Prof. Dr. Katharina Morik,
JProf. Dr. Uwe Ligges
Dipl.-Inform. Felix Jungermann,
Dipl.-Stat. Julia Schiffner

Dortmund, 06.07.10
Abgabe: bis Di, 13.07., 10.00 Uhr an
schiffner@statistik.tu-dortmund.de

Übungen zur Vorlesung

Wissensentdeckung in Datenbanken

Sommersemester 2010

Blatt 12

Aufgabe 12.1 (5 Punkte)
Auf der Homepage liegen die Datensätze `orange.train.txt` und `orange.test.txt`. Es handelt sich um ein künstliches Klassifikationsproblem mit 2 Klassen (Variable `class`) und zehn erklärenden Variablen `F1` bis `F10`.

a) Trainieren Sie eine QDA (Funktion `qda` aus dem Paket `MASS`) auf den Trainingsdaten und berechnen Sie die Testdatenfehlerrate.

b) Versuchen Sie herauszufinden, welche Variablen nützlich sind, um die Klassen zu trennen. Wie Sie dabei vorgehen, ist Ihnen überlassen. Sie können z. B. verschiedene Kennzahlen wie Klassenmittel, Klassenkovarianzen oder die Korrelationen der erklärenden und der Zielvariable etc. berechnen oder Grafiken betrachten wie z. B. eine Scatterplotmatrix, Histogramme der einzelnen Variablen für die beiden Klassen, oder die Klassengrenzen der QDA aus Teil a) mit `partimat` visualisieren etc. Probieren Sie mindestens eine Methode aus und interpretieren Sie die Ergebnisse.

c) Benutzen Sie die Funktion `stepclass` aus dem Paket `klaR`, um eine schrittweise Variablenselektion für die QDA durchzuführen. Trainieren Sie anschließend eine QDA auf den Trainingsdaten, wobei Sie aber nur die von `stepclass` selektierten Variablen benutzen, und berechnen Sie die Fehlerrate auf den Testdaten.

Aufgabe 12.2 (5 Punkte)
Die Datei `aussenhandel.txt` enthält Viermonatsdaten (Trimesterdaten) für die Einnahmen y_t der BRD aus dem Reiseverkehr mit dem Ausland (gemessen in Mio. DM).

a) Stellen Sie die Zeitreihe graphisch dar. Liegt ein Trend vor? Wenn ja, um welche Art von Trend handelt es sich?

b) Schätzen Sie den Trend der Zeitreihe durch eine lineare Trendfunktion nach der Methode der Kleinsten Quadrate. Zeichnen Sie die geschätzte Trendgerade in das Schaubild der Originalzeitreihe mit ein. Hierzu sind in R die Funktionen `lm` und `abline` nützlich.

Betrachten Sie die beiden stochastischen Prozesse

$$y_t = 0.2 - 0.9y_{t-1} + \epsilon_t \qquad \text{und}$$
$$y_t = -0.2 + 1.25y_{t-1} + \epsilon_t$$

Die ϵ_t seien unabhängig und identisch $N(0, 0.25)$ verteilt.

c) Um welche Art von Prozessen handelt es sich? Sind die Prozesse stationär?

d) Simulieren Sie 500 Beobachtungen aus den beiden stochastischen Prozessen und plotten Sie die erzeugten Zeitreihen. Berechnen Sie, falls es sich um einen stationären Prozess handelt, jeweils den Erwartungswert und zeichnen Sie ihn in die Grafik mit ein.

e) Glätten sie die Zeitreihen mithilfe eines einfachen gleitenden Durchschnitts der Länge 20. Zeichnen Sie die geglätteten Zeitreihen in die bereits erzeugten Plots mit ein. Für die Berechnung gleitender Durchschnitte in R stehen mehrere Funktionen in verschiedenen Paketen zur Verfügung, z. B. die Funktion `runmean` im Paket `caTools`.