Assignment - 3

Q1

It is not a valid distance as it does not satisfy triangle inequality.

$d(x,y) = |x-y|^2$

$|x-z| \leq |x-y| + |y-z|$

Squaring $|x-z|^2 \leq |x-y|^2 + |y-z|^2 + 2|x-y||y-z|$

↳ extra term

This is not guaranteed.

Q2

$d_{max}^{PS} = \max_{y \in C} (x,y) = d(x,x_1)$

$(6,4) \& (1.5,1.5)$, $d_{max} = \sqrt{(6-1.5)^2 + (4-1.5)^2}$

$= 5.15$

$d_{min}^{PS}(x,C) = \min_{y \in C} (x,y) = d(x,x_8)$

$(6,4) \& (3.5,3)$, $d_{min} = \sqrt{(6-3.5)^2 + (4-3)^2}$

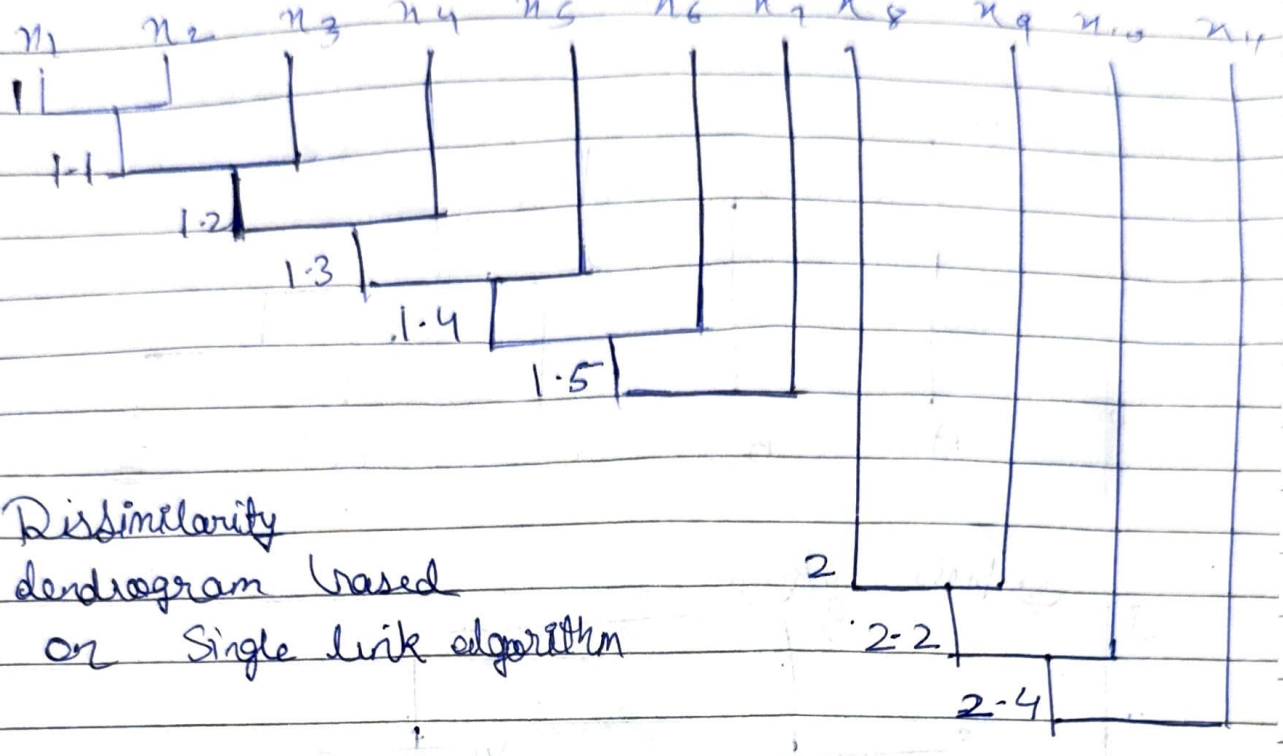
$= 2.69$

$d_{avg}^{PS}(x,C) = \frac{1}{nc} \sum_{y \in C} d(x,y)$

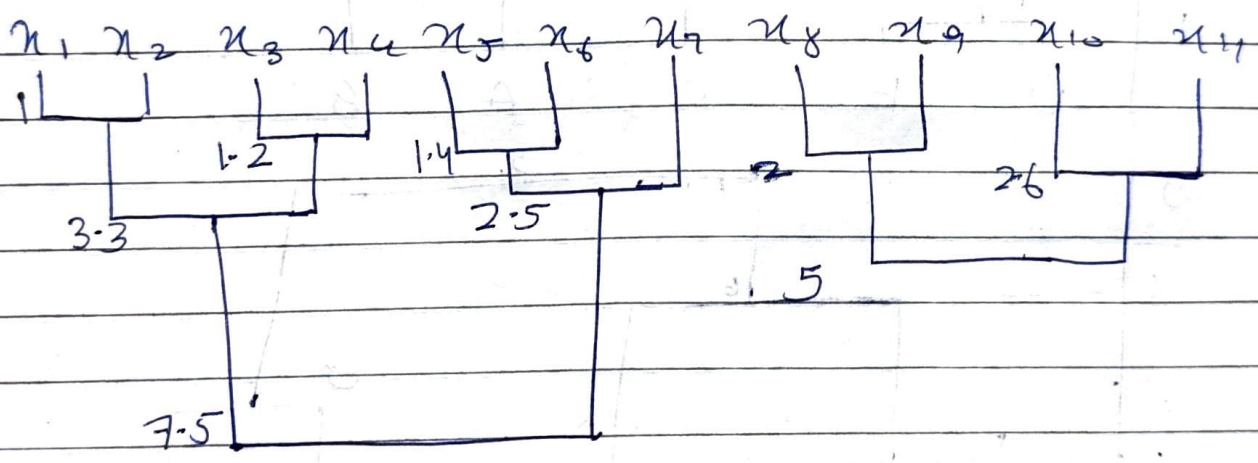
$= \frac{1}{8} \sum_{i=1}^{8} d(x,x_i) = 4.33$

Date :
KALAM Page :

Q3



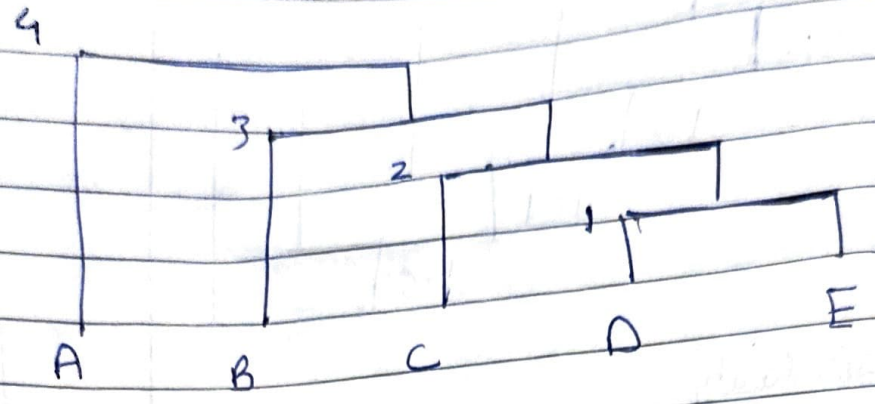
Dissimilarity dendrogram based on Single link algorithm



