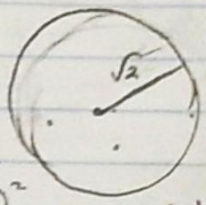
CS6220 - Maha Ashraf AlRhairy

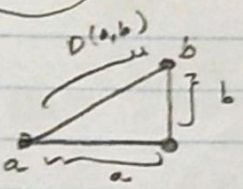
Data Mining HW4 :-

① $\varepsilon = \sqrt{2}$ minpts = 3

using euclidean distance

$$D(a,b) = \sqrt{(a_x - b_x)^2 + (a_y - b_y)^2}$$



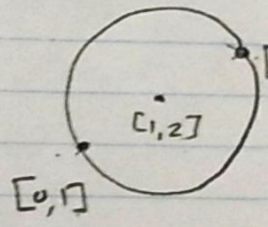


*) choose random point and find all points $\sqrt{2}$ away from it

$q = [1,2]$

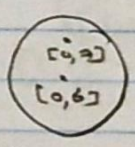
$D(q, [0,1]) = \sqrt{1^2 + (1)^2} = \sqrt{2}$

$D(q, [2,3]) = \sqrt{1^2 + 1^2} = \sqrt{2}$



one cluster

$q = [0,6]$ $D(q, [0,7]) = \sqrt{1}$



number of points within ε of $[0,6]$ is 1

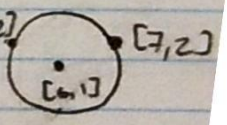
not satisfy minPts

so $[0,6]$ $[0,7]$ are noise

$q = [6,1]$ $D(q, [7,2]) = \sqrt{1^2 + 1^2} = \sqrt{2}$

$D(q, [5,2]) = \sqrt{1+1} = \sqrt{2}$.

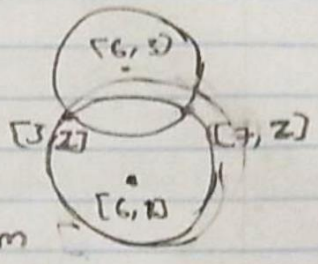
× $D(q, [6,3]) = \sqrt{0+2^2} = 2$.



q = [6, 3]

$D(q, [5,2]) = \sqrt{1^2+1^2} = \sqrt{2}$

$D(q, [7,2]) = \sqrt{1^2+1^2} = \sqrt{2}$.



[6, 3] density reachable from

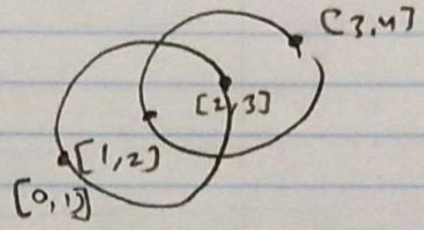
[6, 1] and linked by [5, 2] and [7, 2]

---

q = [2, 3]

[1, 2] is ε neighborhood

$D(q, [3,4]) = \sqrt{1^2+1^2} = \sqrt{2}$



$\begin{bmatrix} \\ \\ \\ \end{bmatrix}_{N \times d}$ $\begin{bmatrix} \\ \\ \\ \\ \end{bmatrix}_{N \times d}$

(a) mean m_1 : $[6.\bar{6}, \ 5.\bar{3}]$

mean m_2 : $[6 \quad 4.2]$

(b) mean m : $[6.25 \quad 4.\bar{6}25]$

(c) $S_1 = \sum_{x \in C_1} (x - m_1)(x - m_1)^T = \begin{bmatrix} 4.\bar{6} & 0.\bar{3} \\ 0.\bar{3} & 8.\bar{6} \end{bmatrix}$

$$(X - m_1) = \begin{bmatrix} -1.\bar{6} & 0.\bar{6} \\ 1.\bar{3} & 1.\bar{6} \\ 0.\bar{3} & -2.\bar{3} \end{bmatrix}$$

$$(X - m_1)^T = \begin{bmatrix} -1.\bar{6} & 1.\bar{3} & 0.\bar{3} \\ 0.\bar{6} & 1.\bar{6} & -2.\bar{3} \end{bmatrix}$$

$$(X - m_1)^T (X - m_1) = \begin{bmatrix} 4.\bar{6} & 0.\bar{3} \\ 0.\bar{3} & 8.\bar{6} \end{bmatrix}_{2 \times 2}$$

$$S_2 = \sum_{X \in C_2} (X - m_2)(X - m_2)^T$$

$$(X - m_2)^T = \begin{bmatrix} 0 & 0.8 \\ -2 & 0.8 \\ 3 & -2.2 \\ -3 & 0.8 \\ 2 & -0.2 \end{bmatrix}_{5 \times 2}$$

$$(X - m_2)^T = \begin{bmatrix} 0 & -2 & 3 & -3 & 2 \\ 0.8 & 0.8 & -2.2 & 0.8 & -0.2 \end{bmatrix}_{2 \times 5}$$

$$(X - m_2)^T (X - m_2) = \begin{bmatrix} 26 & -11 \\ -11 & 6.8 \end{bmatrix}_{2 \times 2}$$

