Dr. Andreas Hotho, Robert Jäschke
Fachgebiet Wissensverarbeitung

20.6.2006

8. Übung „Knowledge Discovery“

Sommersemester 2006

1 Naiver Bayes-Klassifikator

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

1. Klassifizieren Sie mit Hilfe des naiven Bayes-Klassifikators den Datensatz *D15 = (Outlook=Overcast, Temperature=Cool, Humidity=High, Wind=Strong)*.

2. Klassifizieren Sie mit Hilfe des naiven Bayes-Klassifikators den Datensatz *D16 = (Temperature=Cool, Humidity=High, Wind=Strong)* und vergleichen Sie ihr Ergebnis mit dem aus Teilaufgabe a).

3. Diskutieren Sie die praktischen Auswirkungen z.B. auf Spamfilter.

4. Was bewirkt eine Änderung der Berechnung der geschätzten Wahrscheinlichkeit von $P(a|c) = \frac{n_c}{n}$ zu $P(a|c) = \frac{n_c+mp}{n+m}$ (wobei n die Gesamtzahl der Trainingsbeispiele mit Klassifikation c und n_c die Anzahl der Trainingsbeispiele mit Klassifikation c und dem Attributwert a darstellt; m ist eine Konstante und p der geschätzte Wert für $P(a|c)$ – ist dieser unbekannt, wird Gleichverteilung der Attributwerte angenommen)? Vergleichen Sie diesen Ansatz unter dem Gesichtspunkt der in c) diskutierten Probleme.

2 k-nächste Nachbarn Klassifikator

1. Geben Sie die prinzipiellen Schritte eines kNN-Verfahrens an und nennen Sie mindestens je ein Abstandsmaß für numerische und kategorische Werte.

2. Diskutieren Sie die Vor- und Nachteile des Verfahrens.

3. Berechnen Sie für $k = 4$ den Abstand zum Beispiel (sunny, cool, high, true) für den Datensatz vom letzten Übungsblatt.