Tut 9: k-Means Clustering

Feb 2025

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?

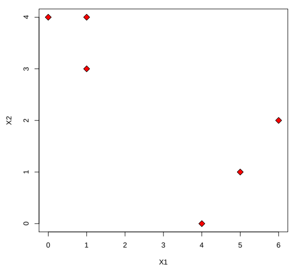
   Solution. No. It can never give more than k clusters, since at every stage every point is assigned to one of k clusters. □

   (b) Less than k clusters?

   Solution. To give fewer than k clusters, we would need there to be a cluster which contain no points at one of the re-assignment stages. This means that its centre would be farther from every point than one of the other cluster centres and results in an empty 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.

   Solution. In Python:

```
import matplotlib.pylab as plt
plt.plot([1,1,0,5,6,4],[4,3,4,1,2,0],'o')
plt.xlabel('$X_1$'); plt.ylabel('$X_2$')
```

   In R:

```
plot(c(1,1,0,5,6,4),c(4,3,4,1,2,0),type='p',xlab="X1",ylab="X2",
  pch=23,bg="red",cex=1.5)
```

   □

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

   Solution. Scale with min-max normalisation in R using

```
d.f = data.frame(x1=c(1,1,0,5,6,4),x2=c(4,3,4,1,2,0))
normdf = scale(df,center=c(0,0),scale=sapply(df,function(x){max(x)-min(x)}))
```

   which gives

Obs	X_1	X_2	Clust_Initial	Norm_X1	Norm_X2
1	1	4	A	0.1667	1.0000
2	1	3	A	0.1667	0.7500
3	0	4	B	0.0000	1.0000
4	5	1	B	0.8333	0.2500
5	6	2	A	1.0000	0.5000
6	4	0	B	0.6667	0.0000

□

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

Solution. $t = 0$:

$$C_1^{(0)} = (0.1667,\ 0.7500); \quad C_2^{(0)} = (1.0000,\ 0.5000)$$

and then find the Euclidean distance for all points to the cluster centres $C_A^{(2)}$ and $C_B^{(2)}$:

Obs	Dist_A	Dist_B	Cluster*
1	0.2500000	0.9718253	1
2	0.0000000	0.8700255	1
3	0.3004626	1.1180340	1
4	0.8333333	0.3004626	2
5	0.8700255	0.0000000	2
6	0.9013878	0.6009252	2

$t = 1$: Compute the cluster centres from the previous table:

$$C_A^{(3)} = (0.1111, 0.9167); \quad C_B^{(3)} = (0.8333, 0.2500)$$

and then find the Euclidean distance for all points to the cluster centres $C_1^{(1)}$ and $C_2^{(1)}$:

Obs	Dist_A	Dist_B	Cluster*
1	0.1002	1.0035	1
2	0.1757	0.8333	1
3	0.1389	1.1211	1
4	0.9829	0.0000	2
5	0.9817	0.3005	2
6	1.0719	0.3005	2

We can see that the clusters do not change, so we have the final cluster centres $C_1^{(1)}$, $C_2^{(1)}$ and stop. □

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

Solution. A "command" for plotting "kmeans" can be found in `practical2.R`.

```
plot(normdf,col=km$cluster+1,pch=20,cex=4)
```

□

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)

Solution. Given the initial centres:

V1	V2	V3	V4
-0.3323	0.7264	2.4691	1.8429
-1.5492	1.4777	5.1921	0.9621

Step 1 : Update table based on distance to cluster centres

V1	V2	V3	V4	dist.1	dist.2	clust.centre
-0.3323	0.7264	2.4691	1.8429	0	3.1993	1
5.5783	5.7211	-3.3731	3.9209	9.9162	12.2851	1
-1.5492	1.4777	5.1921	0.9621	3.1993	0	2
8.0669	-1.1127	1.2409	-0.1392	8.9088	10.7705	1
-0.294	-0.5842	0.7708	1.6414	2.155	5.0829	1
5.5741	3.4215	0.9827	3.8443	6.9544	8.9747	1
-1.838	0.5629	-3.898	4.483	7.0572	9.7952	1
2.6957	-0.2016	0.6947	0.6821	3.8113	6.4144	1
10.7553	0.1658	-0.8895	3.0359	11.6599	13.943	1
6.0329	2.3343	0.8758	2.8348	6.8281	8.9643	1

