Ludwig-Maximilians-Universität München
Institut für Informatik
Dr. Peer Kröger
Johannes Aßfalg, Karsten Borgwardt

Knowledge Discovery in Databases

WS 2005/06

Übungsblatt 7

Abgabe aller mit Hausaufgabe *markierten Aufgaben bis Donnerstag, 22.12.2005, 8:30 Uhr, vor der Vorlesung beim Dozenten oder im Übungsbriefkasten*

Aufgabe 7-1 Entscheidungsbaum Klassifikation
Hausaufgabe

Sie wollen die Entscheidung treffen, ob Sie an einem bestimmten Tag Tennis spielen gehen sollen. Dazu betrachten Sie die letzten 10 Tage, an denen Sie Tennis spielen waren. Aufgrund Ihres guten Gedächtnises erinnern Sie sich für jeden dieser Spieltage an

- die Aussicht aus Ihrem Fenster (sonnig, bedeckt oder regnerisch)
- die ungefähre Temperatur (heiß, mild oder kühl)
- die ungefähre Luftfeuchtigkeit (hoch oder normal)
- die Stärke des Windes (stark oder schwach)

Außerdem wissen Sie noch, ob das Tennis spielen Spaß gemacht hat oder nicht, d.h. ob Sie bei diesen Verhältnissen wieder zum Tennis spielen gehen wollen oder nicht. Die folgende Tabelle fasst Ihre Erinnerungen zusammen:

Tag	Aussicht	Temperatur	Feuchtigkeit	Wind	Tennisspielen
1	sonnig	heiß	hoch	schwach	nein
2	sonnig	heiß	hoch	stark	nein
3	bedeckt	heiß	hoch	schwach	ja
4	regnerisch	mild	hoch	schwach	ja
5	regnerisch	kühl	normal	schwach	ja
6	regnerisch	kühl	normal	stark	nein
7	bedeckt	kühl	normal	stark	ja
8	sonnig	mild	hoch	schwach	nein
9	sonnig	kühl	normal	schwach	ja
10	regnerisch	mild	normal	schwach	ja

(a) Konstruieren Sie anhand dieser Trainingsdaten einen Entscheidungsbaum. Benutzen Sie beim Split den Gini-Index als Maß für die Unreinheit. Erzeugen Sie dabei für jeden Attributwert einen eigenen Ast. Der Entscheidungsbaum soll terminieren, wenn alle Instanzen im Blatt die gleiche Klasse haben. Die Anwendung eines Pruning-Algorithmus ist nicht erforderlich!

(b) Entscheiden Sie mit Hilfe Ihres Entscheidungsbaumes, ob Sie an den folgenden Tagen zum Tennis spielen gehen wollen:
Tag A: sonnig, heiß, normal, stark
Tag B: regnerisch, mild, hoch, schwach
Tag C: sonnig, kühl, hoch, stark

Aufgabe 7-2 Informationsgewinn
Hausaufgabe

In dieser Aufgabe wollen wir den Informationsgewinn (information gain) genauer untersuchen und verstehen.

Im folgenden betrachten wir die Menge T von n Trainingsobjekten, mit den Attributen $A_1, \ldots, A_a$ und den k Klassen c_1 bis c_k.

Sei $\{T_i^A \mid i \in \{1, \ldots, m_A\}\}$ die disjunkte, vollständige Partitionierung von T, die durch einen Split auf dem Attribut A erzeugt wird (wobei m_A die Anzahl von Ausprägungen von A ist).

(a) *Gleichverteilung*
Berechnen Sie *entropie*(T), *entropie*(T_i) für $i \in \{1 \ldots m_A\}$ sowie *informationsgewinn*(T, A) unter der Annahme, dass die Klassenzugehörigkeiten in T gleichverteilt und unabhängig von den Ausprägungen von A sind. Interpretieren Sie Ihr Ergebnis!

(b) *Zusätzliche gleichverteilte Ausprägung*
Wir wollen untersuchen, inwieweit die Anzahl der Ausprägungen den Informationsgewinn beeinflusst. Betrachten wir dazu ein beliebiges Attribut A mit seinen m_A Ausprägungen. Wie ändert sich der *informationsgewinn*(T, A) wenn wir A durch A' mit $m_{A'} = m_A + 1$ Ausprägungen ersetzen, wobei die relativen Häufigkeiten in den Ausprägungen 1 bis m_A von A' identisch zu A sind und in der Ausprägung $m_{A'}$ eine Gleichverteilung der Klassen herrscht? Interpretieren Sie Ihr Ergebnis!

(c) *Attribute mit sehr vielen Ausprägungen*
Sei A ein Attribut mit zufälligen, nicht mit der Klasse der Objekte korrelierten Werten. Weiterhin verfüge A über so viele Ausprägungen, dass keine zwei Objekte der Trainingsmenge zu derselben Ausprägung in A gehören. Was geschieht in dieser Situation beim Aufbau das Entscheidungsbaumes? Was ist daran problematisch?

Aufgabe 7-3 Support Vector Machines

Angenommen, eine Support Vector Machine minimiert beim Lernen der Entscheidungsfunktion lediglich die Zahl der falsch klassifizierten Trainingsobjekte. Welches Problem kann sich daraus potentiell ergeben? Wie lässt sich dieses Problem beheben?