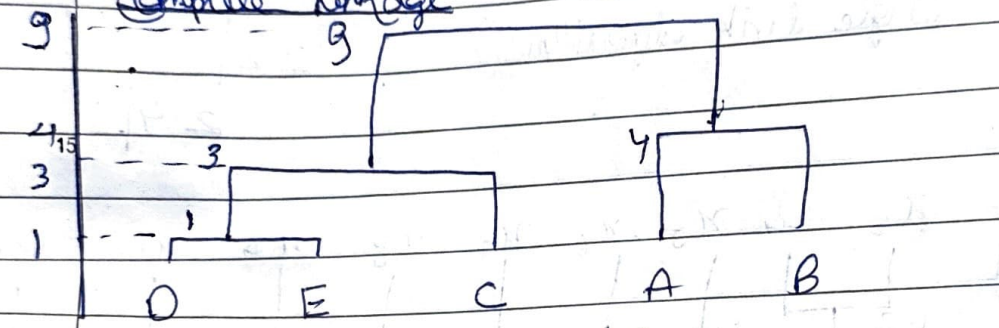
Complete link algorithm

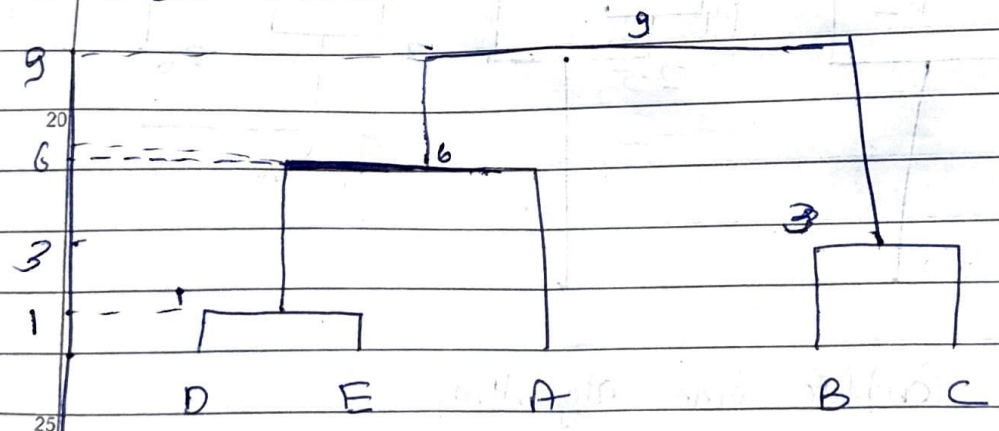
Q-4

Single Linkage Dendrogram



Complete linkage -





Date :
Page :

Q5 $C_1 - (2,3)$ C_2 $(5,8)$ $C_3 (9,4)$

(a) $(1,2) \rightarrow C_1$
$(3,4) \Rightarrow C_1$
$(6,7) \rightarrow C_2$
$(8,3) \rightarrow C_3$
$(5,5) \rightarrow C_2$

C_1 - Mean, $C_1' = \left(\frac{1+3}{2}, \frac{2+4}{2}\right) = (2,3)$

C_2 - Mean, $C_2' = \left(\frac{6+5}{2}, \frac{7+5}{2}\right) = (5.5, 6)$

C_3 - Mean $C_3' = (8,3)$

(b) Distortion -

Initial -

Op 6.0	2 from $(1,2)$ + 2 from $(3,4)$ + 2 from $(6,7)$ + 2 from $(8,3)$ + 9 from $(5,5) = 17$

Final -

2 from $(1,2)$ + 2 from $(3,4)$ + 1.25 from $(6,7)$ + 0 from $(8,3)$ + 1.25 from $(5,5) = 6.5$

distortion decreases.

Date :
KALAM Page :

8-6

Gaussian density -

$$\ln N(x_n | \mu_k, \Sigma_k) = -\frac{D}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}(x_n-\mu_k)^T\Sigma^{-1}(x_n-\mu_k)$$

Maximizing with r.t.d - Σ_k

$$0 = \frac{\ln(N)}{d\Sigma_k} = -\frac{N_k}{2}\Sigma_k^{-1} - \frac{1}{2}\sum_{n=1}^{N}\gamma(z_{nk})\Sigma_k^{-1}(x_n-\mu_k)(x_n-\mu_k)^T\Sigma_k^{-1}$$

$$\Rightarrow \Sigma_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(x_n-\mu_k)(x_n-\mu_k)^T$$

Maximizing wrt π_k

we use - $\sum_{n=1}^{N}\sum_{k=1}^{K}\ln(\pi_k)\gamma(z_{nk})$

Under constraint of $\sum_{k=1}^{K}\pi_k = 1$

So, $\sum_{n=1}^{N}\gamma(z_{nk})\ln(\pi_k) + \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right)$

diff - , $= \frac{N_k}{\pi_k} + \lambda = 0$

$$\Rightarrow \pi_k = -\frac{N_k}{\lambda}$$

$$\sum_{k=1}^{K}\pi_k = \sum_{k=1}^{K}-\frac{N_k}{\lambda} \Rightarrow \pi_k = \frac{N_k}{N}$$