. . . . [1.5 marks]

The new cluster centres are

$$C_1^{(1)} = (4.0265, 1.2259, -0.1252, 2.4607), \quad C_2^{(1)} = (-1.5492, 1.4777, 5.1921, 0.9621).$$

[0.5 mark]

Step 2 : Update table based on distance to cluster centres

V1	V2	V3	V4	dist.1	dist.2	clust.centre
-0.3323	0.7264	2.4691	1.8429	5.1343	3.1993	2
5.5783	5.7211	-3.3731	3.9209	5.941	12.2851	1
-1.5492	1.4777	5.1921	0.9621	7.8531	0	2
8.0669	-1.1127	1.2409	-0.1392	5.5154	10.7705	1
-0.294	-0.5842	0.7708	1.6414	4.8392	5.0829	1
5.5741	3.4215	0.9827	3.8443	3.2183	8.9747	1
-1.838	0.5629	-3.898	4.483	7.2908	9.7952	1
2.6957	-0.2016	0.6947	0.6821	2.7649	6.4144	1
10.7553	0.1658	-0.8895	3.0359	6.8786	13.943	1
6.0329	2.3343	0.8758	2.8348	2.529	8.9643	1

. . . . [1 mark]

The new cluster centres are

$$C_1^{(2)} = (4.5714, 1.2883875, -0.4494625, 2.5379) \quad C_2^{(2)} = (-0.94075, 1.10205, 3.8306, 1.4025)$$

[0.5 mark]

Step 3 : Update table based on distance to cluster centres

V1	V2	V3	V4	dist.1	dist.2	clust.centre
-0.3323	0.7264	2.4691	1.8429	5.7761	1.5997	2
5.5783	5.7211	-3.3731	3.9209	5.5788	11.0485	1
-1.5492	1.4777	5.1921	0.9621	8.474	1.5997	2
8.0669	-1.1127	1.2409	-0.1392	5.2923	9.7533	1
-0.294	-0.5842	0.7708	1.6414	5.4288	3.5611	2
5.5741	3.4215	0.9827	3.8443	3.0518	7.8674	1
-1.838	0.5629	-3.898	4.483	7.5685	8.3855	1
2.6957	-0.2016	0.6947	0.6821	3.239	5.0275	1
10.7553	0.1658	-0.8895	3.0359	6.32	12.7523	1
6.0329	2.3343	0.8758	2.8348	2.2526	7.8059	1

The new cluster centres are

$$C_1^{(3)} = (5.2665, 1.5559, -0.6238, 2.6660) \quad C_2^{(3)} = (-0.7252, 0.5400, 2.8107, 1.4821)$$

[0.5 mark]

Depending how one understands the last question, from Step 2 to Step 3, we find that the **k-means does not converge**. From Step 3 to Step 4, the same applies as illustrated below. .... [0.5 mark]

Step 4 : Update table based on distance to cluster centres

V1	V2	V3	V4	dist.1	dist.2	clust.centre
-0.3323	0.7264	2.4691	1.8429	6.5021	0.6602	2
5.5783	5.7211	-3.3731	3.9209	5.1556	10.5245	1
-1.5492	1.4777	5.1921	0.9621	9.1207	2.7386	2
8.0669	-1.1127	1.2409	-0.1392	5.1293	9.2263	1
-0.294	-0.5842	0.7708	1.6414	6.2043	2.374	2
5.5741	3.4215	0.9827	3.8443	2.7467	7.5436	1
-1.838	0.5629	-3.898	4.483	8.0921	7.4331	2
2.6957	-0.2016	0.6947	0.6821	3.9207	4.1677	1
10.7553	0.1658	-0.8895	3.0359	5.6804	12.1674	1
6.0329	2.3343	0.8758	2.8348	1.863	7.38	1

The new cluster centres are

$$C_1^{(4)} = (6.4505, 1.7214, -0.0781, 2.3631) \quad C_2^{(4)} = (-1.003375, 0.5457, 1.1335, 2.23235)$$

□

4. (May 2020 Final Q3(a)) Given the unlabelled data in Table 3.1.

