Authors: Mhd Jawad Al Rahwanji, Wojciech Ptaś 7038980, 7042843

Problem 4

1.

To prove this equation:

$\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2 = 2\sum_{i\in C_k}\sum_{j=1}^{p}(x_{ij}-\hat{x}_{kj})^2$

Let's first investigate lef-hand side of this equation:

$\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2 =$
$\frac{1}{|C_k|}\sum_{i\in C_k}\sum_{i'\in C_k}\sum_{j=1}^{p}(x_{ij}^2-2x_{ij}x_{i'j}+x_{i'j}^2) =$

$= \sum_{i\in C_k}\sum_{j=1}^{p}\frac{1}{|C_k|}|C_k|x_{ij}^2 - \sum_{i\in C_k}\sum_{j=1}^{p}2x_{ij}\hat{x}_{kj} + \sum_{i\in C_k}\sum_{j=1}^{p}x_{ij}^2 = \sum_{i\in C_k}\sum_{j=1}^{p}x_{ij}^2 - 2|C_k|\sum_{j=1}^{p}\hat{x}_{kj}^2 + \sum_{i\in C_k}\sum_{j=1}^{p}x_{ij}^2 =$

$= 2\sum_{i\in C_k}\sum_{j=1}^{p}x_{ij}^2 - 2|C_k|\sum_{j=1}^{p}\hat{x}_{kj}^2$

Now, let's investigate the right-hand side of the equation:

$2\sum_{i\in C_k}\sum_{j=1}^{p}(x_{ij}-\hat{x}_{kj})^2 =$
$= 2\sum_{i\in C_k}\sum_{j=1}^{p}x_{ij}^2 - 4\sum_{i\in C_k}\sum_{j=1}^{p}x_{ij}\hat{x}_{kj} + \sum_{i\in C_k}\sum_{j=1}^{p}\hat{x}_{kj}^2 = = 2\sum_{i\in C_k}\sum_{j=1}^{p}x_{ij}^2 - 4|C_k|\sum_{j=1}^{p}\hat{x}_{kj}^2 + 2|C_k|\sum_{j=1}^{p}\hat{x}_{kj}^2 =$

$= 2\sum_{i\in C_k}\sum_{j=1}^{p}x_{ij}^2 - 2|C_k|\sum_{j=1}^{p}\hat{x}_{kj}^2$

We shown that both sides of the equation are equal, which concludes the proof.

a)

In each iteration of the algorithm, we assign each point to the closest centroid, which decreases the right-hand side of the equation (distance between the centroid and every point in the cluster). Because we have proved the equality above, we can say that this algorithm also decreases the left-hand side of this equation, which means it also decreases the presented objective.

b)

This equality means, that the average distance between each two points in the cluster is equal to the double sum of distances between the centroid and every point of the cluster. Calculating the centroid and than the distance between it and each point is requires much less computation than calculating distance between each pair of points, it also allows us to use algorithm described in the lecture to perform K-Means clustering.

2.

For the data plotted shown in *Figure 1*, we should not use k-means clustering. This algorithm minimizes the variance inside each cluster, so it leads to compacted, not spectral clusters.

To cluster this data, we should use hierarchical clustering with single linkage. This linkage is the best for our purpose, beacuse it takes the minimum distance between two clusters and in this scenario, the points that form the rings, are always closer to each other than to other rings (other clusters).