Assignment 5

Information Retrieval and Text Mining 17/18

2018-01-16; to be submitted 2018-01-25

Roman Klinger, Florian Strohm

- **Deadline:** This assignment will be discussed on 2018-01-30. Submissions are accepted via Ilias until end of day 2018-01-25.

Task 1 (Feature Selection) 4 points

Given the following documents assigned with classes c_1 and c_2:

c_1 I drink coffee

c_1 I drink tee

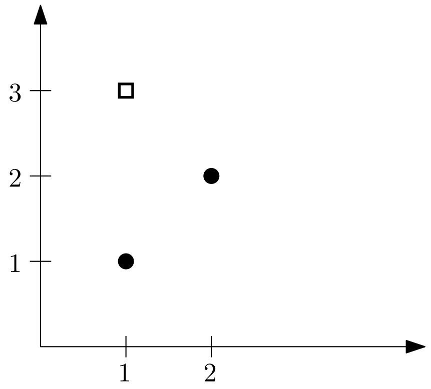
c_2 I take aspirin

c_2 I take paracetamol

The occurrence of which of the words "drink", "take", "coffee", "tee", "aspirin", "paracetamol", "I", helps best to predict if a document belongs to class c_1 or c_2? Please argue based on *mutual information*. (that does not mean that you need to calculate formally, but you can)

Task 2 (Perceptron) 4 points

Given are these instances:



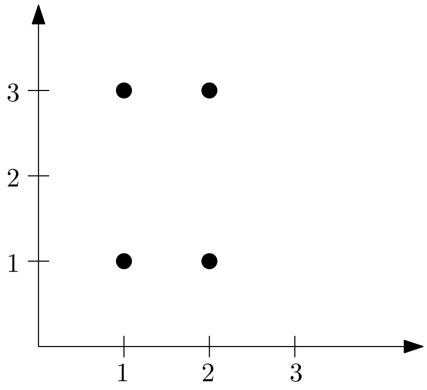
Perceptron learning searches for a solution by iteratively optimizing a weight vector. Interpret $-\theta$ as an additional feature weight in the weight vector which always has the feature value 1. Then, the instances are interpreted as vectors $(1, 1, 1)$, $(1, 2, 2)$ and $(1, 1, 3)$. This is because:

$$\begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix}^T \cdot \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{pmatrix} = \theta \iff \begin{pmatrix} -\theta \\ w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix}^T \cdot \begin{pmatrix} 1 \\ f_1 \\ f_2 \\ \vdots \\ f_n \end{pmatrix} = 0$$

Perform (at least) three iterations of perceptron learning, starting with the vector $(1, 1, 1)$. In each iteration, all instances are processed, that therefore leads to (at least) 9 weight vector updates.

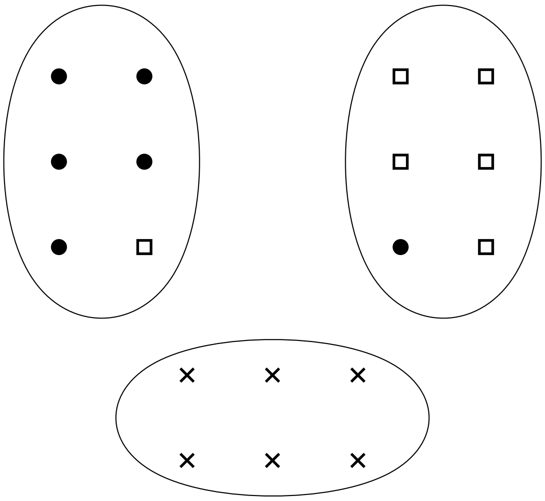
Task 3 (HAC) 3 points

Perform hierarchical agglomerative clustering of these instances with single link and with complete link clustering. Is the result the same? Is there a difference?



Task 4 (Evaluation of Clustering) 2 points

What is the rand index of the following clustering, assuming that cross, circle and box are gold annotations of classes?



Task 5 (PageRank) 7 points

What is the page rank value for two documents d_1 and d_2 in which exactly one link in d_1 points to d_2 and one link points from d_2 to itself?

Programming Task 5 (k means clustering) 10 points

Implement k means clustering and apply it to one of the document collections provided in this class (or a subset).

- Apply it with $k = 2$ and with $k = 20$. Interpret the result!
- Provide the list of documents which is closest to the cluster centers for $k = 20$.

Please provide the source code as before.