Exercises: Clustering

October 16, 2017

1 K-means Clustering

Consider the following data set consisting of the scores of two variables on each of seven individuals:

Subject	A	B
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Given $K = 2$ and $(1.0, 1.0)$, $(5.0, 7.0)$ as the initial selection of points as centroids, perform 2 iterations of the K-means clustering algorithm.
Draw the data points and cluster centroids in a coordinate system.

Algorithm

- Assign each point to the closest centroid. The Euclidean distance between two objects, with n attributes, $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$ is computed as follows:

$$d(x, y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2} \tag{1}$$

  E.g., the Euclidean distance between two points $(2, 4)$ and $(3, 1)$ is $\sqrt{(2-3)^2 + (4-1)^2} = \sqrt{10}$.

- Recompute the centroid of each cluster. The centroid of the ith cluster is defined as follows:

$$\mathbf{c}_i = \frac{1}{m_i} \sum_{x \in C_i} \mathbf{x}, \tag{2}$$

  where m_i is the number of data points in cluster i. E.g., the centroid of a cluster containing three 2-dimensional points, $(1, 1)$, $(2, 3)$, and $(6, 2)$ is $((1 + 2 + 6)/3, (1 + 3 + 2)/3) = (3, 2)$.

Solution

Round numbers to the first digit.

The underline indicates which centroid the point was assigned to. The cluster centroids in the next iteration are updated based on the points assigned to that cluster.

Subject	Distance to centroid of Cluster 1 (1.0, 1.0)	Distance to centroid of Cluster 2 (5.0, 7.0)
1	<u>0.0</u>	7.2
2	<u>1.1</u>	6.1
3	<u>3.6</u>	3.6
4	7.2	<u>0.0</u>
5	4.7	<u>2.5</u>
6	5.3	<u>2.0</u>
7	4.3	<u>2.9</u>

Table 1: Iteration 1

Subject	Distance to centroid of Cluster 1 (1.8,2.3)	Distance to centroid of Cluster 2 (4.1,5.4)
1	<u>1.5</u>	5.4
2	<u>0.4</u>	4.3
3	2.1	<u>1.8</u>
4	5.7	<u>1.8</u>
5	3.2	<u>0.7</u>
6	3.8	<u>0.6</u>
7	2.8	<u>1.1</u>

Table 2: Iteration 2

The colors of data points indicate the cluster they have been assigned to. The crosses show the cluster centroids.

