Prof. Dr. Andreas Hotho, Florian Buckermann
Chair of Computer Science X - Data Science

02.06.2022

3. Exercise „Data Mining“

Summer term 2022

1 Clustering - Basics

1. Name at least two distance measures each for numerical and categorical values!

   - numerical:
     - euclidean metric: $d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} \left(a_r(x_i) - a_r(x_j)\right)^2}$
     - manhattan metric: $d(x_i, x_j) = \sum_{r=1}^{n} |a_r(x_i) - a_r(x_j)|$
   - categorical:
     - Hamming: $dist(x, y) = \sum_{i=1}^{d} \delta(x_i, y_i)$ mit $\delta(x_i, y_i) = 0$ gdw. $x_i = y_i$
     - Jaccard distance: $dist(x, y) = 1 - \frac{|x \cap y|}{|x \cup y|}$

2. Describe the k-means clustering method in your own words.

   a) Choose k random samples. These are the initial centroids
   b) Assign each sample to the nearest centroid, i.e. insert the sample into the corresponding cluster
   c) Calculate a new centroid for each cluster
   d) Repeat step 2-4 until the centroids converge

3. What is the difference between k-means and k-medoids? Which algorithm has a lower run time?

   In every step of k-means the actual centroid of a cluster is updated. In every step of k-medoids it is searched for a new medoid that improves the current clustering. Thus, the k-medoids algorithm has a longer run time.

4. Is it possible to perform the standard k-means or k-medoids algorithms for categorical data? Give a reason for your answer!

- k-means: No. The standard k-means calculates the centroid for a cluster in an euclidean space. Such a centroid can not be calculated for categorical data (e.g. $(blue + yellow)/2 = red$?)
- k-medoids: Yes. k-medoids only needs the a defined distance between two datapoints. There are suitable distance functions for categorical data.

2 Clustering - k-means

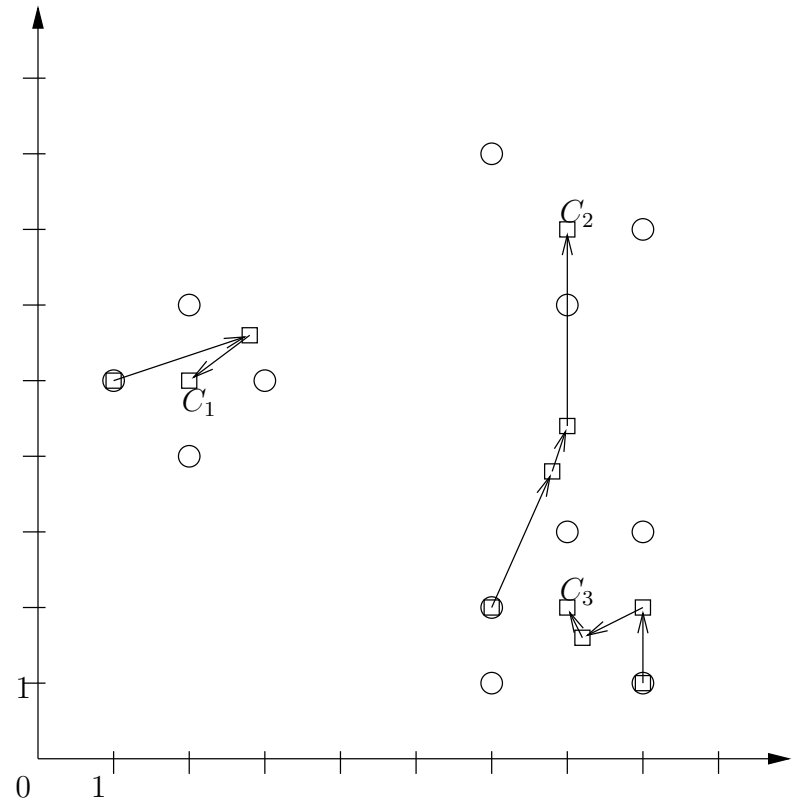
1. The following dataset is given:

x	1	6	8	3	2	2	6	6	7	7	8	8
y	5	2	1	5	4	6	1	8	3	6	3	7

Determine a clustering with the k-means method. Use $k = 3$! Use the first three datapoints as intitial centroids and use the L_2 metric as a distance measure. Update the centroids **after** a full iteration (Lloyd's method)!

