Prof. Dr. Katharina Morik,
Prof. Dr. Claus Weihs
Marco Stolpe, Julia Schiffner

Dortmund, 12.05.09
Abgabe: bis Di, 19.05., 10.00 Uhr an
schiffner@statistik.tu-dortmund.de

Übungen zur Vorlesung
Wissensentdeckung in Datenbanken
Sommersemester 2009

Blatt 5

Aufgabe 5.1 (6 Punkte)
Bearbeiten Sie die Aufgabenteile b) und c) bitte in R und hängen den zugehörigen Quellcode an.

a) Weisen Sie nach, dass sich die trennende Hyperebene bei der Linearen Diskriminanzanalyse im Fall gleicher a priori Wahrscheinlichkeiten in Form einer Ebenengleichung

$$H = \left\{ x \in \mathbb{R}^p \;\middle|\; \left(\Sigma^{-1}(\mu_1 - \mu_2)\right)' \left(x - 0.5(\mu_1 + \mu_2)\right) = 0 \right\}$$

mit Stützvektor $0.5(\mu_1 + \mu_2)$ und Normalenvektor $\Sigma^{-1}(\mu_1 - \mu_2)$ darstellen lässt. Zeigen Sie, dass die Hyperebene genau dann senkrecht auf $\mu_1 - \mu_2$, der Verbindungslinie zwischen den beiden Klassenmitteln, steht, wenn $\mu_1 - \mu_2$ ein Eigenvektor von Σ^{-1} ist.

b) Programmieren Sie eigenständig eine *Lineare Diskriminanzanalyse* nach Fisher in R für die Klassifikation des Banknoten-Datensatzes vom letzten Blatt und wenden diese darauf an.

Wie viele Diskriminanzkomponenten existieren für dieses Klassifikationsproblem?

Stellen Sie die Daten auf der Diskriminanzachse grafisch dar und beschreiben sie kurz das Ergebnis.

c) Mit Hilfe des im Skript beschriebenen F-Tests (Folie 50 in der Datei `Klassifikation2.pdf`) lässt sich beurteilen, welche erklärenden Variablen Einfluss auf die Trennung der Klassen haben. Machen Sie den ersten Schritt einer Rückwärtsselektion auf den Banknoten-Daten, d. h. identifizieren Sie zunächst diejenige Variable, die D_k am wenigsten verkleinert und testen Sie anschließend zum Signifikanzniveau 5%, ob diese Variable tatsächlich irrelvant ist.

Hinweise zu b) und c): Einige nützliche Befehle in R sind

- die Funktion `solve` zur Invertierung von Matrizen,
- der Operator `%*%`, um Matrizen miteinander zu multiplizieren,
- die Funktion `t` zum Transponieren von Matrizen sowie
- der Funktion `qf` zur Berechnung von Quantilen der F-Verteilung.

Aufgabe 5.2 (4 Punkte)
Auf der Homepage steht der Datensatz `spam.txt` sowie die Datei `info.txt`, die einige Informationen zu den Daten enthält. Bearbeiten Sie diese Aufgabe bitte in R und hängen den zugehörigen Quellcode an.

a) Wenden Sie *lineare* sowie *quadratische Diskriminanzanalyse* (R-Funktionen `lda` und `qda` im Paket `MASS`) auf den Datensatz an. Berechnen Sie die Wiedereinsetzungsfehlerrate und die 10-fach kreuzvalidierte Fehlerrate und vergleichen Sie die Ergebnisse.

b) Eine Alternative zur linearen bzw. quadratischen Diskriminanzanalyse stellt die *regularisierte Diskriminanzanalyse* dar (R-Funktion `rda` im Paket `klaR`). Berechnen Sie wie im vorigen Aufgabenteil die Wiedereinsetzungsfehlerrate und die 10-fach kreuzvalidierte Fehlerrate. Varriieren sie dabei die Parameter $(\delta, \lambda)' \in \{0, 0.25, 0.5, 0.75, 1\}^2$ und tabellieren Sie die Fehlerraten.

Hinweise: Für die Erzeugung eines Parametergitters in b) ist die Funktion `expand.grid` nützlich.