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

Dortmund, 23.06.09
Abgabe: bis Di, 30.06., 10.00 Uhr an
schiffner@statistik.tu-dortmund.de

Übungen zur Vorlesung

# Wissensentdeckung in Datenbanken

Sommersemester 2009

Blatt 10

**Aufgabe 10.1 (3 Punkte)**
Ein Hobbygärtner möchte die Stiellänge seiner selbst gezüchteten Rosensorte vergrößern. Er vermutet, dass

- die Art der Beleuchtung,
- die Art des Wassers und
- die Düngung

einen Einfluss auf das Wachstum der Pflanzen haben. Da der Gärtner Wechselwirkungen ausschließt und sein Gewächshaus relativ klein ist, entscheidet er sich, ein Screening-Experiment durchzuführen und dabei einen Plackett-Burman-Plan mit 8 Versuchen zu verwenden. Dazu variiert er die 3 Einflussfaktoren jeweils auf zwei Niveaus:

- normale Beleuchtung (kodiert mit -1) und Zusatzbeleuchtung (kodiert mit +1),
- Leitungswasser (-1) und Regenwasser (+1) sowie
- keine Düngung (-1) und Düngung (+1).

In der Datei `rosen.txt` finden Sie die durchschnittlichen Stiellängen der Rosen in cm unter den verschiedenen Versuchsbedingungen.

a) Der Gärtner verwendet die Spalten 1, 3 und 6 des Plackett-Burman-Plans. Stellen sie die Planmatrix $A$ und die Designmatrix $X$ für dieses Experiment auf.

b) Bestimmen Sie die Halbeffekte und die Effekte der drei Einflussfaktoren und interpretieren Sie diese.

**Aufgabe 10.2 (4 Punkte)**
Laden Sie den Iris-Datensatz, den Sie in der R-Einführung kennengelernt haben. Dies ist in R mit Hilfe des Befehls `data(iris)` möglich. Betrachten Sie den Datensatz als Grundgesamtheit. Ziel ist es, Stichproben aus dem Datensatz zu ziehen und damit eine Schätzung für den Erwartungswert der Variable `Petal.Width` zu gewinnen.

Ziehen Sie zunächst eine optimal geschichtete Stichprobe der Größe $N = 20$ aus dem Iris-Datensatz. Betrachten Sie dabei die drei verschiedenen Spezies *virginica*, *versicolor* und *setosa* als Schichten. Gehen Sie folgendermaßen vor: Ermitteln Sie zunächst die Schichtgewichte $w_l$ und Schichtstandardabweichungen $\sigma_l$ in der Grundgesamtheit, d.h. im gesamten Datensatz. Berechnen Sie daraus mit Hilfe des Satzes auf Folie 63 in der Datei `Datengenerierung.pdf` die Größe $N_l$ der Schichtstichproben bei optimaler Schichtung. Ziehen Sie anschließend die geschichtete Stichprobe und berechnen Sie den Mittelwert $\bar{X}_S$ des Merkmals `Petal.Width` sowie seine geschätzte Varianz $\widehat{\text{var}}(\bar{X}_S)$ nach den Formeln in der Vorlesung.

Ziehen Sie außerdem zum Vergleich eine einfache Zufallsstichprobe der Größe $N = 20$ und berechnen ebenfalls den Mittelwert $\bar{X}$ des Merkmals `Petal.Width` und seine geschätzte Varianz. Diese ist gegeben durch

$$\widehat{\text{var}}(\bar{X}) = \frac{1}{N(N-1)} \sum_{i=1}^{N} (X_i - \bar{X})^2.$$

Welcher Schätzer hat die kleinere Varianz?

**Aufgabe 10.3 (3 Punkte)**
Im Netz liegt der Datensatz `adidas.dat`. Dieser enthält die beiden metrischen Variablen `X1` und `X2` sowie eine Klassenvariable `y`.

Führen Sie eine Hauptkomponentenanalyse auf den Variablen `X1` und `X2` durch. Dies ist in R mit Hilfe der Funktion `princomp` möglich.

Führen Sie anschließend eine lineare Diskriminanzanalyse durch.

Stellen Sie die Daten grafisch dar (mit unterschiedlichen Farben für Beobachtungen der verschiedenen Klassen). Zeichnen Sie in diese Grafik die erste Hauptkomponente und die Diskriminanzachse ein. Vergleichen Sie die beiden Projektionen der Daten.