Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Thomas Seidl
Max Berrendorf, Julian Busch

Knowledge Discovery and Data Mining I

WS 2018/19

Exercise 5: Sequential Patterns, k-Means, Silhouette Coefficient

Exercise 5-1 Sequential Pattern Mining

Let D be a database that contains the following five sequences.

SID	Sequence
1	ABBA
2	BBACA
3	CBAA
4	ACA
5	BAAB

In addition let *min_sup* = *40%*, i.e. there need to be 2 sequences supporting a pattern.

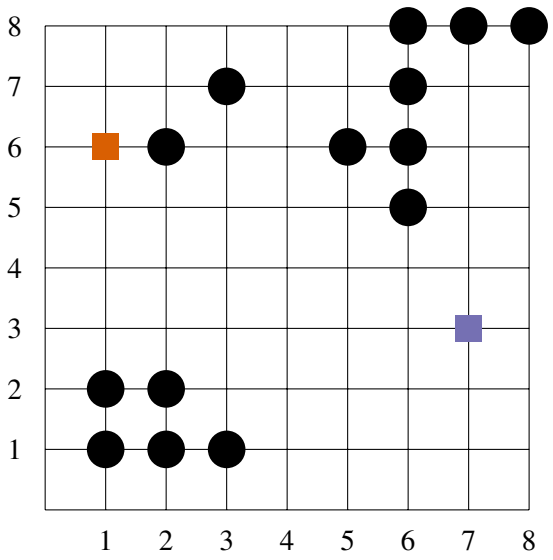
(a) Find all frequent sequence patterns using the *PrefixSpan* algorithm.

(b) Which patterns are maximal? Which are closed?

(c) Considering consecutive patterns, how can they be mined using the framework of PrefixSpan?

Exercise 5-2 k-Means

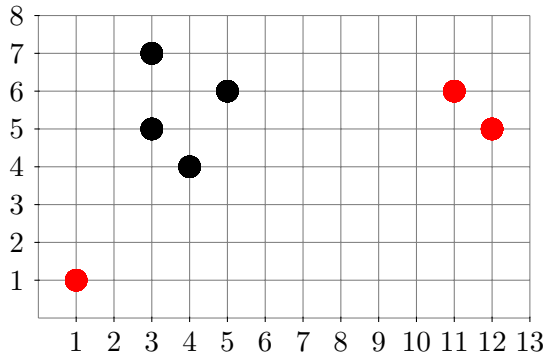
Given the following data set with 14 objects in $\mathbb{R}^2$ (the black dots):



Compute a partitioning into $k = 2$ clusters using the k-means algorithm. As initial representatives use the red and violet square. Start with computing the initial assignment. Explain and draw the assignments as well as the updated centroids after each step.

Exercise 5-3 k-Means

Given the following data set with 7 objects in $\mathbb{R}^2$ represented by the black and red dots:



In the following, we would like to compute complete partitionings of the dataset into $k = 3$ clusters using the k-means algorithm.

Let the initial cluster centers be given by the points marked in red. Carry out the k-Means algorithm as presented in the lecture. Which problem arises?

Exercise 5-4 Silhouette-Coefficient and K-Means

Construct a two-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 4 data points.