Prof. Dr. Katharina Morik,
Prof. Dr. Claus Weihs
Dipl.-Inform. Christian Bockermann,
Dipl.-Stat. Julia Schiffner

Dortmund, 03.05.11
Abgabe: bis Di, 10.05., 10.00 Uhr an
schiffner@statistik.tu-dortmund.de

Übungen zur Vorlesung
Wissensentdeckung in Datenbanken
Sommersemester 2011

Blatt 4

Aufgabe 4.1 (4 Punkte)
Den weltberühmten Iris-Datensatz haben Sie bereits in der Software-Einfürung kennengelernt. Laden Sie ihn in R mit Hilfe des Befehls `data(iris)`. Betrachten Sie den Datensatz als Ihre (zugegebenermaßen kleine) Grundgesamtheit/Zielpopulation. Ihr Ziel ist es, Stichproben aus dem Datensatz zu ziehen und damit eine Schätzung für den Erwartungswert der Variable `Petal.Width` zu gewinnen.

In einer Simulation soll die Güte der Schätzer basierend auf einer einfachen Zufallsauswahl und auf einer geschichteten Stichprobe verglichen werden.

- Ziehen Sie zunächst eine einfache Zufallsstichprobe der Größe $N = 20$ und berechnen den Mittelwert $\bar{X}$ des Merkmals `Petal.Width`. Berechnen Sie außerdem seine geschätzte Varianz. Diese ist gegeben durch

$$\widehat{var}(\bar{X}) = \frac{1}{N(N-1)} \sum_{i=1}^{N} (X_i - \bar{X})^2$$

- Ziehen Sie zum Vergleich eine *optimal geschichtete Stichprobe* der Größe $N = 20$ aus dem Iris-Datensatz. Betrachten Sie dabei die verschiedenen Spezies virginica, versicolor und setosa als Ihre $L = 3$ Schichten. Gehen Sie folgendermaßen vor: Ermitteln Sie zunächst die Schichtgewichte w_l und Schichtstandardabweichungen σ_l in der Grundgesamtheit, d.h. im gesamten Datensatz. Berechnen Sie daraus mit Hilfe des Satzes auf Folie 120 in der Datei `Statistik_Teil1.pdf` die Größe N_l der Schichtstichproben bei optimaler Schichtung. (Hierbei ist geeignetes Runden ist erforderlich.) Ziehen Sie anschließend die geschichtete Stichprobe und berechnen Sie den Mittelwert $\bar{X}_S$ des Merkmals `Petal.Width` sowie seine geschätzte Varianz $\widehat{var}(\bar{X}_S)$ nach den Formeln in der Vorlesung.

Wiederholen Sie beides jeweils 100-mal. Welcher Schätzer hat im Mittel die kleinere Varianz?

Hilfreiche Befehle in R:

- `sample` – zum Ziehen der Stichproben,
- `table` – zum Ermitteln von Häufigkeiten.

Aufgabe 4.2 (6 Punkte)
Gegeben sei ein Klassifikationsproblem mit zwei Klassen. Nehmen Sie an, dass die Daten aus zwei univariaten Normalverteilungen mit $\mu_0 = 0$, $\mu_1 = 2$ und $\sigma_0 = \sigma_1 = 2$ stammen. Die a priori Wahrscheinlichkeiten π_0 und π_1 der beiden Klassen seien zunächst gleich.

a) Stellen Sie die beiden Dichtefunktionen $f(x \mid \mu_0, \sigma_0)$ und $f(x \mid \mu_1, \sigma_1)$ 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 identischen 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 $f(x \mid \mu_0, \sigma_0)$ und $f(x \mid \mu_1, \sigma_1)$ 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) Wie groß wäre vermutlich in dieser Situation die Fehlerrate der datenunabhängigen Regel: 'Wähle die häufigste Klasse im Lerndatensatz'?

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

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