Outline the movements of the centroids visually!

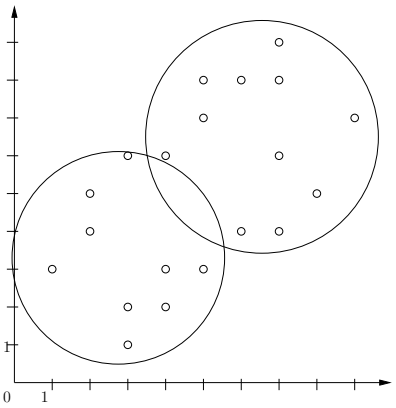
It.	C_1	$K(C_1)$	C_2	$K(C_2)$	C_3	$K(C_3)$
1	$(1,5)$	$\{(1,5),(3,5),(2,4),(2,6),(6,8)\}$	$(6,2)$	$\{(6,2),(6,1),(7,3),(7,6),(8,7)\}$	$(8,1)$	$\{(8,1),(8,3)\}$
2	$(\frac{14}{5},\frac{28}{5})$	$\{(1,5),(3,5),(2,4),(2,6),(6,8)\}$	$(\frac{34}{5},\frac{19}{5})$	$\{(6,2),(7,3),(7,6),(8,7)\}$	$(8,2)$	$\{(6,1),(8,1),(8,3)\}$
3	$(\frac{14}{5},\frac{28}{5})$	$\{(1,5),(3,5),(2,4),(2,6)\}$	$(\frac{14}{2},\frac{9}{2})$	$\{(6,8),(7,6),(8,7)\}$	$(\frac{22}{3},\frac{5}{3})$	$\{(6,1),(6,2),(7,3),(8,1),(8,3)\}$
4	$(2,5)$	$\{(1,5),(3,5),(2,4),(2,6)\}$	$(7,7)$	$\{(6,8),(7,6),(8,7)\}$	$(7,2)$	$\{(6,1),(6,2),(7,3),(8,1),(8,3)\}$



2. Analyse the following two-dimensional labeled dataset without the class information!

x	3	3	4	4	5	6	7	7	8	9	1	2	2	3	4	5	5	6	7	7
y	1	2	2	3	3	4	4	6	5	7	3	4	5	6	6	7	8	8	8	9
Class	a	a	a	a	a	a	a	a	a	a	b	b	b	b	b	b	b	b	b	b

Which challenges arise when the k-means algorithm is used with $k = 2$ and the L_2 distance measure?
Hint: Think about the desired outcome! What is the actual outcome of the algorithm? You do not have to actually calculate the algorithm. A qualitative description is sufficient.



The algorithm produces clusters as shown above. The found cluster do not correspond with the predefined labels. The longish form of the desired clusters is an instable solution for the k-means method. Points from one class may be closer to a centroid that lies within another class. Small changes to the centroid result in a translation of the centroid of the upper cluster.

