Ludwig-Maximilians-Universität München
Institut für Informatik
Dr. Peer Kröger, Dr. Matthias Schubert
Markus Bundschus

Knowledge Discovery in Databases

WS 2007/08

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 Lineare Regression

Das Gehalt einer Person hängt von den Jahren ab, in denen die Person ihren Beruf ausgeübt hat. Um diesen Zusammenhang genauer zu untersuchen, kann man ein lineares Regressionmodel lernen. Als Trainingsmenge stehen uns die Jahre an Berufserfahrung und die Gehälter folgender Personen zur Verfügung.

Erfahrung in Jahren	Gehalt in (1000$)
3	30
8	57
9	64
13	72
3	36
6	43
11	59
21	90
1	20
16	83

(a) Berechnen Sie eine Regressionsgerade, die dazu dienen soll, das voraussichtliche Gehalt auf Basis der Berufserfahrung abzuschätzen. Bestimmen Sie hierzu die Gerade, die den quadratischen Fehler minimiert.

(b) Bestimmen Sie den quadratischen Fehler der berechneten Gerade, um abzuschätzen, wie gut die Regressiongerade den Zusammenhang erklärt.

(c) Berechnen Sie mit Hilfe Ihrer Regressionsgerade das vorraussichtliche Gehalt für Personen mit den folgenden Jahren an Beruferfahrung:
Person A: 20
Person B: 8
Person C: 11

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