Tut 9: k-Means Clustering

May/June 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?

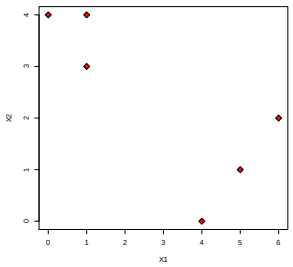
   Solution. No. It can never give more than k clusters, since at every stage every point is assigned to one of k clusters. □

   (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

□