# BIG DATA HW 4A

Nagabharan Nagendran nxn141730

Q1>

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| A | | 3 | | 1 | 2 | 3 |
| B | | 5 | 4 | | 4 | |
| C | | | 2 | 5 | | 1 |
| D | 4 | | 1 | 4 | | 2 |

a> (A, B): I = 2, U = 5
Similar = 2/5
Distance = 3/5

(A, C): I = 2, U = 5
Similar = 2/5
Distance = 3/5

(A, D): I = 3, U = 5
Similar = 3/5
Distance = 2/5

(B, C): I = 1, U = 5
Similar = 1/5
Distance = 4/5

(B, D): I = 1, U = 6
Similar = 1/6
Distance = 5/6

(C, D): I = 2, U = 5
Similar = 2/5
Distance = 3/5

b) $(A,B)$:

$$\frac{3\times5 + 2\times4}{\sqrt{3^2+1^2+2^2+3^2}\times\sqrt{5^2+4^2+4^2}}$$

$$= 23/(4.8\times7.55) = 0.63$$

$(A,C)$:

$$\frac{1\times5 + 1\times3}{\sqrt{3^2+1^2+2^2+3^2}\times\sqrt{2^2+5^2+1^2}}$$

$$= 8/(4.8\times5.48) = 0.304$$

$(A,D)$:

$$\frac{1\times3 + 1\times4 + 2\times3}{\sqrt{3^2+1^2+2^2+3^2}\times\sqrt{4^2+1^2+4^2+2^2}}$$

$$= 13/(4.8\times6.08) = 0.445$$

$(B,C)$:

$$\frac{2\times4}{\sqrt{5^2+4^2+4^2}\times\sqrt{2^2+5^2+1^2}}$$

$$= 8/(7.55\times5.48) = 0.193$$

$(B,D)$:

$$\frac{1\times5}{\sqrt{5^2+4^2+4^2}\times\sqrt{4^2+1^2+4^2+2^2}}$$

$$= 5/(7.55\times6.08) = 0.109$$

$(C,D)$:

$$\frac{4\times5 + 2\times1}{\sqrt{2^2+5^2+1^2}\times\sqrt{4^2+1^2+4^2+2^2}}$$

$$= 22/(5.48\times6.08) = 0.66$$

c> By replacing 3,4,5 stars as 1 & 1,2 stars as 0 we get

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| A | | 1 | | | | 1 |
| B | | 1 | 1 | | 1 | |
| C | | | | 1 | | |
| D | 1 | | | 1 | | |

Iteration 1 - Distances

| (1,2) | (1,3) | (1,4) | (1,5) | (1,6) |
|---|---|---|---|---|
| 1 | 1 | 1/2 | 1 | 1 |

| (2,3) | (2,4) | (2,5) | (2,6) |
|---|---|---|---|
| 1/2 | 1 | 1/2 | 1/2 |

| (3,4) | (3,5) | (3,6) |
|---|---|---|
| 1 | 0 | 1 |

| (4,5) | (4,6) |
|---|---|
| 1 | 1 |

| (5,6) |
|---|
| 1 |

Combining 3 & 5 to form a cluster (3,5)

Iteration 2 - Distances

| (1,2) | (1,35) | (1,4) | (1,6) |
|---|---|---|---|
| 1 | 1 | 1/2 | 1 |

(2,35) (2,4) (2,6)
1/2 1 1/2

(4,35) (4,6)
1 1

(6,35)
1

Combining 1 & 4 to form cluster(1,4)

Iteration 3 → Distances

(14,2) (14,35) (14,6)
1 1 1

(2,35) (2,6)
1/2 1/2

(6,35)
1

∴ The final 4 clusters are

<1,4> 2 <3,5> 6

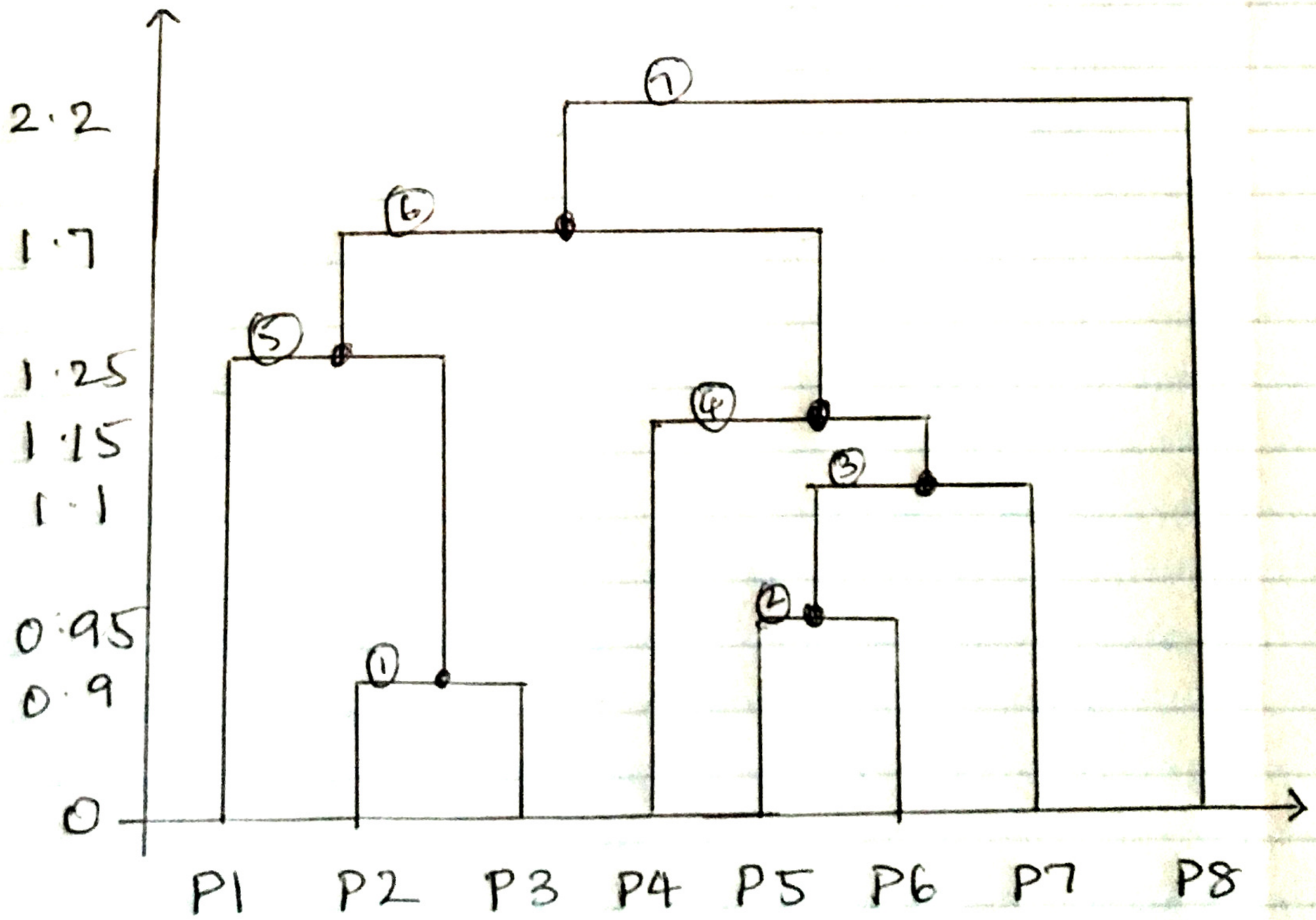
Q27

a.

| | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 |
|---|---|---|---|---|---|---|---|---|
| P1 | 0 | 1.25 | 2.15 | 3.85 | 5 | 5.95 | 7.05 | 9.25 |
| P2 | | 0 | 0.9 | 2.6 | 3.75 | 4.7 | 5.8 | 8 |
| P3 | | | 0 | 1.7 | 2.85 | 3.8 | 4.9 | 7.1 |
| P4 | | | | 0 | 1.15 | 2.1 | 3.2 | 5.4 |
| P5 | | | | | 0 | 0.95 | 2.05 | 4.25 |
| P6 | | | | | | 0 | 1.1 | 3.3 |
| P7 | | | | | | | 0 | 2.2 |
| P8 | | | | | | | | 0 |