Date :
Page :

6-7 Conditional density

$$p(x_b|x_a) = \frac{p(x_a, x_b)}{p(x_a)}$$

$$= \sum_{k=1}^{K} \left(\frac{\pi_k p(x_a|k)}{\sum_{j=1}^{K} \pi_j p(x_a|j)} \right) p(x_b|x_a, k)$$

Mixing coefficient - $\frac{\pi_k p(x_a|k)}{\sum_{j=1}^{K} \pi_j p(x_a|j)}$

Component densities - $p(x_b|x_a, k)$

6.8 (a) Complete log-likelihood

$$\log p(X, Z|\theta) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \left[\log \pi_k - \frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x_n - \mu_k)^T \Sigma_k^{-1} (x_n - \mu_k) \right]$$

(b) MLE -

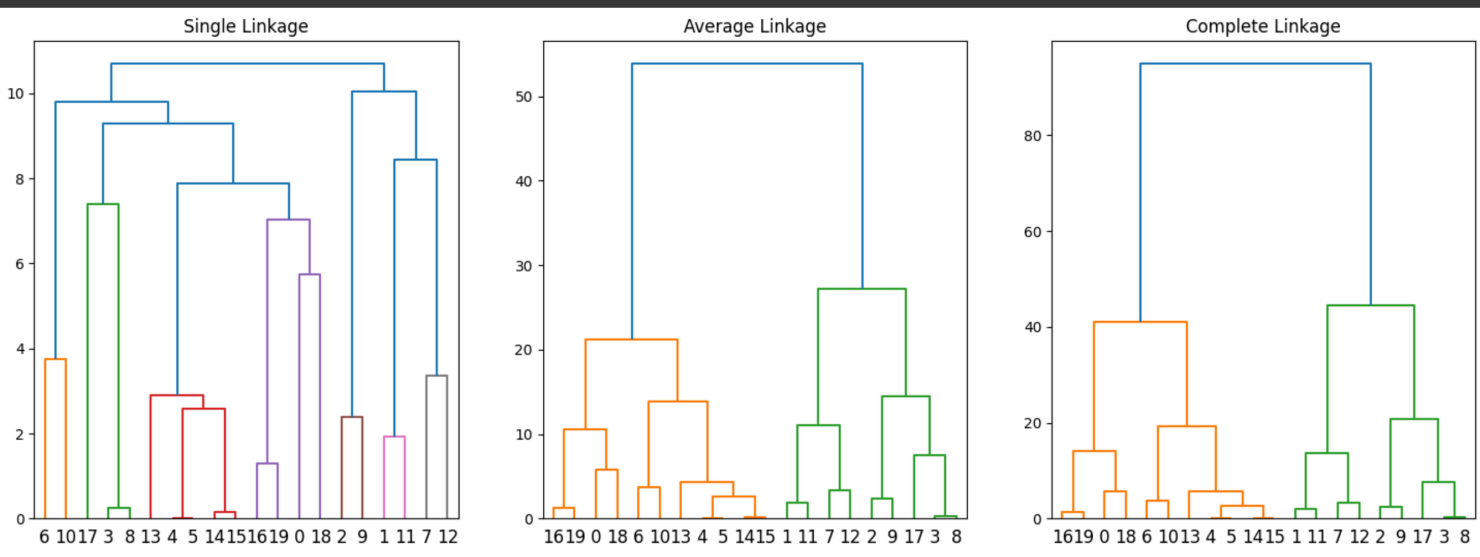
Maximise $\sum_k \pi_k \log \pi_k$ under $\sum_k \pi_k = 1$

