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

Dortmund, 26.04.11
Abgabe: bis Di, 03.05., 10.00 Uhr an
schiffner@statistik.tu-dortmund.de

Übungen zur Vorlesung
Wissensentdeckung in Datenbanken
Sommersemester 2011

Blatt 3

Aufgabe 3.1 (5 Punkte)
In der Vorlesung haben Sie bisher die eindimensionale oder univariate Normalverteilung kennengelernt. Meist beobachten wir aber mehrere Merkmale, z. B. $p \geq 2$ Stück, gleichzeitig. Daher haben wir es mit einem p-dimensionalen Vektor von Zufallsvariablen $\boldsymbol{X} = (X_1, \ldots, X_p)'$ zu tun und betrachten die gemeinsame Verteilung von $X_1, \ldots, X_p$.

Mit $\boldsymbol{x} = (x_1, ..., x_p)' \in \mathbb{R}^p$ lautet die Dichtefunktion der mehrdimensionalen oder multivariaten Normalverteilung

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2} \det(\boldsymbol{\Sigma})^{1/2}} \cdot \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right).$$

Dabei ist $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)' \in \mathbb{R}^p$ der Vektor der Erwartungswerte und $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$ die Kovarianzmatrix.

a) Ziehen Sie 500 Beobachtungen aus der zweidimensionalen Normalverteilung mit Erwartungswertvektor $\boldsymbol{\mu} = (0, 0)'$ und Kovarianzmatrix

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}.$$

Hierbei ist die Funktion `rmvnorm` aus dem Paket `mvtnorm` hilfreich. (Das Paket können Sie in R mit dem Befehl `install.packages("mvtnorm")` installieren. Um es verwenden zu können, müssen Sie das Paket zunächst mit `library(mvtnorm)` laden.)

b) Erstellen Sie einen Scatterplot (Funktion `plot`) der 500 Beobachtungen und zeichnen Sie die Konturlinien der Normalverteilungsdichte ein. Hierzu sind die Funktionen `expand.grid`, `dmvnorm` und `contour` nützlich.

c) Betrachten Sie nur die zur ersten Variable X_1 gehörigen Beobachtungen. Erstellen Sie ein Histogramm (R-Funktion `hist` mit der Option `freq = FALSE`) und zeichnen Sie die Dichte der univariaten Normalverteilung mit Erwartungswert 0 und Varianz 2 ein. Nützliche Funktionen sind `dnorm` und `curve`.

d) Berechnen Sie außerdem die Summe der Beobachtungen zu den beiden Variablen X_1 und X_2. Plotten Sie das Histogramm und zeichnen Sie die Dichte der univariaten Normalverteilung mit Erwartungswert 0 und Varianz 4 ein.

e) Ändern Sie die Kovarianzmatrix in

$$\mathbf{\Sigma}_2 = \begin{pmatrix} 2 & 1.5 \\ 1.5 & 2 \end{pmatrix}$$

und erzeugen Sie dieselben Plots wie in Teil a). Was ändert sich?

Aufgabe 3.2 (5 Punkte)
Ein Hobbygärtner widmet sich der Züchtung von Rosen. Sein Ziel ist es, möglichst langstielige Exemplare zu erhalten. Er vermutet, dass

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

einen Einfluss auf die Stiellänge der gebildeten Blüten 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:

- Dünger A (kodiert mit -1) und Dünger B (kodiert mit $+1$),
- Leitungswasser (-1) und Regenwasser ($+1$) sowie
- kein Rückschnitt (-1) und regelmäßiger Rückschnitt ($+1$).

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.

c) Bei drei Einflussgrößen hätte der Gärtner auch einen Plackett-Burmann-Plan mit nur 4 Versuchen durchführen können. Was sind Vor- und Nachteile seines Vorgehens?