



Statistik III: Multivariate Stat. Verfahren

Übungsblatt 8

Bearbeitung: Do. 8.7.2004, 16.00 Uhr.

Alle Aufgaben können im Mac-Cip Pool bearbeitet werden.

- Wir betrachten den **Crabs** Datensatz.
 - Erstellen Sie mit der Funktion `rpart` aus der `rpart` Library einen Klassifikationsbaum für die Variable `Sex`.
 - Plotten Sie für die zwei im Baum am häufigsten vorkommenden Variablen einen Sectioned Scatterplot, in dem Sie alle Splitpunkte dieser Variablen als Linien (mittels `lines`, oder `abline`) eintragen.
 - Erstellen Sie die Konfusionsmatrix der Lösung aus (a). Wie würden Sie das Ergebnis bewerten?
- Betrachtet wird erneut der **Crabs** Datensatz.
 - Erstellen Sie einen Klassifikationsbaum wie in 1.(a), benutzen Sie aber nun nicht die rohen Daten, sondern die Hauptkomponenten der 5 Messgrößen.
 - Welche Komponenten werden (warum?) im Baum verwendet? Erstellen Sie wieder einen Sectioned Scatterplot für die meist benutzten Komponenten, und berechnen Sie die Konfusionsmatrix.
 - Erstellen Sie nun einen Baum auf den linearen Diskriminanten der 4 Gruppen (`sex * species`) der Crabs Daten (die nötige Rotationsmatrix erhalten sie aus der Komponente `scaling` der `lda` Funktion). Wie gut ist die dadurch gewonnene Klassifikation?
- Erstellen Sie mittels `rpart` einen Klassifikationsbaum für die *Studienrichtung* des **Math Students** Datensatz. Was schließen Sie aus dem erhaltenen Baum? Wie könnten Sie den Baum verbessern?
Erstellen Sie nun einen Regressionsbaum für die *Diplomnote*. Welche Input Variablen sind dafür sinnvoll? Wie gut ist die Vorhersage?
- Analysieren Sie die **Credit** Daten mit Mondrian mit Barcharts und Mosaic Plots. Suchen Sie dabei durch geschickte Selektionen nach möglichst großen Gruppen R_m , in denen die Zielvariable y_i , $i \in R_m$ *creditability* so homogen wie möglich ist.
Bestimmen Sie für mindestens 3 dieser Gruppen:
 - Support:** Größe der Gruppe R_i
 - Confidence:** Prozent der "richtig" klassifizierten Daten in R_i .
- Erstellen Sie in KLIMT einen Klassifikationsbaum für die Zielvariable *creditability* der **Credit** Daten. Welche Variablen verwenden Sie dabei — warum? Analysieren Sie das Ergebnis der Klassifikation in KLIMT und bewerten Sie die Qualität der Vorhersage.