Prof. Dr. Katharina Morik, Prof. Dr. Claus Weihs
Übungen: Christian Bockermann, Julia Schiffner

Dortmund, den 8. April 2011

Übungen zur Vorlesung

Wissensentdeckung in Datenbanken

Sommersemester 2011
Blatt 0 (Präsenzübung)

Aufgabe 1

Starten Sie *RapidMiner* und importieren Sie die CSV-Datei `pilze.csv` in ihr Repository. Die CSV-Datei finden Sie unter der URL

`http://kirmes.cs.uni-dortmund.de/wid2011/data/pilze.csv`

Erzeugen Sie einen RapidMiner-Prozeß, der diesen Datensatz einliest.
Betrachten Sie den Datensatz und überlegen Sie sich:

1. Wieviele Attribute/Variablen enthält der Datensatz?
2. Welchen Typ haben die Variablen?
3. Welchen Fehler würde ein Lernverfahren auf dem Datensatz erzeugen, das immer nur die häufigste Klasse vorhersagt? (Unter der Annahme, dass der Datensatz sowohl zum Training als auch zum Testen verwendet wird.)

Aufgabe 2

Im *Samples* Repository von RapidMiner finden Sie den *Iris*-Datensatz. Bei diesem Datensatz handelt es sich um einen beliebten Datensatz zur Demonstration einfacher Lernverfahren.
Die Daten beschreiben Exemplare der *Schwertlilien*, bei denen jeweils die Längen und Breiten des Kelch- sowie des Kronblattes gemessen wurden. Anhand dieser vier Attribute sollen nun die verschiedenen Familien *iris-setosa*, *iris-versicolor* und *iris-virginica* auseinandergehalten werden.

(a) Iris Setosa

(b) Iris Versicolor

(c) Iris Virginica

1. Laden Sie den Datensatz in RapidMiner!
2. Betrachten Sie den Datensatz im Plotter. Probieren Sie die Plotter *Scatter Plot*, *Scatter Matrix* und *Distribution Plot* aus. Welche Attribute eignen Sich zum Klassifizieren der Pflanzen?

Aufgabe 3

In dieser Aufgabe liegt das Augenmerk auf den Pre-Processing Fähigkeiten von RapidMiner. Der Datensatz ist in zwei Tabellen geteilt, wie z.B. die Situation zweier Datenbanktabellen, die eine umfangreichere Entität beschreiben.
Die Daten beschreiben Situationen in denen Golf gespielt wird oder nicht – in Abhängigkeit vom Wetter. Insgesamt besteht der Datensatz aus 14 Wetter-Situationen mit dem Klassenlabel *play*.
Der Datensatz befindet sich in den Dateien:

`http://kirmes.cs.uni-dortmund.de/wid2011/data/golf1.csv`
`http://kirmes.cs.uni-dortmund.de/wid2011/data/golf2.csv`

Importieren Sie die beiden Datensätze in Ihr RapidMiner Repository.

1. Welche Attribute/Variablen enthalten die Datensätze?
2. Erzeugen Sie aus den Datensätzen mit Hilfe des *Join*-Operators einen neuen Datensatz, der die beiden importierten Datensätze vereint!
3. Das Attribut `play` enthält die Klasse der Beispiele. Wechseln Sie in die *Plot*-Ansicht. Findet sich ein Attribut, das die Klassen gut trennt?