Information Mining - winter semester 2019

Exercise sheet 9

Exercise 1: Clustering: k-Means

Within information retrieval the vector space model is very important. All documents will be represented as vectors of terms and the parts of the vector represent the weight of the term inside the document.

This model can be used for clustering. Assuming we have the terms T = {`house`, `car`}; the corresponding vector space has 2 dimensions `house` and `car`. We have the following 5 documents, who can be described as vectors $\vec{d}$.

$$\vec{d_1} = \begin{pmatrix} 0,8 \\ 0,9 \end{pmatrix}, \vec{d_2} = \begin{pmatrix} 0,9 \\ 0,65 \end{pmatrix}, \vec{d_3} = \begin{pmatrix} 0,5 \\ 0,5 \end{pmatrix}, \vec{d_4} = \begin{pmatrix} 0,2 \\ 0,25 \end{pmatrix}, \vec{d_5} = \begin{pmatrix} 0,25 \\ 0,1 \end{pmatrix}$$

The document d_1 has the weight 0,8 for `house` and the weight 0,9 for `car`.

(a) Cluster the documents with k-Means-Clustering. Two clusters should be created and as initial seeds d_4 and d_5 should be used. Do just one iteration!

(b) Graphically sketch the document space and the movement of the centroids. What can be seen?

(c) Calculate for the resulting clusters *Purity*. Assume the following manual classification: $C_1 = \{d_1, d_2, d_3\}$ and $C_2 = \{d_4, d_5\}$.

Exercise 2: Hierarchical clustering

What distance functions do you know in combination with hierarchical clustering? Explain each of them.

Exercise 3: k-NN

The data set in Table 1 is given. For this data set we aim to apply K-NN with a normal Euclidian distance function:

$$Euclidiandistance : D(\vec{a}, \vec{b}) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + ... + (a_n - b_n)^2},$$

where $\vec{a}$ and $\vec{b}$ are the two instances in vector space. The a_n and b_n are single values in each respective vector. The aim is to predict the class label (y) given a test instance x.

What is the leave-one-out (1 instance is used as testing the remaining ones for training) cross-validation error of 1-NN and 3-NN classifications on this dataset? Give your answer as the number of misclassifications.

Tabelle 1: Data for K-NN.

x	y
-0.1	-
0.7	+
1.0	+
1.6	-
2.0	+
2.5	+
3.2	-
3.5	-
4.1	+
4.9	+