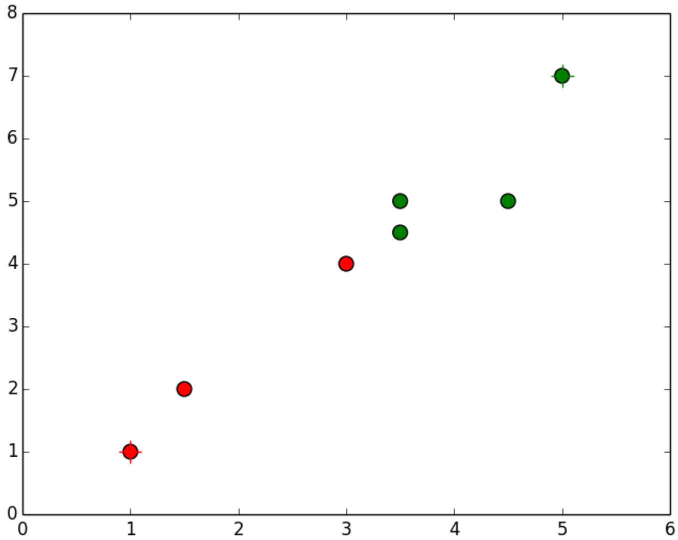

Figure 1: Iteration 1

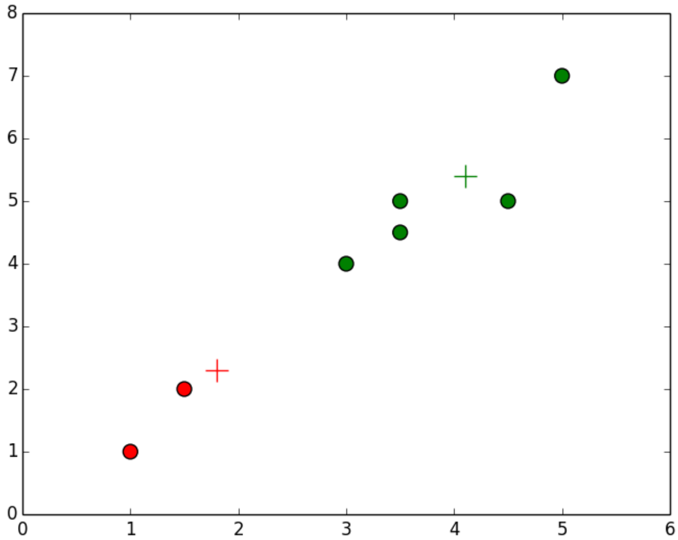

Figure 2: Iteration 2

2 Hierarchical Agglomerative Clustering

Perform a hierarchical clustering of some Italian cities, based on their distances, using the single-linkage method.
Draw a dendogram after each step, showing the distance on the y-axis.

Algorithm

- Compute the proximity matrix and begin with a disjoint clustering.
- Find the most similar pair of clusters in the current clustering, c_i and c_j, based on $\min d(c_i, c_j)$.
- Merge c_i and c_j into a single cluster and update the proximity matrix.
  - Delete the rows and columns corresponding to c_i and c_j and add a new row and new column for the merged cluster c_k.
  - The proximity between the merged cluster (c_i, c_j) and an other (old) cluster c_k is: $d(c_k, (c_i, c_j)) = \min\{d(c_k, c_i), d(c_k, c_j)\}$.
- Repeat until all objects are in a single cluster.

Solution

Init: Compute the proximity matrix (This is already given.)
Each item corresponds to a (singleton) cluster.

	BA	**FI**	**MI**	**NA**	**RM**	**TO**
BA	0	662	877	255	412	996
FI	662	0	295	468	268	400
MI	877	295	0	754	564	138
NA	255	468	754	0	219	869
RM	412	268	564	219	0	669
TO	996	400	138	869	669	0

Repeat: Find the nearest pair of cities, merge them, and update the proximity matrix.
The proximities are symmetric so it is enough to fill in the upper triangle of the proximity matrix.

	BA	FI	MI,TO	NA	RM
BA	0	662	877	255	412
FI		0	295	468	268
MI,TO			0	754	564
NA				0	219
RM					0

Table 3: Step 1 (5 clusters) The closest two clusters are merged (minimal value in the initial proximity matrix): MI and TO

	BA	FI	MI,TO	NA,RM
BA	0	662	877	255
FI		0	295	268
MI,TO			0	564
NA,RM				0

Table 4: Step 2 (4 clusters) The closest two clusters are merged (minimal value in Table 1): NA and RM

	BA,NA,RM	FI	MI,TO
BA,NA,RM	0	268	564
FI		0	295
MI,TO			0

Table 5: Step 3 (3 clusters) The closest two clusters are merged (minimal value in Table 2): BA and NA,RM

	BA,FI,NA,RM	MI,TO
BA,FI,NA,RM	0	295
MI,TO		0

Table 6: Step 4 (2 clusters) The closest two clusters are merged (minimal value in Table 3): FI and BA,NA,RM

In the last step we merge BA,FI,NA,RM and MI,TO and arrive at a single cluster.

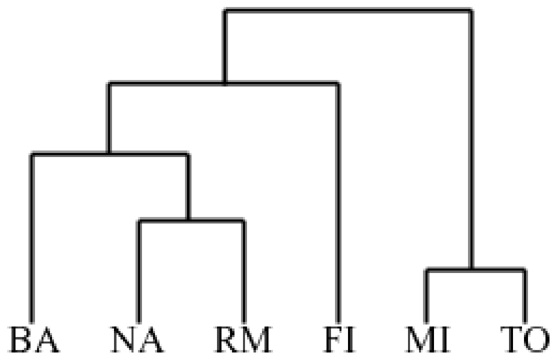

Figure 3: Dendogram