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

Dortmund, 07.07.09
Abgabe: bis Di, 14.07., 10.00 Uhr an
schiffner@statistik.tu-dortmund.de

Übungen zur Vorlesung

Wissensentdeckung in Datenbanken

Sommersemester 2009

Blatt 12

Aufgabe 12.1 (3 Punkte)
Im Netz liegen die Datensätze `mpg_train.txt` und `mpg_test.txt` sowie eine Datei mit Informationen zu den Daten (`mpg_info.txt`). Der Verbrauch mpg (miles per gallon) soll anhand der Merkmale cylinders, displacement, horsepower, weight und acceleration vorhergesagt werden. Passen Sie eine Hauptkomponenten- und eine PLS-Regression auf der Basis von Korrelationen an (in R mit den Funktionen `pcr` und `plsr` im Paket pls möglich) und erstellen Sie die zugehörigen Biplots.

Berechnen Sie die Prognosegüte der Modelle auf dem Testdatensatz (mit den Funktionen `MSEP`, `R2` möglich). Wieviele Komponenten sollten jeweils verwendet werden?

Aufgabe 12.2 (3 Punkte)
Der Datensatz `Tuba.txt` enthält ca. 0.4 Sekunden eines Tons, der auf einer B-Tuba gespielt wurde. Ziel ist herauszufinden, um welche Note es sich handelt. Dazu hilfreiche Funktionen finden Sie im R-Paket tuneR.

Berechnen Sie das Periodogramm der Zeitreihe (mit der Funktion `periodogram` möglich) und plotten Sie es. Beachten Sie, dass Sie die Zeitreihe zunächst mittels der Funktion `Wave` in ein Objekt der Klasse `Wave` umwandeln müssen. Da 4402 Beobachtungen in ca. 0.4 Sekunden gemacht wurden, setzen Sie die sampling rate auf `samp.rate = 11025`. Setzen Sie außerdem `bit = 16`. Um auf einzelne Werte des berechneten Periodogramms wie z. B. `freq` zuzugreifen, muss das @-Zeichen und nicht wie sonst gewohnt das $-Zeichen verwendet werden.

Beurteilen Sie anhand des Periodogramms, aus welchen Grund- und Obertönen das Signal besteht. Um die zu den Frequenzen gehörigen Notennamen herauszufinden, sind die Funktionen `noteFromFF` und `notenames` nützlich.

Aufgabe 12.3 (4 Punkte)
Programmieren Sie den Hastings-Metropolis-Algorithmus, um 1000 Zufallszahlen aus der t-Verteilung mit 10 Freiheitsgraden zu ziehen. Verwenden Sie als Übergangfunktion die Independence Chain Variante ($q(x, y) = q_l(y)$) mit q_l als Dichte der stetigen Gleichverteilung auf dem Intervall $(-4, 4)$. Schreiben Sie eine Funktion, die neben den erzeugten Zufallszahlen auch die Akzeptanzrate ausgibt.
Um die Qualität der erzeugten Zufallszahlen zu kontrollieren, zeichnen Sie das Histogramm der erzeugten Zufallszahlen und überlagern Sie es mit der Dichtefunktion der t-Verteilung mit 10 Freiheitsgraden.

Nützlich in R sind die Funktionen `runif` zur Erzeugung stetig gleichverteilter Zufallszahlen, `dunif` und `dt` zur Auswertung der Dichtefunktionen der stetigen Gleichverteilung und der t-Verteilung sowie `if` und `else` für logische Abfragen. Ein Histogramm kann mit Hilfe des Befehls `hist` erzeugt werden. Achten Sie darauf, die Option `freq = FALSE` zu setzen. Zum Einzeichnen der Dichtefunktion ist der Befehl `curve` hilfreich.