Ludwig-Maximilians-Universität München
Institut für Informatik
Dr. Matthias Schubert

Knowledge Discovery in Databases

WS 2008/09

Übungsblatt 5

Aufgabe 5-1 Entscheidungsbäume
Hausaufgabe

Sie wollen die Risikoklasse einer(s) Autofahrerin(s) 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 Informationgewinn 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 5-2 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?

Aufgabe 5-3 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 zwangsläufig, dass die Klassen aufgrund der Trainingsmenge linear separierbar sind, d.h. wir können mit einem Hard Margin im Eingaberaum arbeiten. 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 $\vec{w}$. Das Inverse des Betrags 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:

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

(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.

Aufgabe 5-4 Kernel-Funktionen

Wie in der Vorlesung erklärt, zeichnet sich eine Kernel-Funktion ("Kernel") durch positive (Semi-)Definitheit aus. Eine Matrix A ist positiv definit, falls ihre Eigenwerte nichtnegativ sind, oder alternativ formuliert, falls für all $x \in \mathbb{R}^d$ gilt: $x^\top \cdot A \cdot x \geq 0$

Zeigen Sie, dass folgende Funktionen Kernels sind, falls x und $\hat{x}$ Vektoren im $\mathbb{R}^d$ sind:

(a) $k(x, \hat{x}) = 1$

(b) $k(x, \hat{x}) = 3 * x^\top \cdot \hat{x}$

(c) $k(x, \hat{x}) = 3 * x^\top \cdot \hat{x} + 5$