes	No
High	2/9	2/5
Low	3/9	1/5
Medium	4/9	2/5

Student	yes	No
Yes	6/9	1/5
No	3/9	4/5

Credit_rating	yes	No
fair	6/9	2/5
Excellent	3/9	3/5

Map Rule

$$P(X'|yes) = P(age = "<=30"/yes) * P(income = "medium"/yes) * P(student = "yes"/yes) * P(credit\_rating = "fair"/yes) * P(yes)$$

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

$$= 0.02821$$

$P(x') \text{ No} = P(\text{age} = \text{"<=30"} / \text{No}) *$
$P(\text{income} = \text{"medium"} / \text{No}) *$
$P(\text{Student} = \text{"yes"} / \text{No}) *$
$P(\text{credit\_rating} = \text{"fair"} / \text{No}) *$
$P(\text{No})$

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

$= 0.007$

Since $0.028 > 0.007$, therefore the naive Bayes classifier predicts bayes_computer = "yes" for sample X.

Q:14. Construct a decision tree using classification algorithm and decide on which day you can play tennis.

Day	Outlook	Temp	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	S	H	H	Strong	N
D3	Overcast	H	H	W	Yes
D4	Rain	Mild	H	W	Y.
D5	R	cool	Normal	W	Y.
D6	R	C	N	S	N
D7	O	C	N	S	Y
D8	S	M	H	W	N
D9	S	C	N	W	Y.
D10	R	M	N	W	Y.
D11	S	M	N	S	Y
D12	O	M	H	S	Y.
D13	O	H	N	W	Y
D14	R	M	H	S	N