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

Dortmund, 18.05.10
Abgabe: bis Di, 25.05., 10.00 Uhr an
schiffner@statistik.tu-dortmund.de

Übungen zur Vorlesung
Wissensentdeckung in Datenbanken
Sommersemester 2010

Blatt 5

Aufgabe 5.1 (7 Punkte)
Gegeben sei ein Klassifikationsproblem mit zwei Klassen. Nehmen Sie an, dass die Daten aus zwei univariaten Normalverteilungen mit $\mu_0 = 0$, $\mu_1 = 1$ und $\sigma_0 = \sigma_1 = 1$ stammen. Die a priori Wahrscheinlichkeiten π_0 und π_1 der beiden Klassen seien zunächst gleich.

a) Stellen Sie die beiden Funktionen $\pi_0 \cdot f(x \,|\, \mu_0, \sigma_0)$ und $\pi_1 \cdot f(x \,|\, \mu_1, \sigma_1)$ gemeinsam in einem Diagramm dar. Dabei bezeichnet f die Dichtefunktion der univariaten Normalverteilung. (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 datenabhängige Bayes-Regel bei symmetrischen Kosten $c(i, j) = I(j \neq i)$ (mit I der Indikatorfunktion und $i, j \in \{0, 1\}$)?

   Zeichnen Sie die Entscheidungsgrenze zur Vorhersage der Klassenzugehörigkeit in Ihre Grafiken mit ein (in R ist z. B. die Funktion `abline` nützlich).

d) Leiten Sie eine Formel für die Fehlklassifikationswahrscheinlichkeit

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

   in Abhängigkeit von den Dichtefunktionen $f(x \,|\, \mu_0, \sigma_0)$ und $f(x \,|\, \mu_1, \sigma_1)$ und den a priori Wahrscheinlichkeiten π_0 und π_1 her.

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

Nehmen Sie nun an, dass die Beobachtungen mit einer Wahrscheinlichkeit von $\pi_1 = 4/5$ aus Klasse 1 stammen.

e) Stellen Sie die beiden Funktionen $\pi_0 \cdot f(x \,|\, \mu_0, \sigma_0)$ und $\pi_1 \cdot f(x \,|\, \mu_1, \sigma_1)$ sowie die a posteriori Wahrscheinlichkeiten der Klassen jeweils gemeinsam in einem Diagramm dar.

f) Wie ändert sich die optimale Klassifikationregel? Zeichnen Sie die Entscheidungsgrenze zur Vorhersage der Klassenzugehörigkeit in Ihre Grafiken mit ein. Wie ändert sich die Fehlklassifikationswahrscheinlichkeit?

Aufgabe 5.2 (3 Punkte)
Erzeugen Sie mithilfe des folgenden R-Codes einen Datensatz.

```
x0 <- rnorm(40)
x1 <- rnorm(160, mean = 1)
daten <- data.frame(x = c(x0, x1), y = factor(c(rep(0, 40), rep(1, 160))))
```

Schätzen Sie die a posteriori Wahrscheinlichkeiten der Klasse 1 mithilfe einer logistischen Regression (in R sind die Funktionen `glm` mit Argument `family = binomial` und `predict` mit `type = "response"` nützlich). Sagen Sie anhand der a posteriori Wahrscheinlichkeiten die Klassenzugehörigkeit der Beobachtungen in `daten` vorher. Benutzten Sie dabei den Schwellenwert $\tau = 0.5$.

Wie groß ist die Fehlerrate? Vergleichen Sie sie mit der Fehlerrate der Bayes-Regel aus Aufgabe 5.1 f). (Sie beträgt ca. 0.19.)