Assignment 4 – Solution

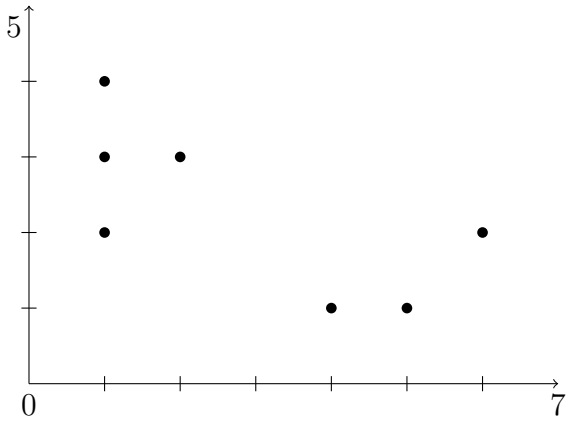
Machine Learning
MSc Business Analytics

Wolfram Wiesemann

1 Individual Assignment

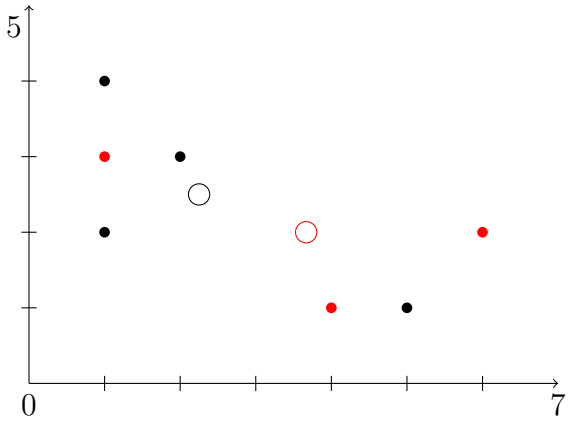
1. *Plot the observations in a two-dimensional graph.*

The graph looks as follows:



2. *Perform K-means clustering with $K = 2$ using the Euclidean norm. Toss a coin 7 times to initialise the algorithm.*

First we assign randomly $C_1 = \{2, 6, 7\}$ (red) and $C_2 = \{1, 3, 4, 5\}$ (black):



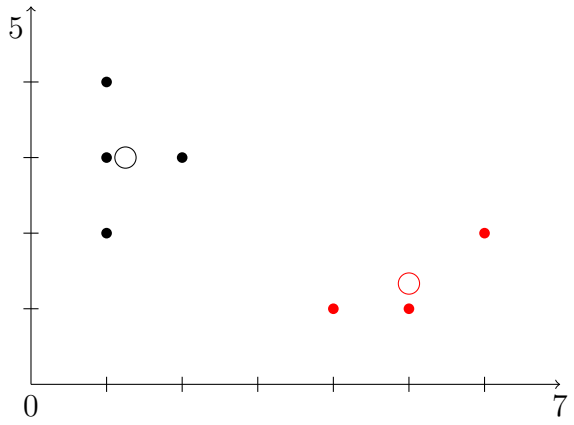
We compute the centroids of the two classes as $\boldsymbol{c}_1 = (\frac{11}{3}, 2)$ and $\boldsymbol{c}_2 = (\frac{9}{4}, \frac{5}{2})$. We now recompute the distances:

Obs. i	x_{i1}	x_{i2}	dist$(\boldsymbol{x}_i, \boldsymbol{c}_1)$	dist$(\boldsymbol{x}_i, \boldsymbol{c}_2)$
1	1	4	2.84	1.95
2	1	3	2.84	1.95
3	1	2	4.01	2.79
4	5	1	1.33	2.79
5	2	3	4.33	3.5
6	6	2	3.07	4.03
7	4	1	2.02	3.05

After reassignment, the new clusters are $C_1 = \{4, 6, 7\}$ and $C_2 = \{1, 2, 3, 5\}$. The new centroids of the two clusters are $\boldsymbol{c}_1 = (5, \frac{4}{3})$ and $\boldsymbol{c}_2 = (\frac{5}{4}, 3)$.

Obs. i	x_{i1}	x_{i2}	dist$(\boldsymbol{x}_i, \boldsymbol{c}_1)$	dist$(\boldsymbol{x}_i, \boldsymbol{c}_2)$
1	1	4	4.01	2.01
2	1	3	4.01	2.01
3	1	2	5.42	2.01
4	5	1	0.67	3.88
5	2	3	5.55	3.09
6	6	2	2.85	4.85
7	4	1	1.66	4.06

The clusters are still $C_1 = \{4, 6, 7\}$ and $C_2 = \{1, 2, 3, 5\}$. The algorithm thus terminates with the following result:



3. *Cluster the data using hierarchical clustering with complete linkage and the Euclidean norm. Draw the resulting dendrogram.*

We calculate the following pairwise distances between the observations:

	{1}	{2}	{3}	{4}	{5}	{6}	{7}
{1}	**0**						
{2}	1	**0**					
{3}	2	1	**0**				
{4}	5	4.47	4.12	**0**			
{5}	1.41	1	1.41	3.6	**0**		
{6}	5.38	5.09	5	1.41	4.12	**0**	
{7}	4.24	3.6	3.16	1	2.82	2.23	**0**

(Empty cells can be inferred from symmetry.) We first merge the 'clusters' {1} and {2}:

	{1, 2}	{3}	{4}	{5}	{6}	{7}
{1, 2}	**0**					
{3}	2	**0**				
{4}	5	4.12	**0**			
{5}	1.41	1.41	3.6	**0**		
{6}	5.38	5	1.41	4.12	**0**	
{7}	4.24	3.16	1	2.82	2.23	**0**

We now merge the 'clusters' {4} and {7}:

	{1, 2}	{3}	{5}	{6}	{4, 7}
{1, 2}	**0**				
{3}	2	**0**			
{5}	1.41	1.41	**0**		
{6}	5.38	5	4.12	**0**	
{4, 7}	5	4.12	3.6	2.23	**0**

We now merge the 'clusters' {5} and {1, 2}:

	{1, 2, 5}	{3}	{6}	{4, 7}
{1, 2, 5}	**0**			
{3}	2	**0**		
{6}	5.38	5	**0**	
{4, 7}	5	4.12	2.23	**0**

We now merge the 'clusters' {3} and {1, 2, 5}:

	{1, 2, 3, 5}	{6}	{4, 7}
{1, 2, 3, 5}	**0**		
{6}	5.38	**0**	
{4, 7}	5	2.23	**0**

We now merge the 'clusters' {6} and {4, 7}:

	{1, 2, 3, 5}	{4, 6, 7}
{1, 2, 3, 5}	**0**	
{4, 6, 7}	5.38	**0**

After merging the clusters $\{1, 2, 3, 5\}$ and $\{4, 6, 7\}$, we obtain the following result: