Prof. Dr. Claus Weihs, Prof. Dr. Katharina Morik
Julia Schiffner, Marco Stolpe

Dortmund, 18. Mai
Abgabe: bis zum 26.05. um 10:00 Uhr an
marco.stolpe@tu-dortmund.de

Übungen zur Vorlesung

Wissensentdeckung in Datenbanken

Sommersemester 2009

Blatt 6

Aufgabe 6.1 (5 Punkte)
Gegeben sei folgender fiktiver Datensatz, der beschreibt, bei welchem Wetter — beschrieben durch die Attribute *Outook*, *Temperature*, *Humidity* und *Windy* — ein bestimmtes Spiel, z. B. Golf, gespielt wird (Label *Play*). Nehmen wir an, wir wollten für diesen Datensatz einen Entscheidungsbaum konstruieren.

ID	Outlook	Temperature	Humidity	Windy	Play
1	sunny	hot	high	false	no
2	sunny	hot	high	true	no
3	overcast	hot	high	false	yes
4	rainy	mild	high	false	yes
5	rainy	cool	normal	false	yes
6	rainy	cool	normal	true	no
7	overcast	cool	normal	true	yes
8	sunny	mild	high	false	no
9	sunny	cool	normal	false	yes
10	rainy	mild	normal	false	yes
11	sunny	mild	normal	true	yes
12	overcast	mild	high	true	yes
13	overcast	hot	normal	false	yes
14	rainy	mild	high	true	no

(a) Bestimmen Sie das Attribut (ausgenommen das *ID*-Attribut!), welches die Wurzel des Entscheidungsbaums bilden würde. Führen Sie die dazu notwendigen Rechenschritte per Hand durch, so dass das Prinzip der Konstruktion von Entscheidungsbäumen ersichtlich wird.

(b) Erläutern Sie stichpunktartig den Zusammenhang zwischen der Reinheit von Knoten und dem Informationsgewinn. Warum sind die Unterknoten eines Attributs mit hohem Informationsgewinn reiner als die eines Attributs mit niedrigem Informationsgewinn?

(c) Gegeben sei ein Knoten in einem Entscheidungsbaum. Der Knoten beschreibe einmal ein nominales und einmal ein numerisches Attribut. Kann es passieren, dass das gleiche Attribut noch einmal in einem Unterbaum ab diesem Knoten verwendet wird? Begründen Sie Ihre Antwort!

(d) Berechnen Sie den Informationsgewinn für das *ID*-Attribut! Auf welches Problem bezüglich des Maßes für den Informationsgewinn stoßen Sie? Geben Sie einen Vorschlag für eine Verbesserung an.

Aufgabe 6.2 (2 Punkte)
Neben der Transformation von Daten durch Basisfunktionen zur Approximation nichtlinearer Funktionen können diese auch global durch sogenannte Interpolationspolynome angenähert werden. Zu vorgegebenen Stützpunkten (x_j, y_j), $0 \leq j \leq n$, gibt es genau ein **Interpolationspolynom** $P_n(x)$ vom Grad höchstens n, das $P(x_j) = y_j$ erfüllt. Es kann z. B. rekursiv über

$$\begin{aligned} P_0(x) &= y_0 \\ P_{k+1}(x) &= P_k(x) + (y_{k+1} - P_k(x_{k+1})) \prod_{j=0}^{k} \frac{x - x_j}{x_{k+1} - x_j} \end{aligned}$$

ermittelt werden.

(a) Bestimmen Sie das Interpolationspolynom zu den Stützpunkten $(0, 1), (1, 3)$ und $(2, 13)$.

(b) Die durch das Interpolationspolynom gegebene Funktion geht durch alle Stützpunkte. Interpretiert man diese Punkte als Trainingsbeispiele, so werden alle korrekt klassifiziert. Es scheint, als hätten wir ein perfektes Klassifikationsverfahren gefunden. Warum trifft dies jedoch nicht zu? Tip: Überlegen Sie sich, wie Funktionen aussehen, die auf Polynomen höheren Grades basieren und was dies für die Klassifikation neuer Beispiele bedeuten kann.

Aufgabe 6.3 (3 Punkte)
Lesen Sie die Spam-Daten (siehe Blatt 5) über den `ExampleSource`-Operator in RapidMiner ein. Konstruieren Sie eine 10-fache Kreuzvalidierung, innerhalb derer in den folgenden Teilaufgaben einige Lernverfahren angewendet werden sollen (tauschen Sie jeweils den Lerner aus bzw. deaktivieren Sie diejenigen, die sie gerade nicht benötigen). Geben Sie Ihre Experimente bitte als XML-Dateien mit ab!

(a) Klassifizieren Sie Spam einmal anhand des Operators `NaiveBayes` und einmal mit Hilfe eines Entscheidungsbaums `DecisionTree` und notieren Sie die jeweils resultierenden Konfusionsmatrizen (bitte mit abgeben). Welchen Lerner würden Sie für die Klassifikationsaufgabe bevorzugen? Auf welche Werte kommt es innerhalb der Konfusionsmatrix für die Klassifikation von Spam besonders an?

(b) Tauschen Sie den Lerner nun gegen den Meta-Lerner `AdaBoost` aus und machen Sie den `DecisionTree` zu einem Unteroperator von AdaBoost. Starten Sie den Prozess, der unter Umständen lange dauert. Notieren Sie auch hier die resultierende Konfusionsmatrix und vergleichen Sie sie mit den Ergebnissen aus der ersten Teilaufgabe.