Exercises: Clustering

September 21, 2016

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, $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ 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.

Subject	Distance to centroid of Cluster 1 (1.0, 1.0)	 of Cluster 2 (5.0, 7.0)
1	0.0	7.2
2	1.1	6.1
3		
4		
5		
6		
7		

Table 1: Iteration 1

Subject	Distance to centroid of Cluster 1 (,)	 of Cluster 2 (,)
1		
2		
3		
4		
5		
6		
7		

Table 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

MI
TO
FI
RM
NA
BA

Repeat: Find the nearest pair of cities, merge them, and update the proximity matrix.

	0				
		0			
			0		
				0	
					0

Table 3: Step 1 (5 clusters)

	0			
		0		
			0	
				0

Table 4: Step 2 (4 clusters)

	0		
		0	
			0

Table 5: Step 3 (3 clusters)

	0	
		0

Table 6: Step 4 (2 clusters)