Ludwig-Maximilians-Universitaet Muenchen 30.06.2015
Institute for Informatics
Prof. Dr. Volker Tresp
Gregor Jossé
Johannes Niedermayer

Machine Learning and Data Mining

Summer 2015

Exercise Sheet 11

Presentation of Solutions to the Exercise Sheet on the 08.07.2015

Exercise 11-1 Document Distance

Consider four documents from a document dataset, which has been mapped onto an lexicon of size $M = 100$ w.r.t. word frequency $x_{i,j} \in \{1,2,\dots\}$.

Let A denote the lexicon itself, i.e. $\forall j \in \{1,\dots,M\} : x_{A,j} = 1$. Let B be a document containing only the first word of A ($x_{B,1} = 1 \wedge \forall j \in \{2,\dots,M\} : x_{B,j} = 0$). Let C contain the first 50 words of A, and, finally, let D contain the 11th to 60th word twice.

a) Compute the pairwise distance of vectors A, B, C, D, w.r.t. the following distance measures:

$$
\begin{aligned}
dist_{\text{eucl}}(x,y) &= \Big(\sum_{i=1}^{M}(x_j - Y_j)^2\Big)^{1/2} \\
dist_{\text{simple}}(x,y) &= \frac{1}{M}\sum_{i=1}^{M}(1 - I(x_j = y_j)) \\
dist_{\text{simple00}}(x,y) &= \frac{1}{M-F}\sum_{i=1}^{M}(1 - I(x_j = y_j)) \\
dist_{\cos}(x,y) &= 1 - \frac{x^T y}{\|x\|\|y\|} \\
dist_{\text{pearson}}(x,y) &= 1 - \frac{\tilde{x}^T \tilde{y}}{\|\tilde{x}\|\|\tilde{y}\|}
\end{aligned}
$$

where $I(\text{condition})$ is the indicator function which is 1 iff the condition is fulfilled and 0 otherwise, F is the number of components in which both vectors are 0, and $\tilde{x} := x - mean(x)$.

b) How do the distances change if it is also known that the first fifty words are contained in 750 of the total $N = 1000$ documents in the set, while all other words only appear in 5 documents? *Remark:* You know the "term frequency", which measures the absolute frequency of words in a document. When there is additional information about the global frequency of a term (i.e., is it common or rare among all documents), it should also be taken into account. This is often done using the inverse document frequency:

$$\text{idf}_j = \log\left(\frac{N}{n_j}\right)$$

where N is the number of documents and n_j is the number of documents in which the word j occurs. The measures tf and idf are often combined by multiplication: $\text{tf idf} = \text{tf} \cdot \text{idf}$. What does this measure reflect? Use this measure for this exercise.