Table 3.1: Unlabelled data.

	V1	V2	V3
1	7.5205	4.6564	-0.1947
2	-1.1824	-1.1174	1.8383
3	-0.3576	-0.4739	-1.1603
4	-1.422	-0.5891	-0.8287
5	3.2287	0.7141	0.6208
6	3.2926	3.1609	2.7553
7	8.2304	3.8832	-1.7378
8	4.2079	0.4964	4.361
9	3.8443	5.7565	1.0293
10	1.493	3.525	-2.9904

Use the k-means algorithm with $k = 2$ (unsupervised learning) to find the final cluster centres if the **first** and **sixth** rows are chosen as the **initial cluster centres**. (4 marks)

Solution. Given the initial centres:

V1	V2	V3
7.5205	4.6564	-0.1947
3.2926	3.1609	2.7553

Step 1 : Update table based on distance to cluster centres

V1	V2	V3	dist.1	dist.2	clust.centre
7.5205	4.6564	-0.1947	0	5.3679	1
-1.1824	-1.1174	1.8383	10.64	6.2586	2
-0.3576	-0.4739	-1.1603	9.4508	6.4705	2
-1.422	-0.5891	-0.8287	10.3868	7.0096	2
3.2287	0.7141	0.6208	5.8844	3.2476	2
3.2926	3.1609	2.7553	5.3679	0	2
8.2304	3.8832	-1.7378	1.8663	6.715	1
4.2079	0.4964	4.361	7.0024	3.2428	2
3.8443	5.7565	1.0293	4.0278	3.1655	2
1.493	3.525	-2.9904	6.7399	6.0319	2

. . . . [1.5 marks]

The new cluster centres are

$$C_1^{(1)} = (7.87545, 4.2698, -0.96625), \quad C_2^{(1)} = (1.6380625, 1.4340625, 0.7031625) \quad [0.5 \text{ mark}]$$

Step 2 : Update table based on distance to cluster centres

V1	V2	V3	dist.1	dist.2	clust.centre
7.5205	4.6564	-0.1947	0.9331	6.767	1
-1.1824	-1.1174	1.8383	10.9056	3.9691	2
-0.3576	-0.4739	-1.1603	9.5039	3.331	2
-1.422	-0.5891	-0.8287	10.4914	3.9754	2
3.2287	0.7141	0.6208	6.0625	1.7479	2
3.2926	3.1609	2.7553	6.0068	3.1513	2
8.2304	3.8832	-1.7378	0.9331	7.4442	1
4.2079	0.4964	4.361	7.4879	4.5676	2
3.8443	5.7565	1.0293	4.7374	4.8639	1
1.493	3.525	-2.9904	6.737	4.2468	2

. . . . [0.5 mark]

The new cluster centres are

$$C_1^{(2)} = (6.5317, 4.7654, -0.3011), \quad C_2^{(2)} = (1.3229, 0.8166, 0.6566) \qquad [0.5 \text{ mark}]$$

Step 3 : Update table based on distance to cluster centres

V1	V2	V3	dist.1	dist.2	clust.centre
7.5205	4.6564	-0.1947	1.0004	7.3403	1
-1.1824	-1.1174	1.8383	9.9344	3.3783	2
-0.3576	-0.4739	-1.1603	8.6978	2.7911	2
-1.422	-0.5891	-0.8287	9.6026	3.4229	2
3.2287	0.7141	0.6208	5.3078	1.9089	2
3.2926	3.1609	2.7553	4.7337	3.7122	2
8.2304	3.8832	-1.7378	2.3933	7.9279	1
4.2079	0.4964	4.361	6.7349	4.7062	2
3.8443	5.7565	1.0293	3.1582	5.5587	1
1.493	3.525	-2.9904	5.8446	4.5459	2

. . . . [0.5 mark]

There is no change in the clustering, the final cluster centres are

$$C_1(6.5317, 4.7654, -0.3011), \quad C_2(1.3229, 0.8166, 0.6566) \qquad [0.5 \text{ mark}]$$

□

5. (Final Exam Jan 2023, Q3(c), 13 marks) Given the three-dimensional data in Table 3.3.

