University of Southern Denmark
IMADA
Arthur Zimek

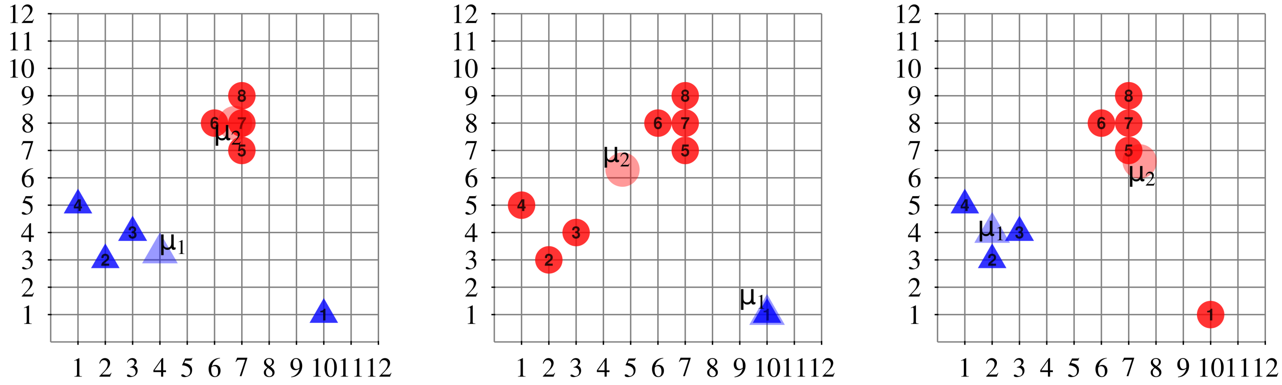
DM868/DM870/DS804: Data Mining and Machine Learning

Spring term 2023

Exercise 5: Clustering: k-means and Silhouette

Exercise 5-1 Silhouette Coefficient (1 point)

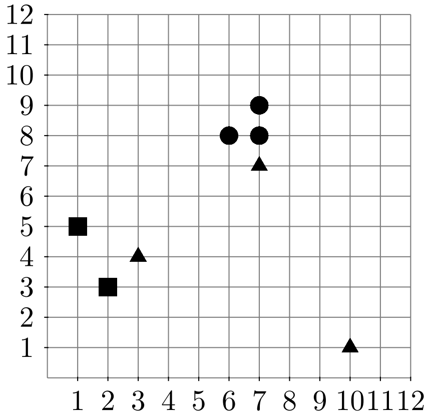
We derived three different clustering solutions for the toy data set in the lecture:



Compute the simplified silhouette coefficient for each solution. Compare the result with the ranking by the k-means objective function (TD^2), that we determined in the lecture.

Exercise 5-2 k-means, choice of k, and compactness (1 point)

Given the following data set with 8 objects (in $\mathbb{R}^2$) as in the lecture:



Compute a complete partitioning of the data set into $k = 3$ clusters using the basic k-means algorithm (due to Forgy and Lloyd). The initial assignment of objects to clusters is given using the triangle, square, and circle markers.

Objects x are assigned to the cluster with the least increase in squared deviations $SSQ(x, c)$ where c is the cluster center.

$$SSQ(x,c) = \sum_{i=1}^{d} |x_i - c_i|^2$$

Start with computing the initial centroids, and draw the cluster assignments after each step and explain the step. Remember to use the least squares assignment!

Give the final quality of the clustering (TD^2). How does it compare with the solutions for $k = 2$ discussed in the lecture? Can we conclude on $k = 3$ or $k = 2$ being the better parameter choice on this data set?

Also compute solutions with $k = 4$, $k = 5$, starting from some random initial assignments of objects to clusters. What do you observe in terms of the TD^2 measure?

Exercise 5-3 Silhouette and k-means implementations in scikit-learn (1 point)

Explore the code on `http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html`. To do this as lab session in class, bring your laptop with Python, NumPy, SciPy, and SciKit-Learn installed.

(a) What is the termination criterion in k-means in the scikit-learn implementation?

(b) Why can we get negative Silhouettes in this example?