(d) $$S_W = \sum_{i=1}^{2} S_i = \begin{bmatrix} 26 + 4.\bar{6} & -11 + 0.3 \\ -11 + 0.3 & 8.\bar{6} + 6.8 \end{bmatrix} = \begin{bmatrix} 30.\bar{6} & -10.\bar{6} \\ -10.\bar{6} & 15.\overline{46} \end{bmatrix}$$

$m_1 = [6.5 \quad 5.\bar{3}]_{1\times 2}$

$m_2 = [6 \quad 4.2]_{1\times 2}$

$m = [6.25 \quad 4.625]_{1\times 2}$

(e)

$$S_B = \sum_{i=1}^{2} N_i (m_i - m)(m_i - m)^T$$

$$= 3(m_1 - m)^T (m_1 - m) + 5(m_2 - m)^T (m_2 - m)$$

...

$$= \begin{bmatrix} 0.8\bar{3} & 1.41\bar{6} \\ 1.41\bar{6} & 2.408\bar{3} \end{bmatrix}$$

(f) $tr(S_W) = 30.\bar{6} + 15.4\bar{6} = 46.1\bar{3}$ (want low)

$tr(S_B) = 0.8\bar{3} + 2.408\bar{3} = 3.241\bar{6}$ (want high)

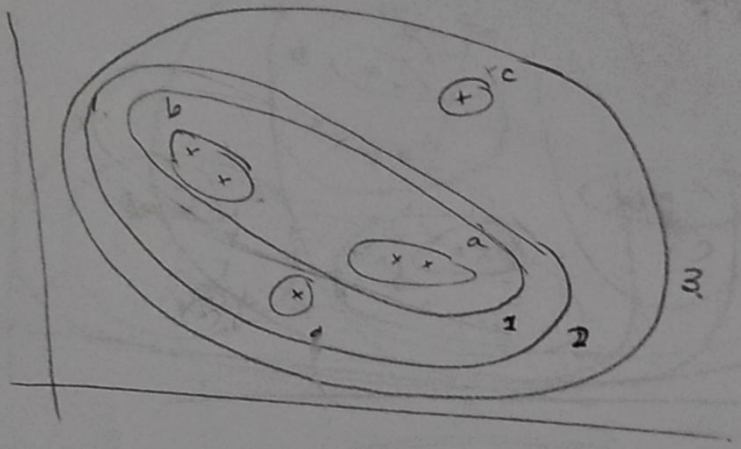
$$\frac{tr(S_B)}{tr(S_W)} = 0.070267 \quad \text{(want high)}$$

based on the scattering criteria, this is <u>not</u> a good clustering.

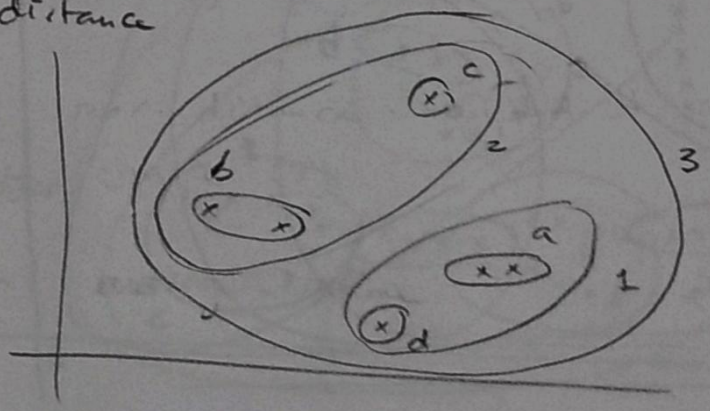
Data Mining - HW4 - Question 3

Graph 1: - * numbers indicate the merging order of clusters.