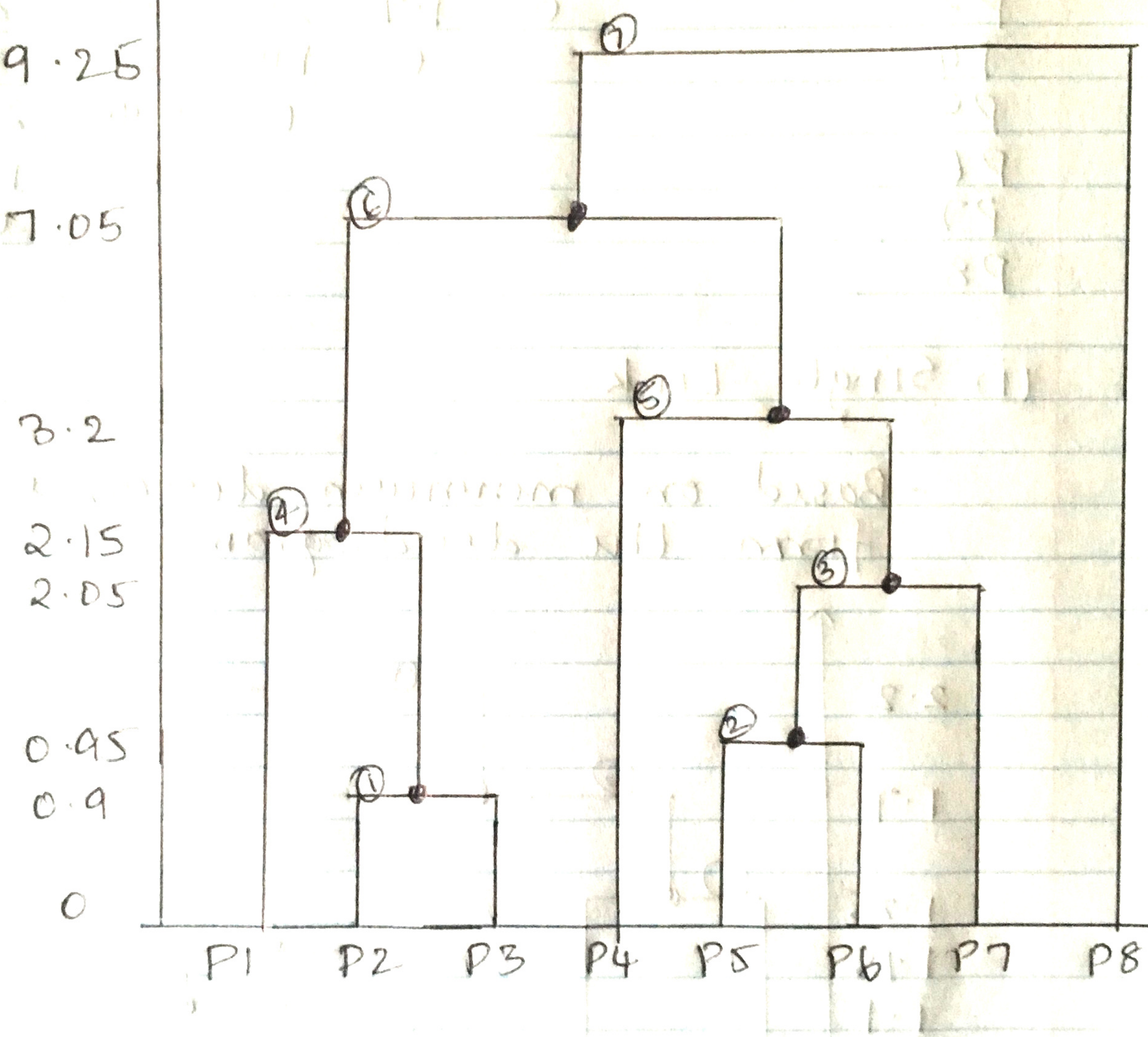
(i) Single Link

- Based on minimum distance we form the dendrogram



(ii) Complete Link

- Based on maximum distance we form the dendrogram



b) (i) $d\{(1,2,3,4),(5,6,7,8)\}$

$$= \min\begin{pmatrix} (1,5) & (1,6) & (1,7) & (1,8) \\ (2,5) & (2,6) & (2,7) & (2,8) \\ (3,5) & (3,6) & (3,7) & (3,8) \\ (4,5) & (4,6) & (4,7) & (4,8) \end{pmatrix}$$

