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

30.06.2015

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 two 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:
$dist_{\text{eucl}}(x,y)$, $dist_{\text{simple}}(x,y)$, $dist_{\text{simple00}}(x,y)$, $dist_{\cos}(x,y)$, $dist_{\text{pearson}}(x,y)$

b) How do the distance 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?