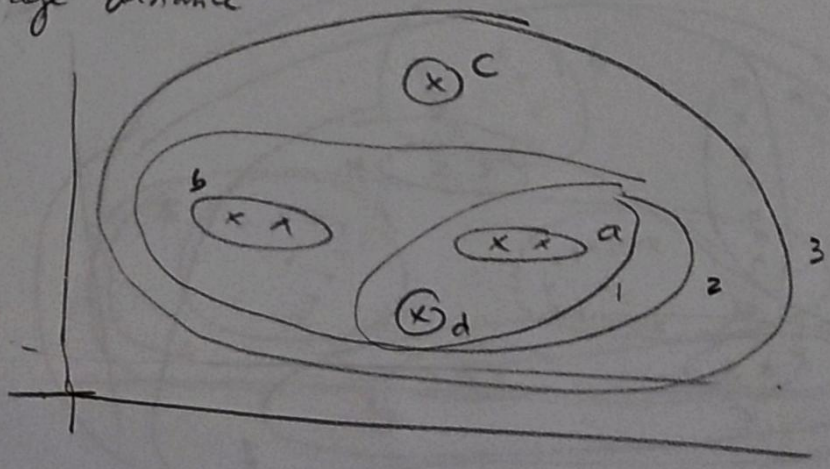
min distance



max distance

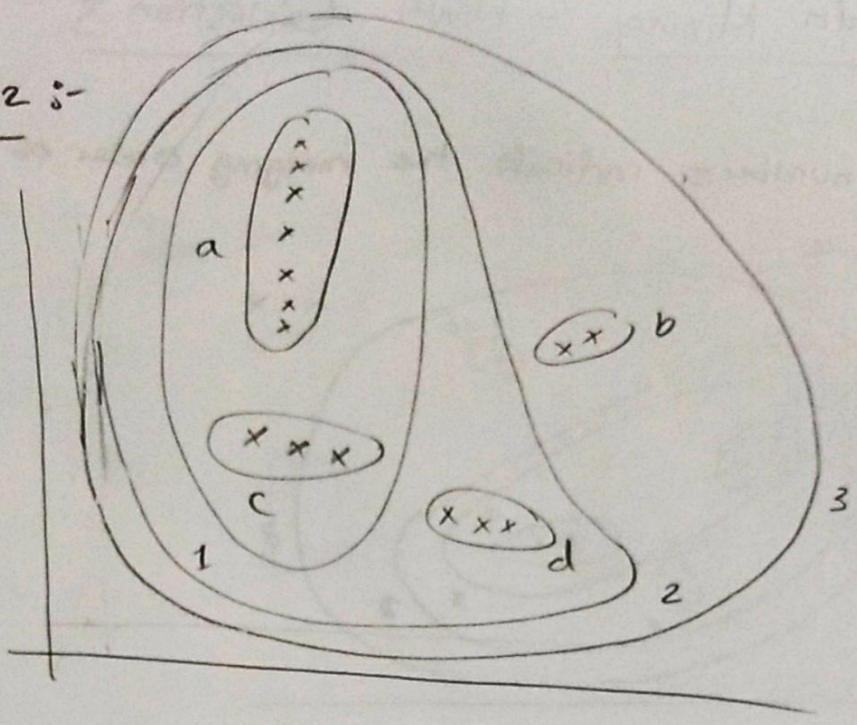


average distance

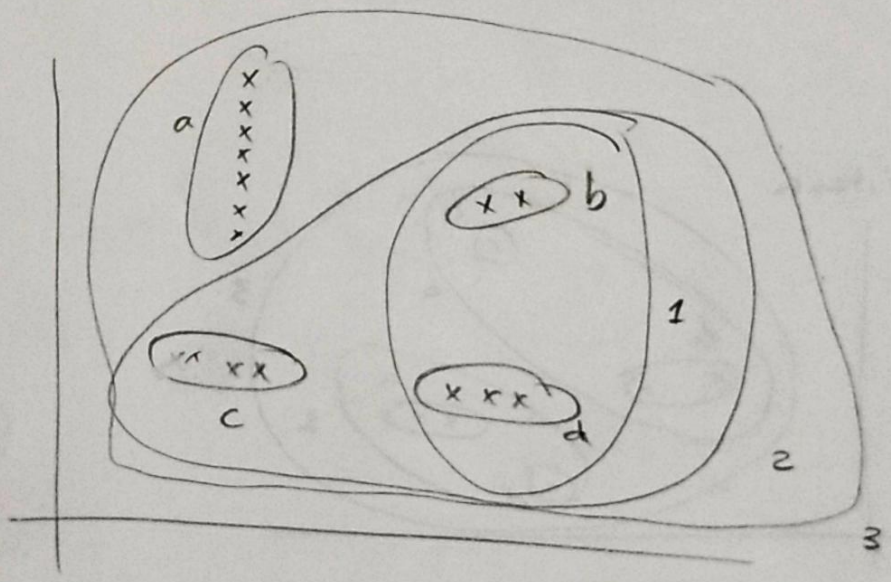


Graph2 :-