$$= \min\begin{pmatrix} 5 & 5.95 & 7.05 & 9.25 \\ 3.75 & 4.7 & 5.8 & 8 \\ 2.85 & 3.8 & 4.9 & 7.1 \\ 1.15 & 2.1 & 3.2 & 5.4 \end{pmatrix} = 1.15$$

(ii) $\max\begin{pmatrix} (1,5) & (1,6) & (1,7) & (1,8) \\ (2,5) & (2,6) & (2,7) & (2,8) \\ (3,5) & (3,6) & (3,7) & (3,8) \\ (4,5) & (4,6) & (4,7) & (4,8) \end{pmatrix}$

$$= \max\begin{pmatrix} 5 & 5.95 & 7.05 & 9.25 \\ 3.75 & 4.7 & 5.8 & 8 \\ 2.85 & 3.8 & 4.9 & 7.1 \\ 1.15 & 2.1 & 3.2 & 5.4 \end{pmatrix} = 9.25$$

(iii) $\text{avg}\begin{pmatrix} (1,5) & (1,6) & (1,7) & (1,8) \\ (2,5) & (2,6) & (2,7) & (2,8) \\ (3,5) & (3,6) & (3,7) & (3,8) \\ (4,5) & (4,6) & (4,7) & (4,8) \end{pmatrix}$

$$= \text{avg}\begin{pmatrix} 5 & 5.95 & 7.05 & 9.25 \\ 3.75 & 4.7 & 5.8 & 8 \\ 2.85 & 3.8 & 4.9 & 7.1 \\ 1.15 & 2.1 & 3.2 & 5.4 \end{pmatrix}$$

$$= 80/16 = 5$$

Iteration 2-

| P1C1 | P2C1 | P3C1 | P4C1 | P5C1 |
|---|---|---|---|---|
| $\sqrt{3^2+.5^2}$ | $\sqrt{2^2+2.5^2}$ | $\sqrt{1^2+5.5^2}$ | $\sqrt{1.5^2}$ | $\sqrt{1^2+5.5^2}$ |
| 3.04 | 3.2 | 5.59 | 1.5 | 5.59 |

| P6C1 | P7C1 | P8C1 |
|---|---|---|
| $\sqrt{3^2+5.5^2}$ | $\sqrt{3^2+.5^2}$ | $\sqrt{4^2+1.5^2}$ |
| 6.27 | 3.04 | 4.27 |

| P1C2 | P2C2 | P3C2 | P4C2 | P5C2 |
|---|---|---|---|---|
| $\sqrt{3.83^2+3.17^2}$ | $\sqrt{2.83^2+1.17^2}$ | $\sqrt{1.83^2+1.83^2}$ | $\sqrt{.83^2+2.17^2}$ | $\sqrt{.17^2+1.83^2}$ |
| 4.97 | 3.06 | 2.59 | 2.32 | 1.84 |

| P6C2 | P7C2 | P8C2 |
|---|---|---|
| $\sqrt{2.17^2+1.83^2}$ | $\sqrt{2.17^2+4.17^2}$ | $\sqrt{3.17^2+2.17^2}$ |
| 2.84 | 4.7 | 3.84 |

Cluster 1- {1,4,7} Cluster 2- {2,3,5,6,8}

C1 - (5,2)
C2 - (6,5.6)

Iteration 3-

| P1C1 | P2C1 | P3C1 | P4C1 | P5C1 | P6C1 |
|---|---|---|---|---|---|
| $\sqrt{3^2}$ | $\sqrt{2^2+2^2}$ | $\sqrt{1^2+5^2}$ | $\sqrt{1^2}$ | $\sqrt{1^2+5^2}$ | $\sqrt{3^2+5^2}$ |
| 3 | 2.83 | 5.1 | 1 | 5.1 | 5.83 |

| P7C1 | P8C1 |
|---|---|
| $\sqrt{3^2+1^2}$ | $\sqrt{4^2+1^2}$ |
| 3.16 | 4.12 |