3. Determine a clustering with the k-means method. Use $k = 2$! Use the first two datapoints as intitial centroids and update the centroids **after** a full iteration. Instead of using the L_2 metric, use the cosine distance in this case:

$$cosdist(x, y) = 1 - \frac{\langle x, y \rangle}{|x| \cdot |y|} = 1 - \frac{\sum_{i=1}^{d} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{d} x_i^2 \cdot \sum_{i=1}^{d} y_i^2}} \tag{1}$$

#	Centroids	Members
1	$(3, 1)$	p_1
	$(3, 2)$	$p_2, p_3, p_4, p_5, p_6, p_7, p_8, p_9, p_{10}, p_{11}, p_{12}, p_{13}, p_{14}, p_{15}, p_{16}, p_{17}, p_{18}, p_{19}, p_{20}$
2	(3, 1)	p_1, p_3, p_5, p_7, p_9
	(5, 5.26)	$p_2, p_4, p_6, p_8, p_{10}, p_{11}, p_{12}, p_{13}, p_{14}, p_{15}, p_{16}, p_{17}, p_{18}, p_{19}, p_{20}$
3	(5.4, 3)	$p_1, p_2, p_3, p_4, p_5, p_6, p_7, p_9, p_{10}$
	(4.7, 5.7)	$p_8, p_{11}, p_{12}, p_{13}, p_{14}, p_{15}, p_{16}, p_{17}, p_{18}, p_{19}, p_{20}$
4	(5.6, 3.7)	$p_1, p_2, p_3, p_4, p_5, p_6, p_7, p_8, p_9, p_{10}$
	(4.2, 6.4)	$p_{11}, p_{12}, p_{13}, p_{14}, p_{15}, p_{16}, p_{17}, p_{18}, p_{19}, p_{20}$
5	(5.6, 3.7)	$p_1, p_2, p_3, p_4, p_5, p_6, p_7, p_8, p_9, p_{10}$
	(4.2, 6.4)	$p_{11}, p_{12}, p_{13}, p_{14}, p_{15}, p_{16}, p_{17}, p_{18}, p_{19}, p_{20}$

3 Clustering - k-medoids

The following categorical dataset is given:

$$x_1 = \begin{pmatrix} \text{rot} \\ \text{zwei} \\ \text{sonnig} \\ \text{flüssig} \end{pmatrix}, x_2 = \begin{pmatrix} \text{grün} \\ \text{zwei} \\ \text{bewölkt} \\ \text{fest} \end{pmatrix}, x_3 = \begin{pmatrix} \text{rot} \\ \text{drei} \\ \text{sonnig} \\ \text{gas} \end{pmatrix},$$

$$x_4 = \begin{pmatrix} \text{grün} \\ \text{drei} \\ \text{regnerisch} \\ \text{gas} \end{pmatrix} \text{ und } x_5 = \begin{pmatrix} \text{gelb} \\ \text{zwei} \\ \text{bewölkt} \\ \text{gas} \end{pmatrix}$$

Perform k-medois for the points $x_1, \ldots, x_5$! Use x_1 and x_2 as initial medoids and use the hamming distance!

Initial Clustering $M = \{x_1, x_2\}$, $NM = \{x_3, x_4, x_5\}$:
Costs: 7

1st iteration

- Swap x_1 and x_3 ($M = \{x_2, x_3\}$, $NM = \{x_4, x_5, x_1\}$) Costs: 6
- Swap x_1 and x_4 ($M = \{x_2, x_4\}$, $NM = \{x_3, x_5, x_1\}$) Costs: 7
- Swap x_1 and x_5 ($M = \{x_2, x_5\}$, $NM = \{x_3, x_4, x_1\}$) Costs: 9
- Swap x_2 and x_3 ($M = \{x_1, x_3\}$, $NM = \{x_4, x_5, x_2\}$) Costs: 8
- Swap x_2 and x_4 ($M = \{x_1, x_4\}$, $NM = \{x_3, x_5, x_2\}$) Costs: 8
- Swap x_2 and x_5 ($M = \{x_1, x_5\}$, $NM = \{x_3, x_4, x_2\}$) Costs: 7

Minimum costs in this iteration: 6

Choose $M = \{x_2, x_3\}$, $NM = \{x_4, x_5, x_1\}$

2nd iteration

- No distance between two points is below 2. Thus, achieving costs < 6 is not possible
- Costs do not change $\rightarrow$ Stop. Medoids are x_2 and x_3