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

02.07.2014

Machine Learning and Data Mining

Summer 2014

Exercise Sheet 8

Presentation of Solutions to the Exercise Sheet on the 03.07.2014

Aufgabe 8-1 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?

Aufgabe 8-2 Example Application for CF+

Consider a dataset D where $d_i = (x_{i,1},\dots,x_{i,M})^T$ for $N = 5$ users of a film database, each of them having evaluated 6 films on a scale from 1 to 5:

	Up	Inception	Iron Sky	Kick-Ass	RED	Paul
1	5	2	5	3	4	5
2	4	5	2	2	1	2
3	2	5	4	5	2	3
4	3	1	1	1	2	1
5	5	1	2	1	3	1

D can be reformulated to a problem of Collaborative Filtering (CF): We define $j \in \{1,M\}$ as the film which is to be predicted for query-user $z^T = d_i$, $i \in \{1,N\}$. That is, column j of D corresponds to output vector$y$. b.w.

Using the formula for CF+ given in the script, the pearson correlation may not be well-defined. In this case, the corresponding training examples are not used for evaluation.

a) Determine the predicted score for the film *RED* for all users. Why is the scale from 1 to 5 not maintained?

b) How would you determine the error of these predictions? Which movie is best predicted according to the model? Which user is most predictable?

c) What scores do you expect for *Up* when randomly input patters? Simulate 10000 random patters for the remaining films and determine the mean value for the assigned scores.