Tut 9: k-Means Clustering

June 2023

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?

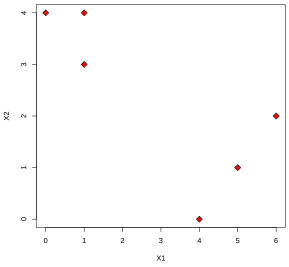
   Solution. No. It can never give more than k clusters, since at every stage every point is assigned to one of k clusters. □

   (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	A
5.5783	5.7211	-3.3731	3.9209	9.9162	12.2851	A
-1.5492	1.4777	5.1921	0.9621	3.1993	0	B
8.0669	-1.1127	1.2409	-0.1392	8.9088	10.7705	A
-0.294	-0.5842	0.7708	1.6414	2.155	5.0829	A
5.5741	3.4215	0.9827	3.8443	6.9544	8.9747	A
-1.838	0.5629	-3.898	4.483	7.0572	9.7952	A
2.6957	-0.2016	0.6947	0.6821	3.8113	6.4144	A
10.7553	0.1658	-0.8895	3.0359	11.6599	13.943	A
6.0329	2.3343	0.8758	2.8348	6.8281	8.9643	A

. . . . [1.5 marks]

The new cluster centres are

V1	V2	V3	V4
4.0265	1.2259	-0.1252	2.4607
-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	B
5.5783	5.7211	-3.3731	3.9209	5.941	12.2851	A
-1.5492	1.4777	5.1921	0.9621	7.8531	0	B
8.0669	-1.1127	1.2409	-0.1392	5.5154	10.7705	A
-0.294	-0.5842	0.7708	1.6414	4.8392	5.0829	A
5.5741	3.4215	0.9827	3.8443	3.2183	8.9747	A
-1.838	0.5629	-3.898	4.483	7.2908	9.7952	A
2.6957	-0.2016	0.6947	0.6821	2.7649	6.4144	A
10.7553	0.1658	-0.8895	3.0359	6.8786	13.943	A
6.0329	2.3343	0.8758	2.8348	2.529	8.9643	A

. . . . [1 mark]

The new cluster centres are

V1	V2	V3	V4
4.5714	1.2883875	-0.4494625	2.5379
-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	B
5.5783	5.7211	-3.3731	3.9209	5.5788	11.0485	A
-1.5492	1.4777	5.1921	0.9621	8.474	1.5997	B
8.0669	-1.1127	1.2409	-0.1392	5.2923	9.7533	A
-0.294	-0.5842	0.7708	1.6414	5.4288	3.5611	B
5.5741	3.4215	0.9827	3.8443	3.0518	7.8674	A
-1.838	0.5629	-3.898	4.483	7.5685	8.3855	A
2.6957	-0.2016	0.6947	0.6821	3.239	5.0275	A
10.7553	0.1658	-0.8895	3.0359	6.32	12.7523	A
6.0329	2.3343	0.8758	2.8348	2.2526	7.8059	A

The new cluster centres are

V1	V2	V3	V4
5.2665	1.5559	-0.6238	2.6660
-0.7252	0.5400	2.8107	1.4821

. . . . [1 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	B
5.5783	5.7211	-3.3731	3.9209	5.1556	10.5245	A
-1.5492	1.4777	5.1921	0.9621	9.1207	2.7386	B
8.0669	-1.1127	1.2409	-0.1392	5.1293	9.2263	A
-0.294	-0.5842	0.7708	1.6414	6.2043	2.374	B
5.5741	3.4215	0.9827	3.8443	2.7467	7.5436	A
-1.838	0.5629	-3.898	4.483	8.0921	7.4331	B
2.6957	-0.2016	0.6947	0.6821	3.9207	4.1677	A
10.7553	0.1658	-0.8895	3.0359	5.6804	12.1674	A
6.0329	2.3343	0.8758	2.8348	1.863	7.38	A

The new cluster centres are

V1	V2	V3	V4
6.4505	1.7214	-0.0781	2.3631
-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	A
-1.1824	-1.1174	1.8383	10.64	6.2586	B
-0.3576	-0.4739	-1.1603	9.4508	6.4705	B
-1.422	-0.5891	-0.8287	10.3868	7.0096	B
3.2287	0.7141	0.6208	5.8844	3.2476	B
3.2926	3.1609	2.7553	5.3679	0	B
8.2304	3.8832	-1.7378	1.8663	6.715	A
4.2079	0.4964	4.361	7.0024	3.2428	B
3.8443	5.7565	1.0293	4.0278	3.1655	B
1.493	3.525	-2.9904	6.7399	6.0319	B

....[1.5 marks]

The new cluster centres are

V1	V2	V3
7.87545	4.2698	-0.96625
1.6380625	1.4340625	0.7031625

......[0.5 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	A
-1.1824	-1.1174	1.8383	10.9056	3.9691	B
-0.3576	-0.4739	-1.1603	9.5039	3.331	B
-1.422	-0.5891	-0.8287	10.4914	3.9754	B
3.2287	0.7141	0.6208	6.0625	1.7479	B
3.2926	3.1609	2.7553	6.0068	3.1513	B
8.2304	3.8832	-1.7378	0.9331	7.4442	A
4.2079	0.4964	4.361	7.4879	4.5676	B
3.8443	5.7565	1.0293	4.7374	4.8639	A
1.493	3.525	-2.9904	6.737	4.2468	B

....[0.5 mark]

The new cluster centres are

V1	V2	V3
6.53173333333333	4.76536666666667	-0.301066666666667
1.32288571428571	0.816571428571429	0.656571428571428

....[0.5 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	A
-1.1824	-1.1174	1.8383	9.9344	3.3783	B
-0.3576	-0.4739	-1.1603	8.6978	2.7911	B
-1.422	-0.5891	-0.8287	9.6026	3.4229	B
3.2287	0.7141	0.6208	5.3078	1.9089	B
3.2926	3.1609	2.7553	4.7337	3.7122	B
8.2304	3.8832	-1.7378	2.3933	7.9279	A
4.2079	0.4964	4.361	6.7349	4.7062	B
3.8443	5.7565	1.0293	3.1582	5.5587	A
1.493	3.525	-2.9904	5.8446	4.5459	B

....[0.5 mark]

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

V1	V2	V3
6.53173333333333	4.76536666666667	-0.301066666666667
1.32288571428571	0.816571428571429	0.656571428571428

. . . . [0.5 mark]

□

5. (Final Exam Jan 2023, Q3(c)) 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. (13 marks)

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

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	B

. . . . [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

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

[1 mark]

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