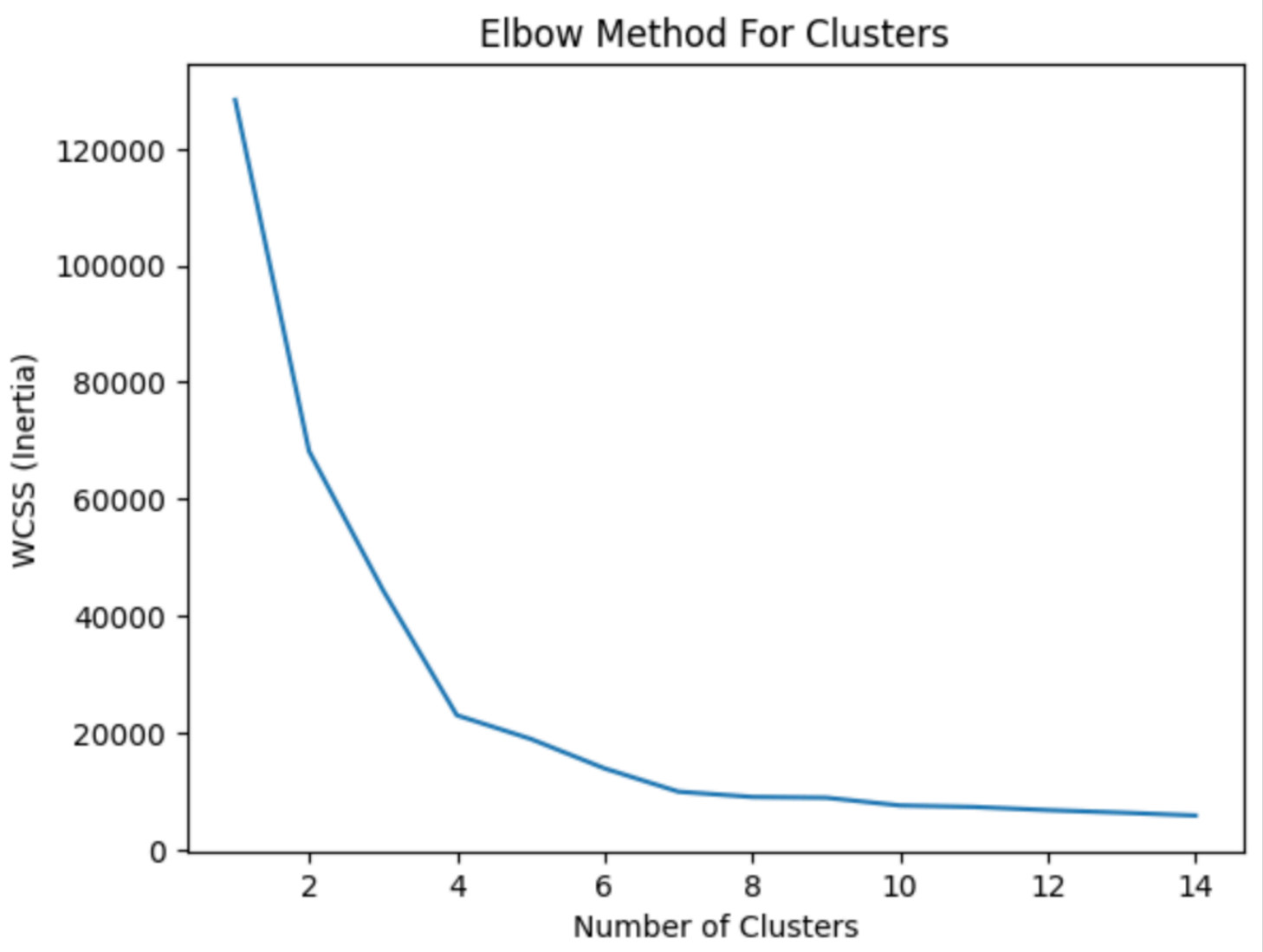
Using Lagrange multiplier -

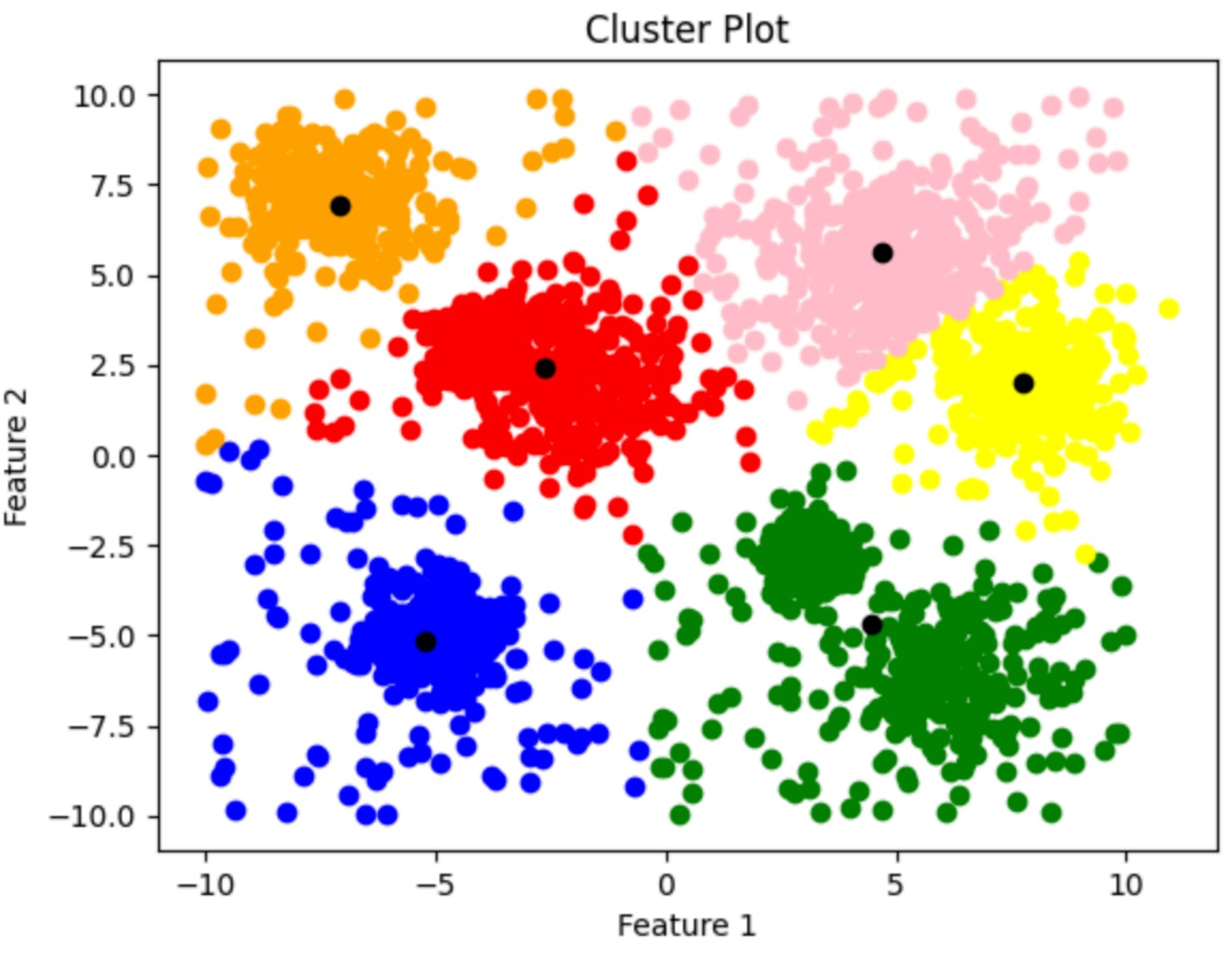
$$\pi_k = \frac{N_k}{N}, \quad N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$

Covariance Σ_k

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} z_{nk} (x_n - \mu_k)(x_n - \mu_k)^T$$


Single Linkage
Average Linkage
Complete Linkage
0
2
4
6
8
10
20
30
40
50
60
80
6 10 17 3 8 13 4 5 14 15 16 19 0 18 2 9 1 11 7 12
16 19 0 18 6 10 13 4 5 14 15 1 11 7 12 2 9 17 3 8
16 19 0 18 6 10 13 4 5 14 15 1 11 7 12 2 9 17 3 8


Elbow Method For Clusters
WCSS (Inertia)
120000
100000
80000
60000
40000
20000
0
2
4
6
8
10
12
14
Number of Clusters


Cluster Plot
Feature 1
Feature 2
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−10
−5
0
5
10