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

Dortmund, 05.05.09
Abgabe: bis Di, 12.05., 10.00 Uhr an
schiffner@statistik.tu-dortmund.de

Übungen zur Vorlesung
Wissensentdeckung in Datenbanken
Sommersemester 2009

Blatt 4

Aufgabe 4.1 (7 Punkte)
Gegeben sei ein Klassifikationsproblem mit zwei Klassen. Nehmen Sie an, dass die Daten aus zwei univariaten Normalverteilungen mit $\mu_1 = 0$, $\mu_2 = 1$ und $\sigma_1 = \sigma_2 = 1$ stammen. Die a priori Wahrscheinlichkeiten π_1 und π_2 der beiden Klassen seien gleich.

a) Stellen Sie die beiden Dichtefunktionen $\phi(x \,|\, \mu_1, \sigma_1)$ und $\phi(x \,|\, \mu_2, \sigma_2)$ gemeinsam in einem Diagramm dar (in R sind die Funktionen `curve` und `dnorm` nützlich).

b) Berechnen Sie die a posteriori Wahrscheinlichkeiten der beiden Klassen und stellen Sie sie ebenfalls gemeinsam in einem Diagramm dar.

c) Wie lautet die Bayes-Regel bei symmetrischen Kosten $c(i, j) = 1 - I_{\{j\}}(i)$ (mit $I_{\{\cdot\}}(\cdot)$ der Indikatorfunktion)?

   Zeichnen Sie die Entscheidungsgrenze in Ihre Grafiken mit ein (in R ist die Funktion `abline` nützlich).

d) Leiten Sie eine Formel für die Fehlklassifikationswahrscheinlichkeit

$$P\big(y_{\text{Regel}}(x) \neq y_{\text{wahr}}(x)\big)$$

   in Abhängigkeit von den Dichten $\phi(x \,|\, \mu_1, \sigma_1)$ und $\phi(x \,|\, \mu_2, \sigma_2)$ und den a priori Wahrscheinlichkeiten π_1 und π_2 her.

   Berechnen Sie die Fehlklassifikationswahrscheinlichkeit für gleiche a priori Wahrscheinlichkeiten der Klassen.

Nehmen Sie nun an, dass die Beobachtungen mit einer Wahrscheinlichkeit von $\pi_2 = 4/5$ aus Klasse 2 stammen.

e) Stellen Sie die beiden Funktionen $\pi_1 \cdot \phi(x \,|\, \mu_1, \sigma_1)$ und $\pi_2 \cdot \phi(x \,|\, \mu_2, \sigma_2)$ sowie die a posteriori Wahrscheinlichkeiten der Klassen jeweils gemeinsam in einem Diagramm dar.

f) Wie ändert sich die optimale Klassifikationregel? Zeichnen Sie die Entscheidungsgrenze in Ihre Grafiken mit ein.

   Wie ändert sich die Fehlklassifikationswahrscheinlichkeit?

Aufgabe 4.2 (3 Punkte)
Auf der Homepage liegt der Datensatz `bank2.txt`, sowie eine weitere Datei `info.txt`.

a) Beschreiben Sie kurz die Bedeutung der Berücksichtigung von Fehlklassifikationskosten für das gegebene Klassifikationsproblem.

   Wie lautet die datenunabhängige Klassifikationsregel und welche Fehlklassifikationswahrscheinlichkeit würden Sie erwarten?

b) Bilden Sie ein Klassifikationsmodell für die Variable `Class` mit Hilfe des *Naïve Bayes* Ansatzes (in R in den Paketen `klaR` bzw. `e1071` zu finden) und sagen Sie die Klassenzugehörigkeit aller Geldscheine vorher.

c) Bilden Sie Klassifikationsmodelle für die Variable `Class` mit Hilfe von *logistischer Regression* (in R durch die Funktion `glm`) basierend auf allen und auf einzelnen erklärenden Variablen und sagen Sie die Klassenzugehörigkeit aller Geldscheine vorher.

   Was fällt Ihnen auf?