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!
2. Describe the k-means clustering method in your own words.
3. What is the difference between k-means and k-medoids? Which algorithm has a lower 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!

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!

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.

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:

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

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!