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.

Start by constructing the project database for the empty prefix and count the support of 1-sequences.

SID	$D_\emptyset$ Sequence
1	ABBA
2	BBACA
3	CBAA
4	ACA
5	BAAB
	A(5)B(4)C(3)

Hence, all 1-sequences are frequent and none of those can be pruned (i.e. A, B, C are frequent). Next, create projected databases for all remaining items.

SID	D_A	D_B	D_C
1	BBA	BA	-
2	CA	BACA	A
3	A	AA	BAA
4	CA	-	A
5	AB	AAB	-
	A(5)B(2)C(2)	A(4)B(3)~~C~~(1)	A(3)~~B~~(1)~~C~~(0)

These yield the following frequent 2-sequences: AA, AB, AC, BA, BB, CA. Continue by constructing the projected databases for the 3-sequences.

SID	D_{AA}	D_{AB}	D_{AC}	D_{BA}	D_{BB}	D_{CA}
1	-	BA	-	-	A	-
2	-	-	A	A	AA	-
3	-	-	-	A	-	A
4	-	-	A	-	-	-
5	B	-	-	AB	-	-
	~~A~~(0)~~B~~(1)~~C~~(0)	~~A~~(1)~~B~~(1)~~C~~(0)	A(2)~~B~~(0)~~C~~(0)	A(3)~~B~~(1)~~C~~(0)	A(2)~~B~~(0)~~C~~(0)	~~A~~(1)~~B~~(0)~~C~~(0)

We can see that the frequent 3-sequences are ACA, BAA, BBA. Finally, the projections for the 4-sequences are given by

SID	D_{ACA}	D_{BAA}	D_{BBA}
1	-	-	-
2	-	-	A
3	-	-	-
4	-	-	-
5	-	B	-
	~~A~~(0)~~B~~(0)~~C~~(0)	~~A~~(0)~~B~~(1)~~C~~(0)	~~A~~(1)~~B~~(0)~~C~~(0)

In total, the frequent patterns are:

k	Pattern	Absolute Support	Closed	Maximal
0	-	5		
1	A	5		
	B	4		
	C	3		
2	AA	5	✓	
	AB	2	✓	✓
	AC	2		
	BA	4	✓	
	BB	3	✓	
	CA	3	✓	
3	ACA	2	✓	✓
	BAA	3	✓	✓
	BBA	2	✓	✓

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

c.f. (a)

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

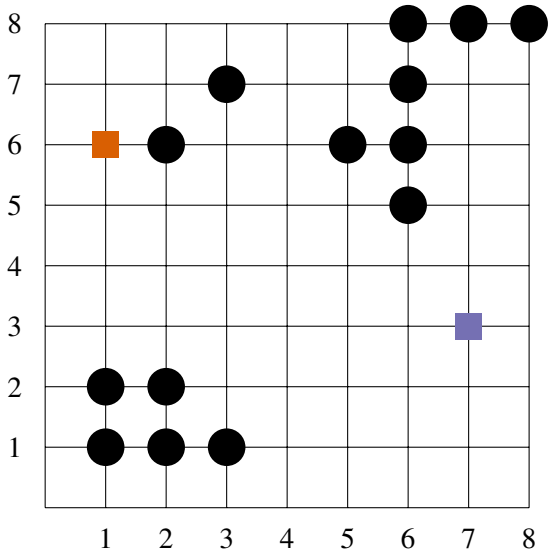
- Large frequent patterns can start later. Consider all starting points when projecting in the first step:

$$ABAAABCD \implies \{AAABCD, CD\} \in D_{AB}$$

- Do not skip objects. Next one is important for later projections!
- Pruning strategy has to be changed

