

Übungen zur Vorlesung
Wissensentdeckung in Datenbanken
Sommersemester 2007
Blatt 1

Aufgabe 1.1

Die erste Aufgabe ist mit dem in der ersten Übungsstunde (13.04.) vorgestellten Programm YALE zu lösen. YALE ist im Statistik-Pool installiert. Sie können sich das Programm allerdings auch für private Versuche oder Vorarbeiten unter

<http://www-ai.cs.uni-dortmund.de/LEHRE/VORLESUNGEN/KDD/SS07/index.html>

(unter Werkzeug) herunterladen. Geben Sie für beide Teile der Aufgabe die Experimentdateien mit ab. Zur Lösung benötigen Sie die Datensätze `iris.xrff` und `irisTest.xrff`, die Sie von den Vorlesungsseiten herunterladen können

(<http://www-ai.cs.uni-dortmund.de/LEHRE/VORLESUNGEN/KDD/SS07/index.html> oder <http://www.statistik.uni-dortmund.de/download/vorlesungen/wissensentdeckung/>).

- (a) Laden Sie den Datensatz `iris.xrff` mit Hilfe des Operators `XrffExampleSource` in YALE ein. Starten Sie das Experiment, und wechseln Sie bei der Ausgabe (im Reiter `ExampleSet`) auf die Ansicht `plot view`. Mittels der Felder `x-Axis` sowie `y-Axis` können Sie Attribute des Datensatzes gegeneinander abbilden.

Wählen Sie verschiedene Attribute und visualisieren Sie diese mit verschiedenen Plottern, die Sie im Feld `Plotter` auswählen können. (Achtung: bei `Plots/Color/...` sollte `class` eingestellt werden)

Welche zwei Attribute reichen aus, um die Daten der jeweils gleichen Klassen des Datensatzes bestmöglich voneinander zu trennen? Welchen `Plotter` haben Sie benutzt, und wie sah das resultierende Bild aus, welches Sie zu Ihrer Entscheidung bewogen hat?

- (b) Für die folgende Aufgabe benötigen Sie zusätzlich zum Datensatz aus der vorangegangenen Aufgabe den Testdatensatz `irisTest.xrff`.

Erweitern Sie den ersten Versuch, indem Sie nach Einlesen des ersten Datensatzes ein Lernverfahren anwenden (z.B. den `/Learner/Trees/J48`). Anschließend müssen Sie den Testdatensatz `irisTest.xrff` mit einem `XrffExampleSource`-Operator einlesen. Daraufhin benutzen Sie den `ModelApplier`-Operator, um das gelernte Verfahren anzuwenden.

Starten Sie das Experiment, und wechseln Sie bei der Ausgabe (im Reiter `ExampleSet`) auf die Ansicht `data view`. Geben Sie bei Ihrer Abgabe die Spalte `prediction(class)` an.

Aufgabe 1.2

Die folgende Aufgabe behandelt Relationenschemata von Datenbanken.

Stellen Sie sich vor, das Prüfungsamt einer Universität hätte zur Erfassung von Prüfungsergebnissen folgendes Relationenschema aufgestellt:

Pruefungen(MatNr, StudName, Studiengang, StudAdresse, Thema, Datum,
PrueferName, Institut, Note)

Hierbei sind **MatrNr**, **StudName** und **Studiengang** die Matrikelnummer, der Name des Studierenden, sowie der belegte Studiengang. Die aktuelle Adresse des Studenten wird unter **StudAdresse** vermerkt. **PrueferName** ist der Name des Prüfers am angegebenen Institut, bei dem der Student seine Prüfung abgelegt hat. **Datum** und **Thema** geben das Datum der Prüfung und die geprüfte Fachrichtung an.

- (a) Geben Sie alle funktionalen Abhängigkeiten an. Welche Normalform liegt vor?
- (b) Welche Attribute kommen als Primärschlüssel in Frage? Gibt es Redundanzen oder liegen Anomalien vor?
- (c) Geben Sie ein geeigneteres Schema in Boyes-Codd Normalform an.