Obs.	x_1	x_2	x_3
A	1	4	3
B	2	6	2
C	4	7	3
D	7	0	2
E	9	3	3
F	8	1	2
G	1	6	3

Table 3.3: Three-dimensional data for clustering.

Perform k-means clustering algorithm (using the Euclidean distance) on the data from Table 3.3 with A and G as the initial centres until **two clusters** are found. Write down the stable cluster centres. You may round the numbers in your calculations to 4 decimal places.

Solution. Given the initial centres: $A(1, 4, 3)$, $G(1, 6, 3)$

Step 1 : Update table based on distance to cluster centres

x_1	x_2	x_3	dist.1	dist.2	clust.centre
1	4	3	0	2	1
2	6	2	2.4495	1.4142	2
4	7	3	4.2426	3.1623	2
7	0	2	7.2801	8.544	1
9	3	3	8.0623	8.544	1
8	1	2	7.6811	8.6603	1
1	6	3	2	0	2

. . . . [5 marks]

The new cluster centres are

$$Centre_1 = (6.25, 2, 2.5)$$
$$Centre_2 = (2.333333, 6.333333, 2.666667)$$

[1 mark]

Step 2 : Update table based on distance to cluster centres

x_1	x_2	x_3	dist.1	dist.2	clust.centre
1	4	3	5.6403	2.7080	2
2	6	2	5.8577	0.8165	2
4	7	3	5.5057	1.8257	2
7	0	2	2.1937	7.8951	1
9	3	3	2.9686	7.4610	1
8	1	2	2.0767	7.8102	1
1	6	3	6.6191	1.4142	2

. . . . [3 marks]

The new cluster centres are

$$Centre_1 = (8, 1.333333, 2.333333)$$
$$Centre_2 = (2, 5.75, 2.75)$$

[1 mark]

Step 3 : Update table based on distance to cluster centres

x_1	x_2	x_3	dist.1	dist.2	clust.centre
1	4	3	7.5203	2.0310	2
2	6	2	7.6085	0.7906	2
4	7	3	6.9682	2.3717	2
7	0	2	1.6997	7.6567	1
9	3	3	2.0548	7.5250	1
8	1	2	0.4714	7.6893	1
1	6	3	8.4393	1.0607	2

. . . . [2 marks]

The stable cluster centres are

$$C_1(8, 1.333333, 2.333333), \quad C_2(2, 5.75, 2.75)$$ [1 mark]

Average: 9.93 / 13 marks in Jan 2023; 16% below 6.5 marks. □

6. (Final Exam May 2023, Q3(c)) Given the three-dimensional data in Table 3.3.

Obs.	x_1	x_2	x_3
A	5	3	8
B	4	1	6
C	3	2	6
D	4	4	9
E	2	1	6
F	3	1	8
G	5	5	8

Table 3.3: Three-dimensional data.

Perform k-means clustering algorithm using the Euclidean distance on the data from Table 3.3 with B and D as the initial centres until **two clusters** are found.

(a) Write down the stable cluster centres. You may round the numbers in your calculations to 4 decimal places. (9 marks)

Solution. Given the initial centres B(4, 1, 6), D(4, 4, 9) which correspond to cluster 1 and cluster 2.
Step 1 : Update table based on distance to cluster centres

x_1	x_2	x_3	dist.1	dist.2	clust.centre
5	3	8	3	1.7321	2
4	1	6	0	4.2426	1
3	2	6	1.4142	3.7417	1
4	4	9	4.2426	0	2
2	1	6	2	4.6904	1
3	1	8	2.2361	3.3166	1
5	5	8	4.5826	1.7321	2

.... [5 marks]

The new cluster centres are $Centre1 = (3, 1.25, 6.5)$, $Centre2 = (4.6667, 4, 8.3333)$.
Step 2 : Update table based on distance to cluster centres

x_1	x_2	x_3	dist.1	dist.2	clust.centre
5	3	8	3.0516	1.1055	2
4	1	6	1.1456	3.8586	1
3	2	6	0.9014	3.496	1
4	4	9	3.8487	0.9428	2
2	1	6	1.1456	4.6428	1
3	1	8	1.5207	3.448	1
5	5	8	4.5069	1.1055	2

