Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Thomas Seidl
Julian Busch, Evgeniy Faerman,
Florian Richter, Klaus Schmid

Knowledge Discovery in Databases

SS 2016

Exercise 6: Clustering

Regarding tutorials on 01.06.-03.06.2016.

Exercise 6-1 K-Medoid (PAM)

Consider the following 2-dimensional data set:

$$x_1 = (1,4), x_2 = (1,6), x_3 = (2,6), x_4 = (3,8), x_5 = (4,3), x_6 = (5,2).$$

(a) Perform the first loop of the PAM algorithm ($k = 2$) using the Euclidian distance. Select x_1 and x_3 as initial medoids and compute the resulting medoids and clusters.

(b) How can the clustering result $C_1 = \{x_1, x_5, x_6\}, C_2 = \{x_2, x_3, x_4\}$ be obtained with the PAM algorithm ($k = 2$) using the weighted Manhattan distance

$$d(x,y) = w_1 \cdot |x_1 - y_1| + w_2 \cdot |x_2 - y_2|?$$

Assume that x_1 and x_3 are the initial medoids and give values for the weights w_1 and w_2 for the first and second dimension respectively.

Exercise 6-2 Silhouette-Coefficient and K-Means

Construct a low dimensional data set D together with a clustering $\{C_1, C_2\}$ computed by k-means with the following property:

There exists an object $o \in D$ with a negative silhouette coefficient $s(o) < 0$.

Provide the means of the clusters and compute the silhouette coefficient for the corresponding point o.

Hint: It is possible to find such an example with 5 data points.

Exercise 6-3 Implementation of EM

Implement the EM algorithm and run it on the datasets introduced in exercise 5-3. What do you observe in comparison to your results with k-Means?