Ludwig-Maximilians-Universität München
Institut für Informatik
PD Dr. Peer Kröger, Dr. Arthur Zimek

Knowledge Discovery in Databases

WS 2010/11

Übungsblatt 6: Entscheidungsbäume, Neuronale Netze

Aufgabe 6-1 *Entscheidungsbäume*

Sie wollen die Risikoklasse eines Autofahrers anhand der folgenden Merkmale vorhersagen:

- Zeit seit Bestehen der Fahrprüfung(1-2 Jahre, 2-7 Jahre, $>$7 Jahre)
- Geschlecht (männlich, weiblich)
- Wohnort(Stadt, Land)

Für Ihre Analyse stehen Ihnen folgende manuell eingeteilte Testbeispiele zu Verfügung:

Person	Zeit seit der Fahrprüfung	Geschlecht	Wohnort	Risikoklasse
1	1-2	m	Stadt	niedrig
2	2-7	m	Land	hoch
3	$>$7	w	Land	niedrig
4	1-2	w	Land	hoch
5	$>$7	m	Land	hoch
6	1-2	m	Land	hoch
7	2-7	w	Stadt	niedrig
8	2-7	m	Stadt	niedrig

(a) Konstruieren Sie anhand dieser Trainingsdaten einen Entscheidungsbaum. Benutzen Sie beim Split den Informationsgewinn 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) Wenden Sie Ihren Entscheidungsbaum auf folgende Autofahrer an:
Person A: 1-2, w, Land
Person B: 2-7, m , Stadt
Person C: 1-2, w, Stadt

Aufgabe 6-2 *Informationsgewinn*

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 $entropie(T)$, $entropie(T_i^A)$ 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 und ein Attribut A' mit $m_{A'} = m_A + 1$ Ausprägungen, 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. Wie unterscheidet sich der *informationsgewinn*(T, A) vom *informationsgewinn*(T, A')? 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 6-3 *Lineare Separierbarkeit*

Geben Sie für die folgenden Booleschen Funktionen an, ob das entsprechende Problem linear separierbar ist.

(a) $A \wedge B \wedge C$

(b) $A \vee B$

(c) $(A \vee B) \wedge (A \vee C)$

(d) $\neg A \wedge B$

Aufgabe 6-4 *Perceptron*

Zeigen Sie, wie das Peceptron-Modell (kein hidden layer!) verwendet werden kann, um über zwei Boolesche Variablen $x_1, x_2 \in \{0, 1\}$ die UND bzw. die ODER Funktion ($x_1 \wedge x_2$ bzw. $x_1 \vee x_2$) zu repräsentieren.