# 7. Übung zur Vorlesung "Internet-Suchmaschinen" im Wintersemester 2007/2008

Dr. Andreas Hotho, Prof. Dr. Gerd Stumme, Dipl.-Wi.Inf. Beate Krause

07. Februar 2008

## 1 Recommender-Systeme

1. Erklären Sie mit eigenen Worten, was der Pearson-Korrelationskoeffizient aussagt!
2. Betrachten Sie die Bewertungen von Filmen durch die Benutzer Alice (A), Bob (B) und Charlie (C) gemäß der folgenden Tabelle:

| Film | Alice | Bob | Charlie |
|---|---|---|---|
| Titanic | 7 | 9 | 5 |
| High Fidelity | 5 | 7 | 5 |
| American Beauty | 5 | 7 | 5 |
| Jarhead | 4 | 6 | 4 |
| Life of Brian | 4 | 6 | 4 |
| Three Kings | 5 | | |
| A Fish Called Wanda | | | 4 |

   Schätzen Sie – gemäß der vorigen Antwort – ab, wie die Größe der Korrelationskoeffizienten $c_{A,B}, c_{B,C}, c_{A,C}$ relativ zueinander aussehen wird!
3. Berechnen Sie die Korrelationskoeffizienten $c_{A,B}, c_{B,C}, c_{A,C}$!
4. Sagen Sie eine Bewertung der Filme "Three Kings" und "A Fish Called Wanda" durch den Anwender Bob voraus! Ziehen sie dazu jeweils den anderen Anwender heran, der den fraglichen Film bewertet hat.

## 2 Klassifikation

1. Geben Sie jeweils eine "gute" Klassifikationsfunktion und eine Funktion, die an Overfitting leidet, an für folgende Klassifikationsaufgaben. Gehen Sie davon aus, das für das Lernen der Funktion eine gewisse Menge Trainingsdaten zur Verfügung steht.

a) Ist ein gegebener Tag X ein Regentag?

b) Ist dieser Zug ein ICE?

c) Kann ich jetzt an dieser Ampel losfahren?

d) Ist diese Mail Spam?

2. Wie könnten (in einem kleinen Beispiel, etwa 10 Dimensionen) Prototyp-Vektoren aussehen für die Klassifikationsaufgabe "Behandelt diese Nachrichtenmeldung das Thema Sport/Basketball/Politik?"

3. Konstruieren Sie einen Fall, wo Nearest-Neighbor bei der Aufgabe, einen Artikel in "Sport" und "Politik" zu klassifzieren, versagt!

4. Sie haben einen Trainingsdatensatz mit 20 Dokumenten. 8 von diesen klassifizieren Sie als Spam ($c = 1$), 12 als Nicht-Spam ($c = 0$). Folgende Wahrscheinlichkeiten konnten für die ersten Wörter in allen Dokumenten erstellt werden. Sie möchten mit Hilfe des Naive Bayes Klassifizierers entscheiden, ob das folgende Dokument Spam oder Nichtspam darstellt: o(web 4, marketing 10, seo 3). Der Naive Bayes Klassifizier ist definiert durch:

$$\arg\max_{c_i \in C} P(c_i|o) = (P(o) * P(o|c_i))/P(o) = P(o)P(o|c_i)$$

Für diese Terme wurden die folgenden bedingten Wahrscheinlichkeiten im Trainingsdatensatz errechnet:

$P(web|c = 1) = 0.2$
$P(web|c = 0) = 0.8$
$P(marketing|c = 1) = 0.5$
$P(marketing|c = 0) = 0.5$
$P(seo|c = 1) = 0.9$
$P(seo|c = 0) = 0.1$

# 3 Clustern

1. Formulieren Sie den k-Means-Algorithmus in Pseudocode!

2. Beschreiben Sie kurz das prinzipielle Vorgehen beim divisiven Clustern.

3. Erläutern Sie warum diese Verfahren typischerweise schlechter arbeiten als die agglomerativen.

# 4 Informationsextraktion

Skizzieren Sie die Informationsextraktions-Aufgaben, die beim Bau eines Webangebots folgender Machart anstehen: auf `http://www.rottentomatoes.com` gibt es zu Filmen eine aggregierte Sicht der Bewertungen verschiedener Kritiker, insbesondere eine numerische Angabe der durchschnittlichen Bewertungen, sowie Textschnipsel aus den Reviews, Spielpläne nahegelegener Kinos und Daten zum Film selbst (Startdatum, Regisseur, usw.).