Maschinelles Lernen
Übungsblatt 5

Prof. Dr. Dr. Lars Schmidt-Thieme, Osman Akcatepe
Wirtschaftsinformatik und Maschinelles Lernen (ISMLL)
Universität Hildesheim

6. Dezember 2011
Abgabe: 12. Dezember bis 14.15

Diskriminanzanalyse (5 Punkte)

Wissenschaftler haben die Böden von Iowa, welche ein bestimmtes Bakterium enthalten (Klasse 1), mit andere Böden, die es nicht enthalten (Klasse 2), verglichen. Dabei haben sie die Variablen x_1 (pH-Wert) und x_2 (Stickstoffgehalt) beobachtet. Die Anzahl der Instanzen pro Klasse, der Mittelwert der Vektoren und die Kovarianzmatrizen für die zwei Bodenarten seien wie folgt gegeben:

$$n_1 = 13, \qquad n_2 = 10$$
$$\mathbf{m}_1 = \begin{pmatrix} 7.8 \\ 45 \end{pmatrix}, \qquad \mathbf{m}_2 = \begin{pmatrix} 5.9 \\ 20.8 \end{pmatrix}$$
$$\mathbf{S}_{W1} = \begin{pmatrix} 0.5 & 4.5 \\ 4.5 & 147.2 \end{pmatrix}, \quad \mathbf{S}_{W2} = \begin{pmatrix} 0.1 & 0.2 \\ 0.2 & 24.2 \end{pmatrix}$$

a) Stellen Sie die Diskriminanzfunktionen für die beiden Klassen auf.

b) Ordnen Sie die Beobachtung $x = \begin{pmatrix} 6 & 52.5 \end{pmatrix}^T$ einer der beiden Klassen zu.

c) Handelt es sich hier um lineare oder quadratische Diskriminanzanalyse? Nennen Sie die Unterschiede zwischen LDA und QDA.

Daten einlesen in R (2 Punkte)

Lesen Sie Kapitel 6 und 7 von „An Introduction to R“.

a) Was ist der Unterschied zwischen einer Liste und einem Array in R? Nennen Sie drei Möglichkeiten, wie man auf die Komponenten einer Liste zugreifen kann. Warum sind *data frames* besonders wichtige Konstrukte in R?

b) Laden Sie den Wein-Datensatz aus dem *UCI Machine Learning Repository* (`http://archive.ics.uci.edu/ml/datasets/Wine`) herunter und laden Sie ihn in R.

LDA und QDA in R (3 Punkte)

Laden Sie die *MASS*-Bibliothek mittels `library(MASS)`. Erstellen Sie zwei Klassifikationsmodelle für den Wein-Datensatz, welches die erste Variable als Zielvariable (Klasse) und die restlichen Variablen als Prädiktorvariablen auffasst: Lineare Diskriminanzanalyse (`lda`) und quadratische Diskriminanzanalyse (`qda`). Die Funktionen `lda` und `qda` werden ähnlich wie `lm` und `glm` aufgerufen, z.B. `glm(Survived ~., data=Titanic, family=binomial)` für eine logistische Regression auf dem in R enthaltenen *Titanic*-Datensatz.

a) Wenn Sie `lda` und `qda` mit dem Parameter `CV=1` aufrufen, bekommen Sie für jeden Eintrag in Ihrem Datensatz eine Vorhersage: `result <- qda(Survived ~., data=Titanic, CV=1)`. Vergleichen Sie die Methoden LDA und QDA, indem Sie jeweils `result$class` mit der ersten Spalte des Datensatzes abgleichen.

b) Warum können Sie nicht einfach ein logistisches Regressionsmodell für den Weindatensatz aufstellen? Begründen Sie Ihre Antwort.