Q 3>

| | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 |
|---|---|---|---|---|---|---|---|---|
| $x_1$ | 2 | 3 | 4 | 5 | 6 | 8 | 8 | 9 |
| $x_2$ | 2 | 4 | 7 | 3 | 7 | 7 | 1 | 3 |

C1 - (2,2) C2 - (3,5)

Distances:
Iteration 1 -

| P1C1 | P2C1 | P3C1 | P4C1 | P5C1 | P6C1 |
|---|---|---|---|---|---|
| 0 | $\sqrt{1^2+2^2}$ | $\sqrt{2^2+5^2}$ | $\sqrt{3^2+1^2}$ | $\sqrt{4^2+5^2}$ | $\sqrt{5^2+6^2}$ |
| | 2.24 | 5.39 | 3.16 | 6.4 | 7.81 |

| P7C1 | P8C1 |
|---|---|
| $\sqrt{6^2+1^2}$ | $\sqrt{7^2+1^2}$ |
| 6.08 | 7.07 |

| P1C2 | P2C2 | P3C2 | P4C2 | P5C2 | P6C2 |
|---|---|---|---|---|---|
| $\sqrt{1^2+3^2}$ | 1 | $\sqrt{1^2+2^2}$ | $\sqrt{2^2+2^2}$ | $\sqrt{3^2+2^2}$ | $\sqrt{5^2+2^2}$ |
| 3.16 | 1 | 2.24 | 2.83 | 3.61 | 5.39 |

| P7C2 | P8C2 |
|---|---|
| $\sqrt{5^2+4^2}$ | $\sqrt{6^2+2^2}$ |
| 6.4 | 6.32 |

Cluster 1 - {1,7} Cluster 2 - {2,3,4,5,6,8}

C1 - (5, 1.5)
C2 - (5.83, 5.17)

| P1C2 | P2C2 | P3C2 | P4C2 | P5C2 | P6C2 |
|---|---|---|---|---|---|
| $\sqrt{4^2+3.6^2}$ | $\sqrt{3^2+1.6^2}$ | $\sqrt{2^2+1.4^2}$ | $\sqrt{1^2+2.6^2}$ | $\sqrt{1+4^2}$ | $\sqrt{2^2+1.4^2}$ |
| 5.83 | 3.4 | 2.44 | 2.78 | 1.4 | 2.44 |

| P7C2 | P8C2 |
|---|---|
| $\sqrt{2^2+4.6^2}$ | $\sqrt{3^2+2.6^2}$ |
| 5.02 | 3.97 |

Cluster 1 - {1, 2, 4, 7}, Cluster 2 - {3, 5, 6, 8}

C1 - (4.5, 2.5)
C2 - (6.75, 6)

Iteration 4 -

| P1C1 | P2C1 | P3C1 | P4C1 | P5C1 | P6C1 |
|---|---|---|---|---|---|
| $\sqrt{2.5^2+.5^2}$ | $\sqrt{1.5^2+1.5^2}$ | $\sqrt{.5^2+4.5^2}$ | $\sqrt{.5^2+.5^2}$ | $\sqrt{1.5^2+4.5^2}$ | $\sqrt{3.5^2+4.5^2}$ |
| 2.55 | 2.12 | 4.53 | .71 | 4.74 | 5.7 |

| P7C1 | P8C1 |
|---|---|
| $\sqrt{3.5^2+1.5^2}$ | $\sqrt{4.5^2+5^2}$ |
| 3.81 | 4.53 |

| P1C2 | P2C2 | P3C2 | P4C2 | P5C2 | P6C2 |
|---|---|---|---|---|---|
| $\sqrt{4.75^2+4^2}$ | $\sqrt{3.75^2+2^2}$ | $\sqrt{2.75^2+1^2}$ | $\sqrt{1.75^2+3^2}$ | $\sqrt{.75^2+1^2}$ | $\sqrt{1.25^2+1^2}$ |
| 6.21 | 4.25 | 2.93 | 3.47 | 1.25 | 1.6 |

| P7C2 | P8C2 |
|---|---|
| $\sqrt{1.25^2+5^2}$ | $\sqrt{2.25^2+3^2}$ |
| 5.13 | 3.75 |

Cluster 1 - {1, 2, 4, 7} Cluster 2 - {3, 5, 6, 8}
Convergence.

