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

Knowledge Discovery in Databases

WS 2005/06

Übungsblatt 9

Abgabe aller mit Hausaufgabe *markierten Aufgaben bis Donnerstag, 19.01.2006, 8:30 Uhr, vor der Vorlesung beim Dozenten oder im Übungsbriefkasten*

Aufgabe 9-1 Lineare Regression
Hausaufgabe

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 9-2 BIRCH
Hausaufgabe

Gegeben seien folgende 1-dimensionale Punktdaten:
1, 2, 5, 10, 4, 7, 15, 9, 3

Im folgenden soll die Euklidische Distanz (L_2-Norm) verwendet werden.

(a) Bauen Sie einen CF-Baum mit $B = 2$ und $L = 3$ auf, indem Sie die Punkte der Reihe nach einfügen. Der Schwellwert T ist dabei wie folgt definiert: "Alle Einträge in einem Blattknoten haben eine Euklidische Distanz ≤ 2 zum Centroiden des Blattknotens".
Zeichnen Sie den CF-Baum nach jeder Einfügung.

(b) Warum ist das Ergebnis abhängig von der Reihenfolge?