Prof. Dr. Katharina Morik, Prof. Dr. Claus Weihs
Gero Szepannek, Martin Scholz

Dortmund, 11. Mai
Abgabe: bis 17.05., 18:00 Uhr an
scholz@ls8.cs.uni-dortmund.de

Übungen zur Vorlesung

Wissensentdeckung in Datenbanken

Sommersemester 2006

Blatt 6

Aufgabe 6.1

Bestimmen Sie die 1-freien Mengen, die 2-freien Mengen und die closed item sets in der folgenden Transaktionstabelle.

tid	A	B	C	D
1	0	1	1	0
2	1	1	0	0
3	1	0	0	1
4	0	0	1	1
5	0	1	1	0
6	0	1	1	1

tid	A	B	C	D
7	1	1	0	1
8	0	0	1	0
9	0	1	0	0
10	1	1	0	0
11	0	1	0	1
12	0	0	1	1

Aufgabe 6.2

Für diese Aufgabe benötigen Sie den Datensatz `house-votes`, den Sie wieder unter

`http://www-ai.cs.uni-dortmund.de/LEHRE/VORLESUNGEN/kdd2006/kdd.html`

beziehen können. Lösen Sie die folgenden Analyseaufgaben mit der YALE Lernumgebung und geben Sie wieder zusätzlich zu den Ergebnissen die Experiment-Dateien ab.

Hinweis: Sie benötigen sowohl die Attributbeschreibungsdatei (Endung `.att`), als auch die Daten selbst (Endung `.dat`). Zum Einlesen verwenden Sie bitte einen `ExampleSource` Operator, der die erste der beiden Dateien als Parameter `attributes` übergeben bekommt. Die zweite Datei muss im gleichen Verzeichnis wie die erste liegen.

(a) Vergleichen Sie die Maße `Accuracy` und `Precision` für die Lernverfahren `NaiveBayes`, `1-NearestNeighbor` und den Entscheidungsbaumlerner `J48` mit einer einfachen ("simple") Validierung. Verwenden Sie eine Testmenge der Größe 30%. Aus technischen Gründen verwenden Sie bitte statt `NaiveBayes` den Operator `AODE`. `1-NearestNeighbor` steht als "lazy learner" `IB1` zur Verfügung. Verwenden Sie die Standardeinstellungen. Welches Verfahren würden Sie hinsichtlich der Ergebnisse bevorzugen?

(b) Validieren Sie nun die gleichen Verfahren mit Hilfe einer 10-fachen Kreuzvalidierung (XValidation). Wie bewerten Sie die Zuverlässigkeit der Validierung (bezieht sich auf die Standardabweichung, kurzer Kommentar reicht aus)? Verändern Sie testweise den Startwert des Zufallszahlengenerators (`random_seed` im "Root" Operator). Ändern sich die Ergebnisse spürbar? Kommentieren Sie kurz den Vorteil der aufwändigeren Kreuzvalidierung gegenüber der einfachen Validierung aus der letzten Teilaufgabe.

(c) Statt `1-NearestNeighbor` können auch k Nachbarn verwendet werden. Hierfür steht der Operator `IBk` zur Verfügung. Optimieren Sie den Wert von k mit Hilfe einer 10-fachen Kreuzvalidierung ($k \leq 5$), ebenso einen beliebigen Parameter des Verfahrens `J48`, der Ihnen intuitiv erscheint. Wie stark ist der Einfluss der Parameter?

(d) Lassen Sie YALE nun einen J48-Entscheidungsbaum mit Standardeinstellungen auf allen Daten lernen, also ohne Validierung des Modells. Was wird vorhergesagt, wenn das Attribut 5 die Ausprägung "n" hat?