Chapter 12 Exercises

I Recitation Exercises

(1)(a) We know that:

$$\frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 = 2 \sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2$$

Let's decompose the left part

$$\sum_{i,i' \in C_k} \sum_{j=1}^{p} (x^2_{ij} - 2x_{ij} + x^2_{i'j}) = |C_k| \sum_{i \in C_k} \sum_{j=1}^{p} x^2_{ij} - 2 \sum_{i \in C_k} \sum_{i' \in C_k} \sum_{j=1}^{p} x_{ij} x_{i'j}$$
$$+ |C_k| \sum_{i \in C_k} \sum_{j=1}^{p} x^2_{ij}$$

Let us focus on middle term

$$-2 \sum_{i \in C_k} \sum_{i' \in C_k} \sum_{j=1}^{p} x_{ij} x_{i'j} = -2 \sum_{j=1}^{p} \sum_{i \in C_k} x_{ij} \left(\sum_{i' \in C_k} x_{i'j}\right)$$

But we also know that

$$C_k : \bar{x}_{kj} = \frac{1}{|C_k|} \sum_{i \in C_k} x_{ij}$$

So, finally we the part as:

$$2|C_k| \sum_{i \in C_k} \sum_{j=1}^{p} x^2_{ij} - 2|C_k|^2 \sum_{j=1}^{p} \bar{x}^2_{kj}$$

Now divide $|C_k|$ and

$$2\sum_{i\in C_k}\sum_{j=1}^{p} x_{ij}^2 - 2|C_k|\sum_{j=1}^{p}\bar{x}_{kj}^2$$

Rewrite the term as

$$-2|C_k|\sum_{j=1}^{p}\bar{x}_{kj}^2 = -4|C_k|\sum_{j=1}^{p}\bar{x}_{kj}^2 + 2|C_k|\sum_{j=1}^{p}\bar{x}_{kj}^2 = -4\sum_{i\in C_k}\sum_{j=1}^{p} x_{ij}\bar{x}_{kj}$$

$$+2\sum_{i\in C_k}\sum_{j=1}^{p}\bar{x}_{kj}^2$$

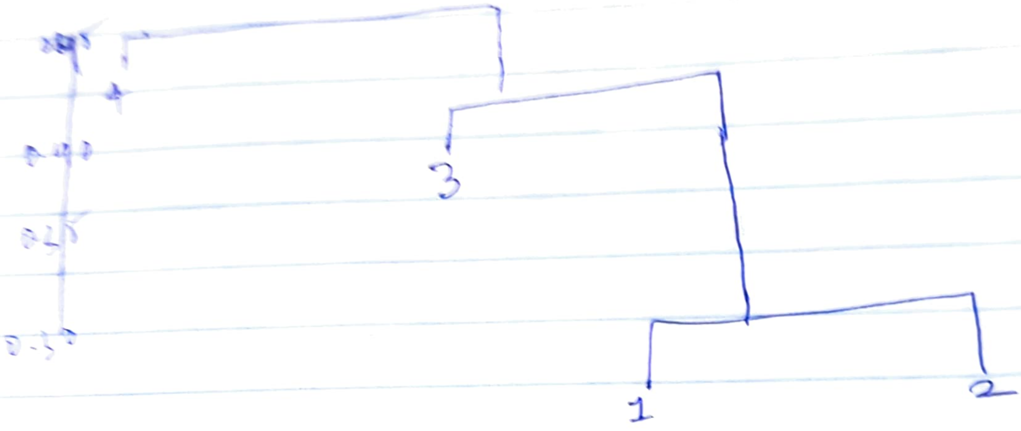
Replace with the rewritten left term and simplify

$$2\sum_{i\in C_k}\sum_{j=1}^{p} x_{ij}^2 - 4\sum_{i\in C_k}\sum_{j=1}^{p} x_{ij}\bar{x}_{kj} + 2\sum_{i\in C_k}\sum_{j=1}^{p}\bar{x}_{kj}$$

