Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Hans-Peter Kriegel
Dr. Alexey Pryakhin

KDD Praktikum
SoSe 2007
Übungsblatt 2

Abnahme am 04.06.2007.

Hinweise:

(1) Die für dieses Übungsblatt relevanten Datensätze sind verfügbar unter: `http://www.dbs.ifi.lmu.de/Lehre/KDD_Praktikum/`.

(2) Die neueste Version von WEKA und YALE kann von der Webseite des KDD-Praktikums heruntergeladen werden.

(3) Fassen Sie Experimentsettings und -ergebnisse **aller** Aufgaben in einer übersichtlichen Form (z.B. tabellarisch) zusammen.

Aufgabe 2-1 Regressionsanalyse mit WEKA und YALE

(a) Welche Methoden für die Regressionsanalyse bietet WEKA und YALE an? Welche Gütemaße für die Regressionsanalyse gibt es in WEKA und YALE? Was bedeuten diese Gütemaße? Welche Variablenselektionsmethoden gibt es in YALE?

(b) Erstellen Sie mit Hilfe von WEKA ein Regressionsmodell für den Datensatz "naehrwerte.arff" bzgl. der Zielvariable "CALORIES". Nutzen Sie dabei folgende Regressionsmethoden: SVMreg, LineareRegression, PaceRegression. Verwenden Sie in Ihren Experimenten zehnfache Überkreuzvalidierung. Untersuchen Sie verschiedene Parametereinstellungen bei diesen Algorithmen. Das Ziel dabei ist es, das "beste" Regressionsmodell (siehe die Gütemaße in der Teilaufgabe 2-1 (a)) zu erreichen.

(c) Wiederholen Sie die Teilaufgabe (b) in YALE. Wenden Sie verschiedene Variablenselektionsmethoden (siehe die Teilaufgabe 2-1 (a)) an, um bessere Ergebnisse zu bekommen. Untersuchen Sie die Signifikanz des Regressionsmodells anhand von ANOVA und T-Test.

Aufgabe 2-2 Regressionsanalyse mit YALE

(a) Stellen Sie sich vor, dass eine Versicherungsgesellschaft uns Ihre Daten "crash.dat" in einem ".dat"-Format zur Verfügung gestellt hat. Transformieren Sie die Daten in ein passendes Format. Laden Sie die Daten in YALE. Nehmen Sie "HEAD INJURY SEVERITY" als Zielvariable und die restlichen Variablen als Prädiktoren. Berechnen Sie das beste multiple Regressionsmodell. Verwenden Sie hierbei die Visualisierungsmethoden (z.B. ROCChat, LiftChat usw.), um die Güte des Modells zu evaluieren.

(b) Kombinieren Sie die vier Variablen, die den Schadensaufwand beschreiben, in eine Variable. Die Wahl der Kombinationsfunktion ist Ihnen überlassen. Berechnen Sie das beste multiple Regressionsmodell für die Zielvariable "HEAD INJURY SEVERITY". Vergleichen Sie Ihre Ergebnisse mit den Ergebnissen aus der Teilaufgabe 2-2 (a).

Aufgabe 2-3 Logistische Regressionsanalyse mit YALE

(a) Laden Sie den "adult.dat" Datensatz in YALE. Erstellen Sie ein logistisches Regressionsmodell für die Response-Variable "$income$". Untersuchen Sie hierbei verschiedene Regressionsmethoden (z.B. SVMreg usw.).

(b) Untersuchen Sie die alternativen Modellierungen der Prädiktor-Variable "age", indem Sie z.B. die nicht linearen Beziehungen durch die Einfürung einer neuen Variable "age^2" modellieren. Vergleichen Sie Ihre Ergebnisse mit den Ergebnissen aus der Teilaufgabe 2-3 (a).