Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Hans-Peter Kriegel
Eshref Januzaj, Peer Kröger, Matthias Schubert

Knowledge Discovery in Databases

WS 2002/03

Übungsblatt 7

Abgabe aller mit Hausaufgabe *markierten Aufgaben bis 12.11.2002, 11:00 Uhr*

Besprechung: 16.12. – 18.12.02

Aufgabe 7-1 Hausaufgabe
Entscheidungsbaum Klassifikation

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 Hausaufgabe
Support Vector Machines

In der Vorlesung wurden Support Vector Machines zur Klassifikation eingeführt. In dieser Aufgabe soll zur Erläuterung des vorgestellten Verfahrens der minimale Fall besprochen werden, in dem für jede Klasse nur ein Vektor gegeben ist. Dies impliziert zwangläufig, dass die Klassen aufgrund der Trainingsmenge linear separierbar sind, d.h. kein Kernel und kein Soft Margin. Als Trainingsmenge sollen uns die beiden Vektoren $(1, 1)$ für Klasse A ($y = -1$) und $(2, 2)$ für Klasse B ($y = 1$) dienen. Das verwendete Skalarprodukt sei das kanonische Skalarprodukt (vgl. Bsp in der Vorlesung).

(a) Formulieren Sie das Problem zunächst als duales Optimierungsproblem mit Lagrange Multiplikatoren. Bestimmen Sie jetzt durch analytische Lösung die Werte der Lagrange Multiplikatoren.

(b) Zur Berechnung der Maximum Margin Hyperplane bestimmen wir jetzt den Normalenvektor $\vec{w}$. Der doppelte Betrag dieses Vektors $\vec{w}$ stellt dabei die Breite des Randes (Margin) dar.
Benutzen Sie zur Bestimmung folgende Formel:

$$\vec{w} = \sum_{i=1}^{n} \alpha_i \cdot y_i \cdot \vec{x}_i$$

(c) Nachdem Sie den Normalenvektor $\vec{w}$ bestimmt haben, können wir jetzt daraus noch den Skalar b berechnen, um die Lage der Hyperebene $H(\vec{w}, b)$ nun endgültig festzulegen. Die Formel zur Berechnung ist:

$$b = -\frac{\min_{i,y_i=-1} \langle \vec{w}, \vec{x}_i \rangle + \max_{i,y_i=1} \langle \vec{w}, \vec{x}_i \rangle}{2}$$

(d) Nachdem Sie die trennende Hyperebene jetzt festgelegt haben, bestimmen Sie jetzt die Klasse der beiden Vektoren: $(3, 5)$ und $(0, 1)$. Benutzen Sie dazu entweder die Entscheidungsregel des Primären OPs oder die des Dualen OPs. Welche Entscheidungsregel sollte bei der Verwendung eines Kernels verwendet werden?