Prof. Dr. Vladimir Shikhman
Professur für Wirtschaftsmathematik
Technische Universität Chemnitz

Übungsleiter: David Müller
david.mueller@mathematik.tu-chemnitz.de

Mathematical Foundation of Big Data Analytics (SS 2019)
Clustering

Ex. 1 You have the following two-dimensional dataset:

$$X = \{(1,1)^T, (2,1)^T, (4,3)^T, (5,4)^T\}$$

Your task is to apply the k-means algorithm on your dataset. In order to do so recall that you have to find proper centers z_ℓ for each cluster C_ℓ, $\ell = 1, \ldots, k$.

a) Take the euclidean dissimilarity measure $d_2(x,z) = \sum_{i=1}^{p}(x_i - z_i)^2$. Calculate

$$\arg\min_z \sum_{x \in C_\ell} d_2(x,z).$$

b) Apply k-means algorithm to X with $k = 2$ and the euclidean dissimilarity measure. Initialize the two cluster centers as the points, which have the largest distance to each other.

c) You receive a new training point $x^{(5)} = (0,6)$. Restart your algorithm from a) and discuss the result.

Ex. 2 You try to improve the result of exercise 1c). Therefore you choose to try another dissimilarity measure

$$d_1(x,z) = \sum_{i=1}^{p} |x_i - z_i|.$$

a) Calculate the cluster centers

$$\arg\min_z \sum_{x \in C_\ell} d_1(x,z)$$

b) Apply k-means algorithm with $k = 2$ and $d_1(x,z)$ for the dataset of exercise 1b).

Ex. 3 Show that for every column stochastic matrix P it holds

$$\lambda_{\max} = 1,$$

where $\lambda_{\max}$ is the largest eigenvalue of the matrix P.