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

Dortmund, 01.06.10
Abgabe: bis Di, 08.06., 10.00 Uhr an
schiffner@statistik.tu-dortmund.de

Übungen zur Vorlesung

Wissensentdeckung in Datenbanken

Sommersemester 2010

Blatt 7

Aufgabe 7.1 (7 Punkte)
Programmieren Sie **eigenständig** in R die folgenden Verfahren aus der Vorlesung:

a) Lineare Diskriminanzanalyse nach Fisher,

b) Allgemeine Lineare Diskriminanzanalyse,

c) Quadratische Diskriminanzanalyse,

Schreiben Sie jeweils eine Trainingsfunktion, in der Sie das Modell an die Daten anpassen, und eine Vorhersagefunktion, die die Klassenzugehörigkeiten für einen Testdatensatz vorhersagt. Die Trainingsfunktion sollte als Argumente die Trainingsdaten und deren Klassenzugehörigkeit haben. Die Vohersagefunktion sollte vom trainierten Modell und dem Testdatensatz abhängen. Bei der Fisher LDA sollte man außerdem die Anzahl der Diskriminanzkomponenten, die zur Vorhersage der Klassenzugehörigkeit benutzt werden, einstellen können.

Nützliche Funktionen in R sind

- `%*%`, um Matrizen miteinander zu multiplizieren,
- `t` zum Transponieren von Matrizen,
- `solve` zur Invertierung von Matrizen,
- `eigen` zur Berechnung von Eigenwerten und -vektoren,
- `det` zur Berechnung von Determinanten,
- `mahalanobis` zur Berechnung von Mahalanobis-Distanzen,
- `by` zur klassenweisen Berechnung von Mittelwerten und Kovarianzmatrizen.

Aufgabe 7.2 (3 Punkte)
Auf der Homepage liegen die Datensätze `wine.train.txt` und `wine.test.txt` sowie eine kurze Beschreibung der Daten (`wine.names`).

Trainieren Sie die drei Verfahren aus Aufgabe 7.1 auf dem Trainingsdatensatz, sagen Sie die Klassenzugehörigkeit der Testdaten vorher und berechnen Sie die Fehlerraten.

Was ist bei der Fisher LDA die maximale Anzahl an Diskriminanzkomponenten für dieses Klassifikationsproblem? Probieren Sie die möglichen Anzahlen an Diskriminanzkomponenten aus und vergleichen Sie die resultierenden Fehlerraten auf dem Testdatensatz. Stellen Sie außerdem jeweils die Projektion der Daten auf die Diskriminanzkomponenten grafisch dar und beschreiben sie kurz das Ergebnis.