Prof. Dr. Katharina Morik,
JProf. Dr. Uwe Ligges
Dipl.-Math. Christian Bockermann,
Dipl.-Stat. Gero Szepannek

Dortmund, 3. Juni
Abgabe: bis **Mi**, 11.6., 23:59 h an
szepannek@statistik.uni-dortmund.de

Übungen zur Vorlesung

Wissensentdeckung in Datenbanken

Sommersemester 2008

Blatt 7

Aufgabe 7.1 (6 Punkte)
Im Netz liegt der Datensatz: `beispiel1.txt`. Dieser enthält in Spalte X Beobachtungswerte und in Spalte Y Klassenlabels $i \in \{1, 2\}$.

a) Schätzen sie die Verteilungsparameter μ_i beider Klassen durch den jeweiligen Klassenmittelwert und stellen Sie die Verteilungen grafisch dar unter Annahme von Normalverteilung mit und $\sigma_i = 1,\ i = 1, 2$.
b) Berechnen Sie die (datenabhängige) Bayes Klassifikationsregel auf Basis der geschätzten Verteilungsparameter bei symmetrischen Kosten $c(i, j) = 1 - I_{\{j\}}(i)$ (mit $I_{\{.\}}(\cdot)$ der Indikatorfunktion) und gleichen a priori Wahrscheinlichkeiten der Klassen?
c) Wie ändert sich die optimale Klassifikationsregel, wenn Ihnen zusätzlich bekannt ist, dass eine Beobachtung – wenn Sie x nicht kennen – mit einer Wahrscheinlichkeit von 2/3 aus Klasse zwei stammt?
d) Bestimmen sie das minimale Risiko, d.h. denjenigen Klassifikationsfehler der unvermeidbar ist.

Aufgabe 7.2 (4 Punkte)
Im Netz liegt der Datensatz `spam.txt`, sowie eine weitere Datei `info.txt`.
a) Beschreiben Sie kurz den Datensatz sowie die Bedeutung der Berücksichigung von Fehlklassifikationskosten für das gegebene Klassifikationsproblem! Wie lautet die datenunabhängige Klassifikationsregel?
b) Bilden Sie ein Klassifikationsmodell für die Variable `type` mit Hilfe des *Naive Bayes* Ansatzes (in `R` in den Paketen `klaR` bzw. `e1071` zu finden)!
c) Bilden Sie ein Klassifikationsmodell für die Variable `type` mit Hilfe von *logistischer Regression* (in `R` durch die Funktion `glm`)!
Beschreiben Sie die Modelle!