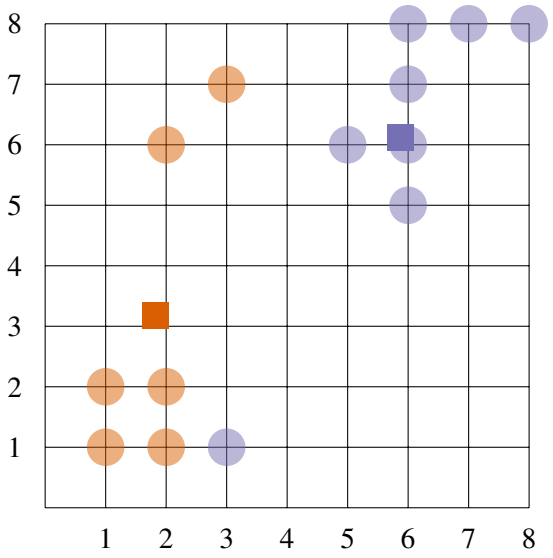
Exercise 5-2 k**-Means**

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

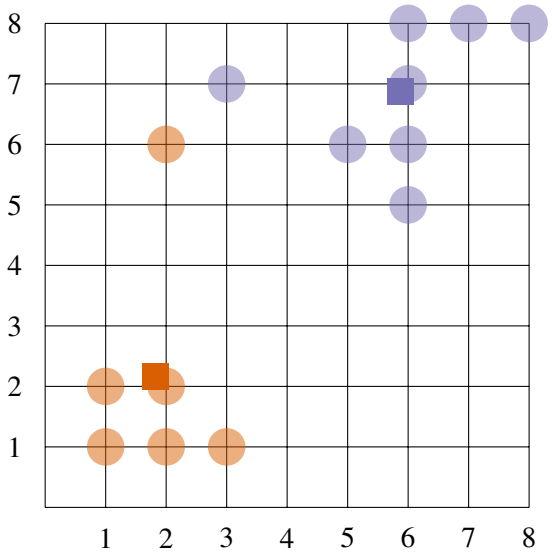


Compute a partitioning into $k = 2$ clusters using the k-means algorithm. As initial representatives use the red and violet square. Start with computing the initial assignment. Explain and draw the assignments as well as the updated centroids after each step.

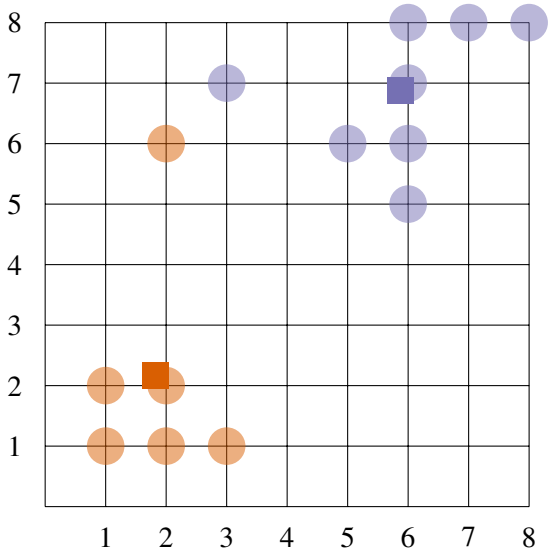
The initial assignment and the updated means are given by



Updating the assignments and subsequently the means yields:

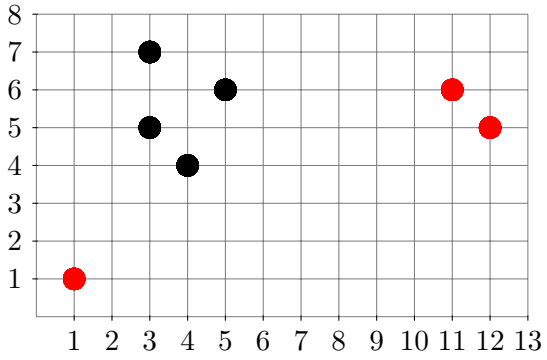


In the next iteration, the assignment does not change anymore. Hence, the means also stay the same and the final result is given by:



Exercise 5-3 k**-Means**

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

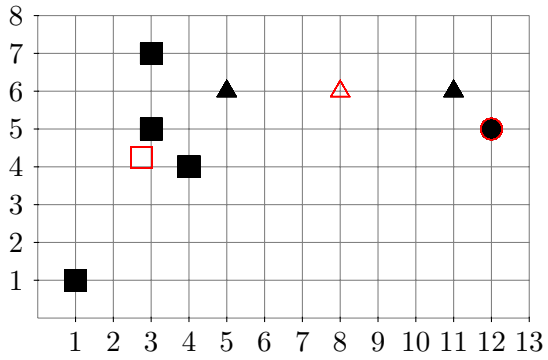


In the following, we would like to compute complete partitionings of the dataset into $k = 3$ clusters using the k-means algorithm.

