Tut 9: k-Means Clustering

Jan 2022

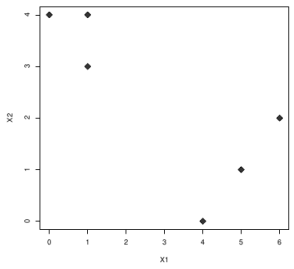
1. The first step of k-means clustering is to decide the number of clusters, k. After a series of iterations, can k-means ever give results which contain

   (a) More than k clusters?

   (b) Less than k clusters?

2. You are given a small example with $n = 6$ observations and $p = 2$ variables. The observations are as follows:

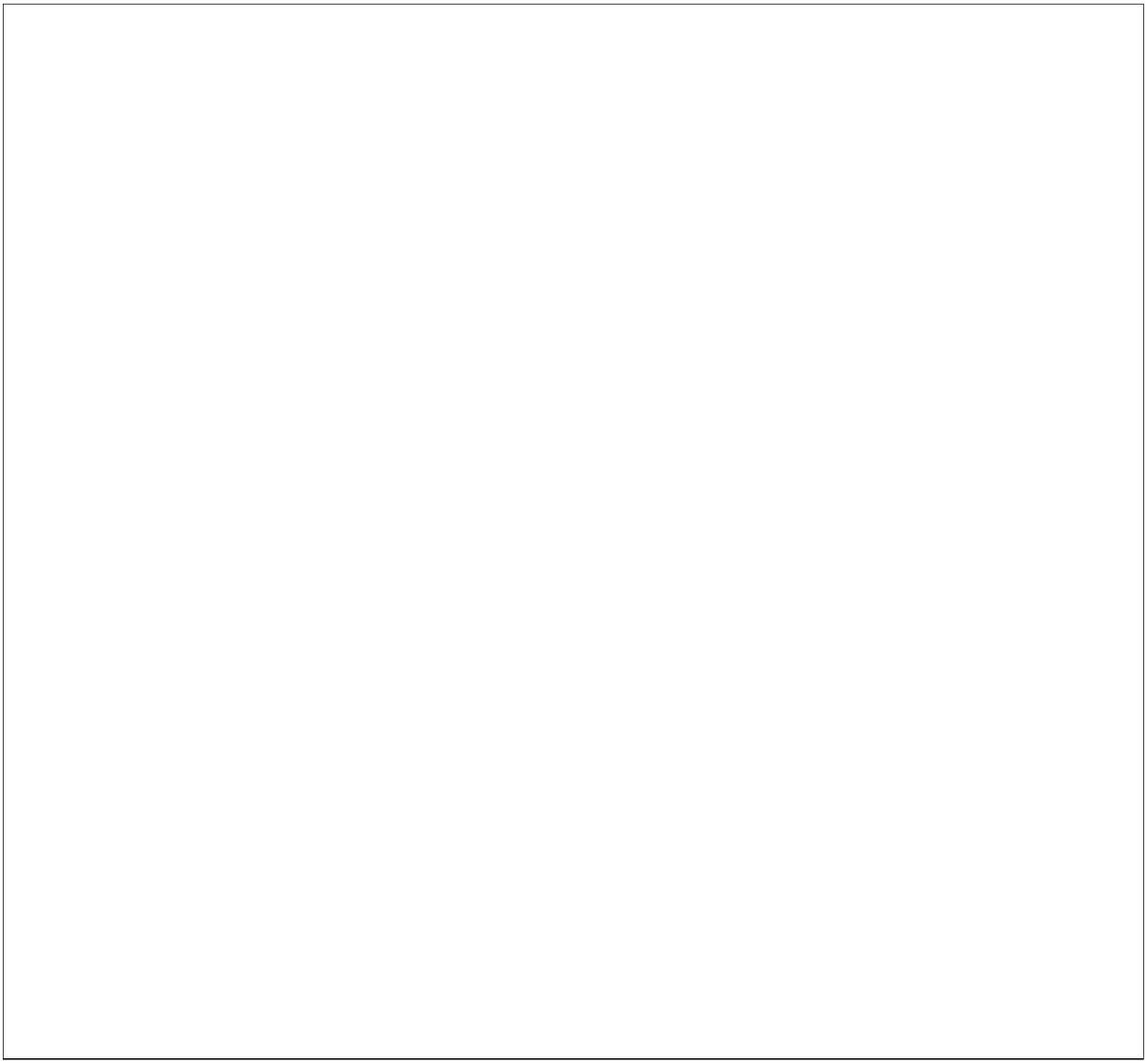
Obs	X_1	X_2
1	1	4
2	1	3
3	0	4
4	5	1
5	6	2
6	4	0



   (a) Plot the observations.

   (b) Rescale the observations to [0,1].

(c) Perform k-means clustering to the observations with $k = 2$. The initial centroids are 2, 5.



(d) In the plot from (a), colour the observations according to the cluster labels obtained.

3. (Jan 2021 Final Q3(b). Need to use Excel/R to perform calculations) Given the unlabelled data in Table 3.2.

Table 3.2: Unlabelled data.

	V1	V2	V3	V4
1	-0.3323	0.7264	2.4691	1.8429
2	5.5783	5.7211	-3.3731	3.9209
3	-1.5492	1.4777	5.1921	0.9621
4	8.0669	-1.1127	1.2409	-0.1392
5	-0.294	-0.5842	0.7708	1.6414
6	5.5741	3.4215	0.9827	3.8443
7	-1.838	0.5629	-3.898	4.483
8	2.6957	-0.2016	0.6947	0.6821
9	10.7553	0.1658	-0.8895	3.0359
10	6.0329	2.3343	0.8758	2.8348

Use the k-means algorithm with $k = 2$ (unsupervised learning) to estimate the final cluster centres in **three steps** if the **first row** and **third row** are chosen as the **initial cluster centres**. Does the algorithm **converges** in three steps? (5 marks)