Q4>

| | P1 | P2 | P3 | P4 |
|---|---|---|---|---|
| $x_1$ | 5 | 3 | 1 | 4 |
| $x_2$ | 2 | 3 | 9 | 1 |
| $x_3$ | 7 | 3 | 6 | 2 |
| $x_4$ | 4 | 3 | 1 | 7 |

$C = (2, 7, 4, 5)$

$N = 4$

SUM = $(13, 15, 18, 15)$

SUMSQ = $(51, 95, 98, 75)$

centroid = $(3.25, 3.75, 4.5, 3.75)$

Variance V -

$x_1$ - $51/4 - (13/4)^2 = 12.75 - 10.56 = 2.19$

$x_2$ - $95/4 - (15/4)^2 = 23.75 - 14.06 = 9.69$

$x_3$ - $98/4 - (18/4)^2 = 24.5 - 20.25 = 4.25$

$x_4$ - $75/4 - (15/4)^2 = 18.75 - 14.06 = 4.69$

SD $\sigma = \sqrt{V}$

$x_1$ - 1.48

$x_2$ - 3.11

$x_3$ - 2.06

$x_4$ - 2.17

$$d(x, c) = \sqrt{\sum_{i=1}^{N} \left( \frac{x_i - c_i}{\sigma_i} \right)^2}$$

$$d(x_1, c) = \sqrt{\left(\frac{5-2}{1.48}\right)^2 + \left(\frac{2-7}{3.11}\right)^2 + \left(\frac{7-4}{2.06}\right)^2 + \left(\frac{4-5}{2.17}\right)^2}$$

$$= 4.109 + 2.585 + 2.121 + 0.212$$

$$= \sqrt{9.027} = 3.005$$

$$d(x_2, c) = \sqrt{\left(\frac{3-2}{1.48}\right)^2 + \left(\frac{3-7}{3.11}\right)^2 + \left(\frac{3-4}{2.06}\right)^2 + \left(\frac{3-5}{2.17}\right)^2}$$

$$= \sqrt{0.457 + 1.654 + 0.236 + 0.849}$$

$$= 1.788$$

$$d(x_3, c) = \sqrt{\left(\frac{1-2}{1.48}\right)^2 + \left(\frac{9-7}{3.11}\right)^2 + \left(\frac{6+4}{2.06}\right)^2 + \left(\frac{1-5}{2.17}\right)^2}$$

$$= \sqrt{0.457 + 0.414 + 0.943 + 3.398}$$

$$= 2.283$$

$$d(x_4, c) = \sqrt{\left(\frac{4-2}{1.48}\right)^2 + \left(\frac{1-7}{3.11}\right)^2 + \left(\frac{2-4}{2.06}\right)^2 + \left(\frac{7-5}{2.17}\right)^2}$$

$$= \sqrt{1.826 + 3.722 + 0.943 + 0.849}$$

$$= 2.709$$

$$d(x_i, c) : (3.005, 1.788, 2.283, 2.709)$$

Q5)

BFR

- Assumes clusters are normally distributed in each dimension
- Axes are fixed & ellipses at an angle are not allowed.

CURE

- Assumes a Euclidian distance
- Allows cluster to assume any shape.

Q6.

