Prof. Dr. Katharina Morik,
Prof. Dr. Claus Weihs
Dipl.-Inform. Klaus Friedrichs,
Dipl.-Stat. Julia Schiffner,
Dr. Issam Ben Khediri

Dortmund, 08.05.12

Abgabe: bis Mi, 16.05., 12.00 Uhr an
schiffner@statistik.tu-dortmund.de

Übungen zur Vorlesung
Wissensentdeckung in Datenbanken
Sommersemester 2012

Blatt 5

Aufgabe 5.1 (4 Punkte)
Den weltberühmten Iris-Datensatz haben Sie bereits bei mehreren Gelegenheiten kennengelernt. Laden Sie ihn in R mit Hilfe des Befehls `data(iris)`. Betrachten Sie den Datensatz als Ihre (zugegebenermaßen kleine) Grundgesamtheit/Zielpopulation der Größe $M = 150$. Ihr Ziel ist es, Stichproben aus dem Datensatz zu ziehen und damit eine Schätzung für den Erwartungswert der Variable `Petal.Width` zu gewinnen.

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

a) Berechnen Sie zunächst für eine *einfache* Zufallsstichprobe der Größe $N = 15$ den minimal und maximal möglichen Erwartungswertschätzer $\bar{X}$.

b) Berechnen Sie zum Vergleich den minimal und maximal möglichen Erwartungswertschätzer $\bar{X}_S = \sum_{l=1}^{L} w_l \bar{X}_l$ für eine *optimal geschichtete Stichprobe* der Größe $N = 15$. 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. basierend auf dem gesamten Iris-Datensatz). Berechnen Sie daraus mit Hilfe des Satzes aus der Vorlesung (Folie 128) die Größe N_l der Schichtstichproben bei optimaler Schichtung. (Hierbei ist geeignetes Runden erforderlich.)

Wie groß ist jeweils die Spannweite?

Hilfreiche Befehle in R:

- `table` – zum Ermitteln von Häufigkeiten,
- `sort` – zum Sortieren.

Aufgabe 5.2 (6 Punkte)
Gegeben sei ein Klassifikationsproblem mit zwei Klassen. Nehmen Sie an, dass die Daten aus zwei univariaten Normalverteilungen mit Erwartungswerten $\mu_0 = 0$, $\mu_1 = 1$ und Standardabweichungen $\sigma_0 = \sigma_1 = 1$ stammen. Die a priori Wahrscheinlichkeiten π_0 und π_1 der beiden Klassen seien zunächst gleich.

a) Stellen Sie die beiden Dichtefunktionen $\pi_0 \cdot f(x \mid \mu_0, \sigma_0)$ und $\pi_1 \cdot f(x \mid \mu_1, \sigma_1)$ gemeinsam in einem Diagramm dar. Hierbei bezeichnet f die Dichtefunktion der Normalverteilung. (In R sind die Funktionen `curve` und `dnorm` nützlich.)

b) Berechnen Sie die a posteriori Wahrscheinlichkeiten $P(A_i \mid x)$ 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 $\pi_1 = 9/10$ aus Klasse 1 stammen.

e) Wie groß wäre in dieser Situation die Fehlerrate der datenunabhängigen Regel: 'Wähle die häufigste Klasse im Lerndatensatz'?

f) Stellen Sie die beiden Funktionen $\pi_0 \cdot f(x \mid \mu_0, \sigma_0)$ und $\pi_1 \cdot f(x \mid \mu_1, \sigma_1)$ 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?