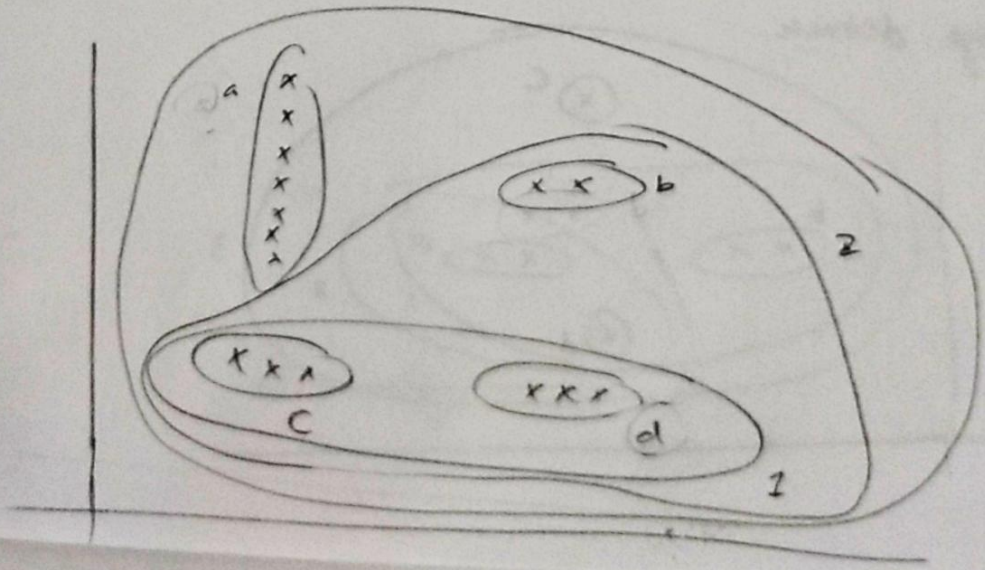
min



max

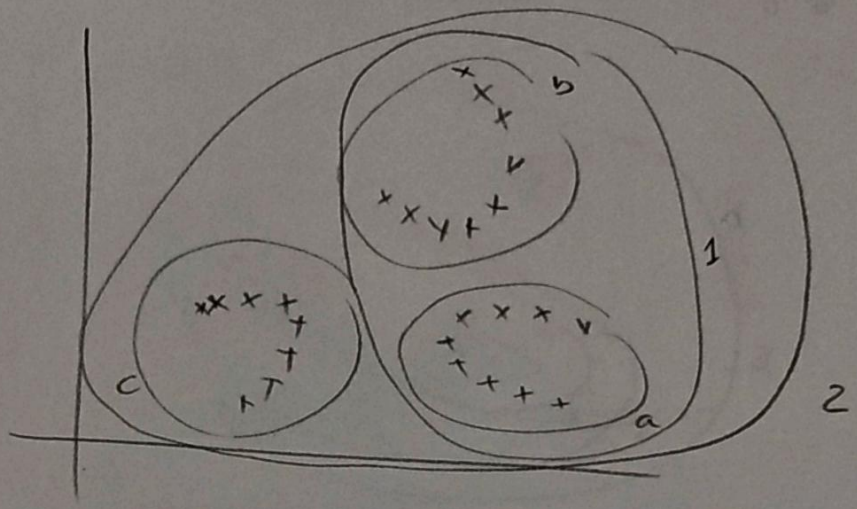


average

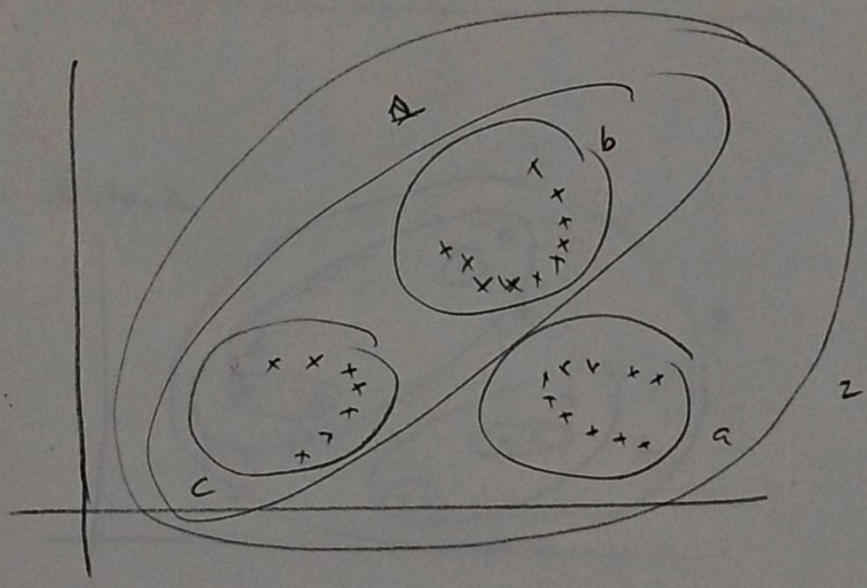


Graphs :-

min



max



averag