Prof. Dr. Katharina Morik,
Prof. Dr. Claus Weihs
Dipl.-Inform. Christian Bockermann,
Dipl.-Stat. Julia Schiffner

Dortmund, 10.05.11
Abgabe: bis Di, 17.05., 10.00 Uhr an
schiffner@statistik.tu-dortmund.de

Übungen zur Vorlesung

Wissensentdeckung in Datenbanken

Sommersemester 2011

Blatt 5

Aufgabe 5.1 (6 Punkte)
Diese Aufgabe ist die Fortsetzung von Aufgabe 4.2: Gegeben sei wieder ein Klassifikationsproblem mit zwei Klassen. Nehmen Sie an, dass die Daten aus zwei univariaten Normalverteilungen mit $\mu_0 = 0$, $\mu_1 = 2$ und $\sigma_0 = \sigma_1 = 2$ stammen. Die a priori Wahrscheinlichkeiten π_0 und π_1 der beiden Klassen seien gleich.

Nehmen Sie nun zusätzlich an, dass die Kosten für eine Fehlklassifikation nicht mehr identisch sind, sondern dass $c(0,1) = 1$ und $c(1,0) = 10$ (und $c(0,0) = c(1,1) = 0$).

a) Berechnen Sie die kostenoptimale Klassifikationsregel.

b) Was ändert sich gegenüber der optimalen Regel für identische Kosten aus Aufgabe 4.2 c) 'Wähle Klasse 0, wenn $x < 1$'? Warum?

c) Berechnen Sie die Fehlklassifikationswahrscheinlichkeiten der beiden Regeln, d.h. der Regel aus a) und der Regel aus 4.2 c). Welche ist größer und warum?

d) Wie groß sind die erwarteten Fehlklassifikationskosten C_0 und C_1 für die Vorhersage von Klasse 0 bzw. 1 (siehe `Statistik_Teil1.pdf`, S. 147)? Wie groß sind die erwarteten Fehlklassifikationskosten insgesamt?

Nehmen Sie nun an, dass die Varianzen σ_0^2 und σ_1^2 nicht mehr gleich sind, sondern dass $\sigma_0 = 2$ und $\sigma_1 = 4$. Die Kosten seien identisch, d.h. $c(i,j) = I(j \neq i)$ (mit I der Indikatorfunktion und $i, j \in \{0,1\}$).

e) Wie lautet die optimale Regel?

Aufgabe 5.2 (4 Punkte)
Auf der Homepage stehen die Datensätze `spam.train.txt` und `spam.test.txt` sowie die Datei `spam.info.txt`, die einige Informationen zu den Daten enthält.

a) Trainieren Sie das naive Bayes Verfahren (R-Funktion `naiveBayes` aus dem Paket `e1071`) auf dem Trainingsdatensatz. Nehmen Sie an, dass die einzelnen Variablen gegeben die Klassen normalverteilt sind. Sagen Sie die Klassenzugehörigkeiten der Beobachtungen im Testdatensatz vorher (R-Funktion `predict`) und berechnen Sie die relative Häufigkeit einer Fehlklassifikation.

b) Verwenden Sie nun eine lineare Diskriminanzanalyse (R-Funktion `lda` aus dem Paket `MASS`). Gehen Sie genauso vor wie in a). Für welches Verfahren ist die Fehlerhäufigkeit kleiner?