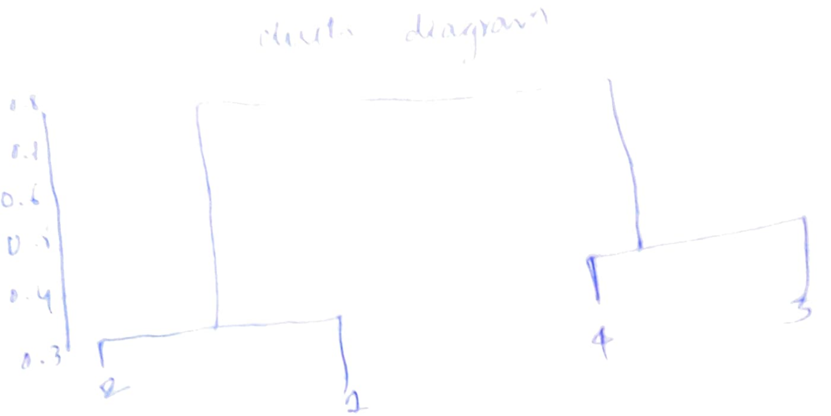
$$= 2\sum_{i\in C_k}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})^2$$

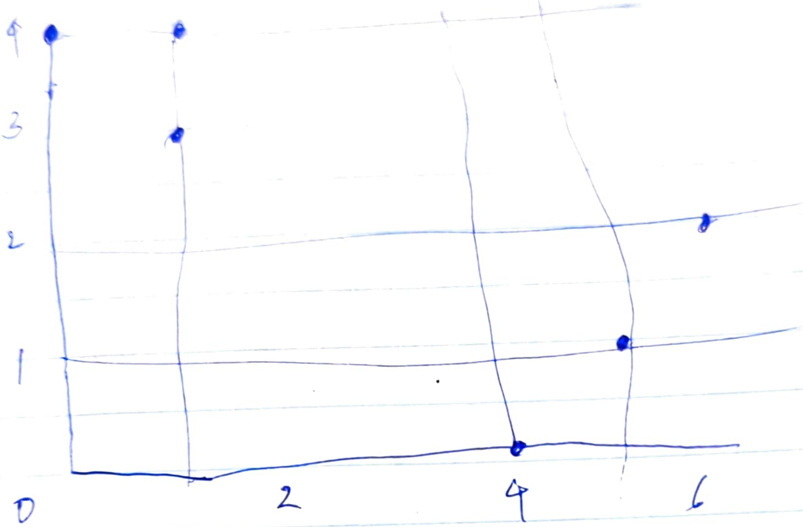
(b) In k means clustering minimization of sum of squared euclidean distance for each cluster and our objective is the minimize the distance with cluster variance for each cluster and both are same.

(b)



(c) Observations 1 and 2 will be in same cluster and observatio





ⓑ The random clusters are

[1 1 1 1 2 2]

ⓒ Centroids of each cluster

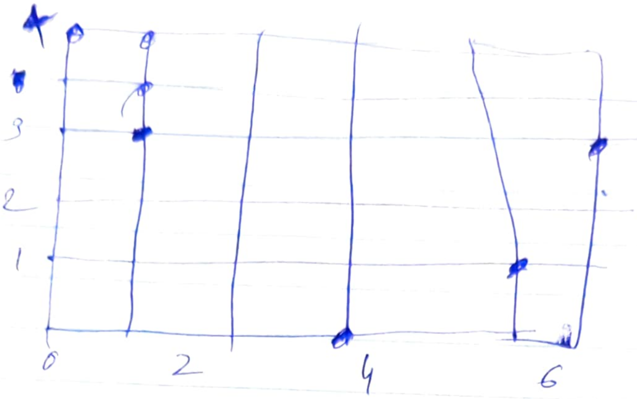
Cluster	X1	X2
1	1.75	3
2	5.00	1

ⓓ After the iterations the 3 points near the left corner are assigned to same, on right corner are another clust

[1, 1, 1, 2, 2, 2]

(e) At iteration 2, it is the same cluster

(f)



Exercise 4

(a) In single linkage clustering, the distance between two clusters is defined as the minimum distance between any two observations belonging to different cluster, in ~~dit~~ complete linkage clustering is defined as maximum & minimum values are equal, will fuse at same height, else the single linkage will fuse at a lower height. But without knowing this information, we cant conclude so we can only say we need more information.

(b) Since we are fusing singleton clusters in both single linkage and complete linkage, the minimum distance between observation is equal to maximum distance

- Same height in both dendograms.