Ludwig-Maximilians-Universität München
Institut für Informatik
Dr. Volker Tresp
Franz Graf, Marisa Thoma

06.07.2010

Maschinelles Lernen und Data Mining

Sommersemester 2009

Übungsblatt 8

Besprechung des Übungsblattes am 15.07.2010

Aufgabe 8-1 Distanzmaße
schriftlich bearbeiten

Gegeben seien folgende Vektoren:
$A = (2, 3)$
$B = (1, 0)$

$$C = (x_1, ..., x_{100}), x_i = \begin{cases} 1 & i = 1 \\ 0 & sonst \end{cases}$$

$$D = (x_1, ..., x_{100}), x_i = \begin{cases} 2 & i = 2 \\ 0 & sonst \end{cases}$$

$$E = (x_1, ..., x_{100}), x_i = \begin{cases} 1 & i \in [1,10] \\ 0 & sonst \end{cases}$$

$$F = (x_1, ..., x_{100}), x_i = \begin{cases} 2 & i \in [6,15] \\ 0 & sonst \end{cases}$$

a) Berechnen Sie die Abstände der Vektorenpaare A, B sowie C, D und E, F 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) Welches Problem kann beim Pearson-Distanzmaß auftreten?

Aufgabe 8-2 Vergleich: Nächster Nachbar Schätzer und das Perceptron
schriftlich bearbeiten

Vergleichen Sie den Nächste Nachbar Schätzer mit dem Perceptron. Wie lassen sich die beiden Klassifikatoren visualisieren?

Aufgabe 8-3 Kernglätter

a) Skizzieren Sie den Verlauf der des Kerndichteschätzers im Bereich $x = [-2{,}5; 2{,}5]$ mit $z = 0$:
Gauss Kern: $K_\lambda(z,x_i) = exp\left(-\frac{|z-x_i|^2}{2\lambda^2}\right)$, $\lambda = 0{,}17$
Epanechnikov Kern:

$$K(z,x_i) = \begin{cases} 3/4 \cdot (1 - (z - x)^2) & , |z - x| < 0 \\ 0 & sonst \end{cases}$$

b) Glätten Sie die gegebenen Punkte (1;5), (2;8), (3;6), (4;5), (5;4), (6;5,5) mit den oben angegebenen Kernen. Berechnen Sie dazu die Werte an den Stellen $x = \{0{,}5; 1; \ldots; 6{,}5\}$ und erläutern sie kurz die Vor-und Nachteile der Kerne.

Aufgabe 8-4 Beispielanwendung für CF+
schriftlich bearbeiten

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

user	300	Juno	Crank2	Milk	Indy4	Wall-E
1	5	1	5	1	4	5
2	5	2	5	1	5	4
3	3	4	2	3	3	2
4	2	5	1	4	3	3
5	1	3	1	2	1	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.

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 *Crank2* für alle User. Warum wird hier die Skala von 1 bis 5 nicht eingehalten?

b) Wie würden 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 *Crank2* bei zufälliger Mustereingabe? Simulieren Sie 10000 zufällige Muster für die verbleibenden Filme und bestimmen Sie den Mittelwert über die zugewiesenen Scores.

d) Was ist der mittlere, auf zufälligen Benutzern vorhergesagte Score über alle Filme? Warum?