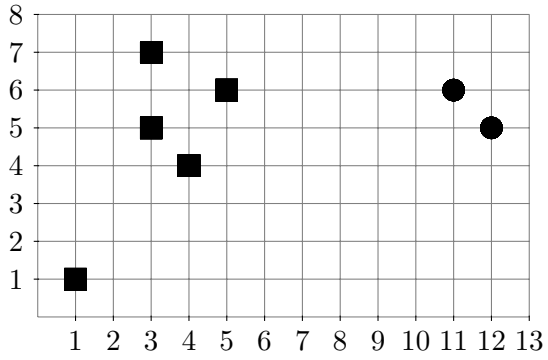
Let the initial cluster centers be given by the points marked in red. Carry out the k-Means algorithm as presented in the lecture. Which problem arises?

One of the clusters becomes empty!

First round of assignments, new centers:



Second round of assignments:



Cluster "triangle" is empty – what should be the new cluster center??

Possible workarounds for such a degenerate case would be to simply restart the algorithm with a different initialization, to remove the empty cluster from consideration and continue with $k-1$ clusters, or to introduce a new cluster center somewhere far away from the existing ones. Empty clusters usually occur as a consequence of bad initialization. Sensible initialization and running the algorithm for multiple iterations are in general important for the success of k-means, not only to prevent empty clusters.

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:

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

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

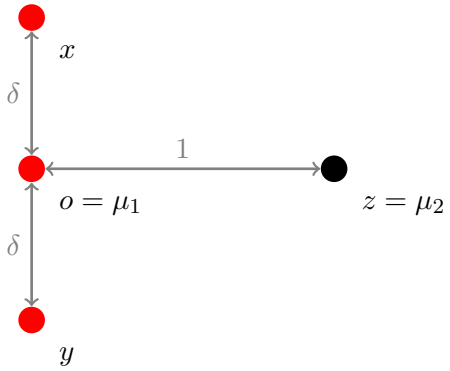
Hint: It is possible to find such an example with 4 data points.

Our example is based on the following ideas:

- All objects need to be closer to the centroid of their own cluster than to centroids of other clusters. Otherwise, k-means would not have terminated.
- In two dimensions, we can symmetrically expand a cluster in one dimension to make the point distances

within the cluster arbitrarily large, but without changing the centroid of the cluster or getting too close to a different cluster

Consider the 4-point dataset $X = \{o, x, y, z\}$ illustrated below with initial centroids $\mu_1 = o$ and $\mu_2 = z$. Then k-means would produce the clusters $C_1 = \{o, x, y\}$ (red) and $C_2 = \{z\}$ (black), irrespective of the value of δ:



The silhouette of o can be computed as follows:

$$a(o) = \frac{1}{|C_1|} \sum_{p \in C_1} d(o, p) = \frac{2\delta}{3}$$

$$b(o) = \frac{1}{|C_2|} \sum_{p \in C_2} d(o, p) = 1$$

$$s(o) = \frac{b(o) - a(o)}{\max\{a(o), b(o)\}} = \frac{1 - \frac{2\delta}{3}}{\frac{2\delta}{3}} = \frac{3}{2\delta} - 1 < 0$$

if we choose δ large enough, such that $a(o) > b(o)$. In particular,

$$\lim_{\delta \to \infty} s(o) = -1,$$

i.e., we can make the silhouette of o arbitrarily bad.

Note that a smaller example is not possible. We need at least 3 data points, since for 1-element clusters, the silhouette is defined to be zero. For 3 data points, the clusters are given w.l.o.g. as $C_1 = \{x_1\}$ and $C_2 = \{x_2, x_3\}$. By definition, $s(x_1) = 0$. For the other cluster, we get

$$a(x_2) = \frac{1}{2} d(x_2, x_3)$$

$$b(x_2) = d(x_1, x_2)$$

$$s(x_2) = \frac{d(x_1, x_2) - \frac{1}{2} d(x_2, x_3)}{\max\{d(x_1, x_2), \frac{1}{2} d(x_2, x_3)\}}$$

For the silhouette to be negative, we need $a(x_2) > b(x_2)$

$$a(x_2) > b(x_2) \Leftrightarrow \frac{1}{2} d(x_2, x_3) > d(x_1, x_2)$$

$$\Leftrightarrow d(x_2, x_3) > 2d(x_1, x_2)$$

On the other hand, termination of k-means requires

$$d(x_2, \mu_2) \leq d(x_2, \mu_1) \Leftrightarrow \frac{1}{2} d(x_2, x_3) \leq d(x_2, x_1)$$

$$\Leftrightarrow d(x_2, x_3) \leq 2d(x_1, x_2)$$

A contradiction. Thus, the silhouette of x_2 cannot be negative.