Prof. Dr. Katharina Morik,
Prof. Dr. Claus Weihs
Dipl.-Inform. Klaus Friedrichs,
Dipl.-Stat. Julia Schiffner,
Dr. Issam Ben Khediri

Dortmund, 01.05.12

Abgabe: bis Mi, 09.05., 12.00 Uhr an
friedrichs@statistik.tu-dortmund.de

Übungen zur Vorlesung
Wissensentdeckung in Datenbanken
Sommersemester 2012

Blatt 4

Aufgabe 4.1 (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 mit `library(mvtnorm)` laden.)

b) Erstellen Sie einen Scatterplot (Funktion `plot`) der 500 Beobachtungen und zeichnen Sie die Konturlinien der Normalverteilungsdichte ein. Erzeugen Sie dafür ein 2-dimensionales Gitter von Punkten für einen angemessenen Wertebereich (R-Funktion `expand.grid`), berechnen Sie für alle Punkte die Werte der Dichtefunktion (`dmvnorm`) und zeichnen Sie mithilfe von `contour` die Linien in den Scatterplot ein.

c) Ändern Sie die Kovarianzmatrix in

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

ziehen Sie erneut 500 Beobachtungen und erzeugen Sie denselben Plot wie in Teil b). Was ändert sich?

d) Ändern Sie die Kovarianzmatrix in

$$\mathbf{\Sigma}_3 = \begin{pmatrix} 2 & -1.5 \\ -1.5 & 2 \end{pmatrix},$$

ziehen Sie erneut 500 Beobachtungen und erzeugen Sie denselben Plot wie in Teil b). Was ändert sich?

e) Im Skript sind Kovarianz- und Korrelationsmatrizen für den 2-dimensionalen Fall definiert. Überlegen Sie, wie diese im p-dimensionalen Fall aussehen, und schreiben Sie die Matrizen für den 3-dimensionalen Fall auf.

f) Von Übungsblatt 3 kennen Sie bereits den Schweizer-Banknoten-Datensatz `bank.txt`. Berechnen Sie die Kovarianz- sowie die Korrelationsmatrix der 100 echten Scheine (in R: `cov` bzw. `cor`).

Interpretieren Sie die Ergebnisse. Dafür kann es hilfreich sein, die Scatterplot-Matrix aus Aufgabe 3.2.1 zu betrachten.

g) Berechnen Sie nun die Kovarianz- sowie die Korrelationsmatrix der 100 gefälschten Scheine. Vergleichen Sie die Ergebnisse mit denen aus Teil f). Was ist anders?

Aufgabe 4.2 (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,
- der Rückschnitt und
- die Art des Lichts

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 4 Einflussfaktoren jeweils auf zwei Niveaus:

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

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, 6 und 7 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 vier Einflussfaktoren und interpretieren Sie diese.