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | Y |
|---|---|---|---|---|
| a | e | j | u | +1 |
| b | f | K | v | +1 |
| b | f | j | v | −1 |
| a | f | K | u | −1 |
| a | f | j | u | +1 |
| a | f | K | u | −1 |
| b | e | K | v | +1 |
| a | f | j | v | +1 |
| a | f | K | v | −1 |
| b | e | K | v | +1 |

$I_E$ Entropy : $\frac{-5}{10}\log\frac{5}{10} - \frac{5}{10}\log\frac{5}{10} = 1$

$X_1$ : #a = 6 #b = 4

a: #+1 = 2 #−1 = 4

b: #+1 = 3 #−1 = 1

$I_a$ : $\frac{-2}{6}\log\frac{2}{6} - \frac{4}{6}\log\frac{4}{6}$ $I_b$ : $\frac{-3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4}$

$= 0.92$ $= 0.81$

$IG(X_1) = I_E - \frac{6}{10}I_a - \frac{4}{10}I_b$

$= 1 - 0.876 = 0.124$

$X_2$: $\#e = 3$ $\#f = 7$

e: $\#+1 = 3$ $\#-1 = 0$ $I_e = 0$

f: $\#+1 = 2$ $\#-1 = 5$ $I_f = -\frac{2}{7}\log\frac{2}{7} - \frac{5}{7}\log\frac{5}{7}$

$= 0.87$

$IG(X_2) = IE - \frac{3}{10}I_e - \frac{7}{10}I_f$

$= 1 - 0.609 = 0.391$

$X_3$: $\#j = 4$ $\#K = 6$

j: $\#+1 = 2$ $\#-1 = 2$ $I_j = 1$

K: $\#+1 = 3$ $\#-1 = 3$ $I_K = 1$

$IG(X_3) = 1 - \frac{4}{10}I_j - \frac{6}{10}I_K = 0$

$X_4$: $\#u = 4$ $\#v = 6$

u: $\#+1 = 2$ $\#-1 = 2$ $I_u = 1$

v: $\#+1 = 3$ $\#-1 = 3$ $I_v = 1$

$IG(X_4) = 1 - \frac{4}{10}I_u - \frac{6}{10}I_v = 0$

$\therefore X_2$ is 1st splitting attribute

$IE = -\frac{5}{10}\log\frac{5}{10} - \frac{5}{10}\log\frac{5}{10} = 1$

$X_1$ : $\#a = 4$ $\#b = 2$

a: $\#+1 = 1$, $\#-1 = 3$

$I_a = \frac{1}{4}\log\frac{1}{4} - \frac{3}{4}\log\frac{3}{4}$

$\approx 0.81$

b: $\#+1 = 1$, $\#-1 = 1$

$I_b = 1$

$IG(X_1) = IE - \frac{4}{6}I_a - \frac{2}{6}I_b$

$= 1 - 0.873 = 0.13$

$X_3$ : $\#j = 3$ $\#k = 3$

j: $\#+1 = 1$, $\#-1 = 2$

$I_j = \frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3}$

$= 0.92$

K: $\#+1 = 1$, $\#-1 = 2$

$I_K = \frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3}$

$= 0.92$

$IG(X_3) = IE - \frac{3}{6}I_j - \frac{3}{6}I_k$

$= 1 - 0.92 = 0.08$

$X_4$ : $\#u = 2$ $\#v = 4$

u: $\#+1 = 1$, $\#-1 = 1$

$I_u = 1$

v: $\#+1 = 1$, $\#-1 = 3$

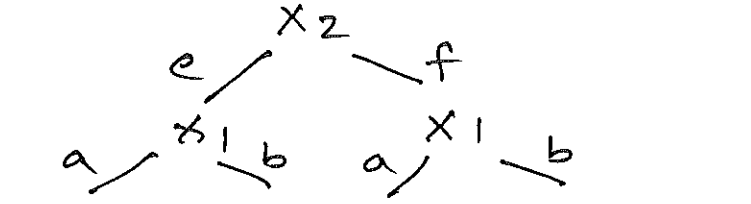
$I_v = \frac{1}{4}\log\frac{1}{4} - \frac{3}{4}\log\frac{3}{4}$

$= 0.81$

$IG(X_4) = IE - \frac{2}{6}I_u + \frac{4}{6}I_v = 1 - 0.873 = 0.13$

$X_1$ or $X_4$ can be splitting attribute at level 2 as they have same IG.

Decision Tree:




Majority class: (+1) (+1) (-1) (+1)

Classification accuracy: 8/10 = 0.8 or 80%