.... [3 marks]

The stable cluster centres are $C_1(3, 1.25, 6.5)$, $C_2(4.6667, 4, 8.3333)$. . . [1 mark] □

(b) Write down the within cluster sum of squares for the two stable clusters you found in part (i). (4 marks)

Solution. For the stable cluster centre (3,1.25,6.5),

$$WSS_1 = 1.1456^2 + 0.9014^2 + 1.1456^2 + 1.5207^2 = 5.7498 \approx 5.75$$ [2 marks]

For the stable cluster centre (4.6667, 4, 8.3333),

$$WSS_2 = 1.1055^2 + 0.9428^2 + 1.1055^2 \approx 3.3333$$ [2 marks]

□

7. (Final Exam May 2024, Q3(b), 9 marks) Given the two-dimensional data in Table 3.2.

3.2: Two-dimensional data

Obs.	x_1	x_2
A	3.3	2.5
B	-0.3	-0.4
C	-0.4	-2.8
D	7.9	4.2
E	3	6.7

Perform k-means clustering algorithm using the Euclidean distance on the data from Table 3.2 with B and D as the initial centres until **two clusters** are found. Write down the stable cluster centres. You may round the numbers in your calculations to 4 decimal places.

Solution. Given the initial centres B(-0.3, -0.4), D(7.9, 4.2) which correspond to cluster 1 and cluster 2.

Step 1 : Update table based on distance to cluster centres

x_1	x_2	dist.1	dist.2	clust.centre
3.3	2.5	4.6228	4.9041	1
-0.3	-0.4	0	9.4021	1
-0.4	-2.8	2.4021	10.8577	1
7.9	4.2	9.4021	0	2
3	6.7	7.8294	5.5009	2

.... [2 marks]

The new cluster centres are

$$Centre1 = (0.8667, -0.2333), \quad Centre2 = (5.45, 5.45)$$ [1 mark]

Step 2 : Update table based on distance to cluster centres

x_1	x_2	dist.1	dist.2	clust.centre
3.3	2.5	3.6595	3.6503	2
-0.3	-0.4	1.1785	8.2027	1
-0.4	-2.8	2.8622	10.1136	1
7.9	4.2	8.314	2.7505	2
3	6.7	7.2541	2.7505	2

.... [2 marks]

The new cluster centres are

$$Centre1 = (-0.35, -1.6), \quad Centre2 = (4.7333, 4.4667)$$ [1 mark]

Step 3 : Update table based on distance to cluster centres

x_1	x_2	dist.1	dist.2	clust.centre
3.3	2.5	5.4893	2.4336	2
-0.3	-0.4	1.201	7.0013	1
-0.4	-2.8	1.201	8.8969	1
7.9	4.2	10.0848	3.1779	2
3	6.7	8.9506	2.8271	2

.... [2 marks]

The stable cluster centres are

$$Centre1 = (-0.35, -1.6), \quad Centre2 = (4.7333, 4.4667)$$ [1 mark]

Average: 6.75 / 9 marks in May 2024. Reason: Careless mistakes starting from the second or third steps. □

8. (Final Exam May 2024, Q3(c)) Lloyd's algorithm is an algorithm for finding the k-means ($k > 1$) clusters for a finite data.

   (a) Will the Lloyd's algorithm always converge to exactly k stable cluster centres when k centres are choosen by the user? Justify your answer. (2 marks)

   > *Solution.* No. .... [0.5 mark]
   >
   > For the poorly choosen initial centres, Lloyd's always may converge to less than k cluster centres. .... [1.5 marks]
   >
   > **Average**: 0.32 / 2 marks in May 2024. Reason: Did not work on Tutorial 9 Q1. □

   (b) Will the Lloyd's algorithm always find the optimal clusters, i.e. finding the set of clusters with minimum sum of within sum of squares? Justify your answer. (2 marks)

   > *Solution.* No. .... [0.5 mark]
   >
   > The algorithm might convert to local minimum. .... [1.5 marks]
   >
   > **Average**: 0.20 / 2 marks in May 2024. Reason: Did not practise with practical 11. □