Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Volker Tresp
Gregor Jossé

02.07.2014

Maschinelles Lernen und Data Mining

Sommersemester 2014

Übungsblatt 8

Besprechung des Übungsblattes am 03.07.2014

Aufgabe 8-1 Dokumentdistanzen

Gegeben seien vier Dokumente eines Dokumentendatensatzes, der nach Worthäufigkeiten $x_{i,j} \in \{1,2,\dots\}$ auf ein Lexikon der Größe $M = 100$ abgebildet wurde.

Dokument A ist das Lexikon selber, d.h. $\forall j \in \{1,\dots,M\} : x_{A,j} = 1$; Dokument B besteht nur aus dem ersten Wort des Lexikons ($x_{B,1} = 1 \wedge \forall j \in \{2,\dots,M\} : x_{B,j} = 0$). Dokument C enthält genau die ersten 50 Wörter und Dokument D enthält genau das elfte bis sechzigste Wort zweimal.

a) Berechnen Sie die paarweisen Abstände der Vektorenpaare A, B, C und D mit Hilfe der Distanzmaße $dist_{euklid}(x,y)$, $dist_{simple}(x,y)$, $dist_{simple00}(x,y)$, $dist_{cos}(x,y)$, $dist_{pearson}(x,y)$

b) Wie ändern sich die Distanzen, wenn zusätzlich bekannt ist, dass die Wörter 1 bis 50 in der Datenbank der Größe $N = 1000$ in jeweils 750 Dokumenten auftreten, die übrigen dagegen nur in jeweils 5?

Aufgabe 8-2 Beispielanwendung für CF+

Gegeben sei ein Datensatz D mit $d_i = (x_{i,1},\dots,x_{i,M})^T$ für $N = 5$ Benutzer einer Filmdatenbank mit je $M = 6$ auf einer Skala zwischen 1 und 5 beurteilten Filmen:

	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 kann umformuliert werden zu einem Problem des Collaborative Filterings (CF): Wir deklarieren ein $j \in \{1,M\}$ als den vorherzusagenden Film für den Anfragebenutzer $z^T = d_i$ für $i \in \{1,N\}$. Somit entspricht Spalte j von D dem Zielvektor y.

b.w.

Verwenden Sie für die folgenden Aufgaben die in den Folien angegebene Formel zu CF+. Hier kann es passieren, dass die Pearson-Korrelation nicht definiert ist - in diesem Fall werden die betroffenen Trainingsbeispiele für gewöhnlich nicht verwendet, werden also als nicht-gewertet betrachtet.

a) Bestimmen Sie die vorhergesagten Scores zum Film *RED* für alle User. Warum wird hier die Skala von 1 bis 5 nicht eingehalten?

b) Wie wï¿½rden Sie den Fehler für diese Vorhersagen bewerten? Welcher Film kann Ihrem Vorhersagemodell nach am besten vorhergesagt werden? Welcher Benutzer votiert am “berechenbarsten”?

c) Was für einen Score erwarten Sie für *Up* bei zufälliger Mustereingabe? Simulieren Sie 10000 zufällige Muster für die verbleibenden Filme und bestimmen Sie den Mittelwert über die zugewiesenen Scores.