Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Peer Kröger
Anna Beer, Florian Richter

Knowledge Discovery in Databases

WS 2017/18

Übungsblatt 11: Classification II

Besprechung: 01. und 02.02.2018

Aufgabe 11-1 Klassifikation vs. Clusteranalyse

Bei welchen der folgenden Aufgabenstellungen handelt es sich um Klassifikationsprobleme, bei welchen um Clusteranalyse?

(a) Emails im Posteingang sollen nach Spam und nicht Spam sortiert werden.

(b) Eine Datenbank von Nutzern soll nach ihrem Kaufverhalten gruppiert werden.

(c) In einem Supermarkt sollen Produkte, die oft zusammen gekauft werden, in einem Regal nebeneinander platziert werden, um so die Verkäufe zu steigern.

(d) Das Spam-Vorkommen soll analysiert werden, um zu erkennen, ob es darin unterschiedliche Gruppen / Typen von Werbung gibt.

(e) Basierend auf der DNA einer Person soll vorhergesagt werden, ob sie in den nächsten 10 Jahren an Diabetes leiden wird.

(f) Daten von Patienten mit Herzkrankheiten sollen analysiert werden, ob es darin Gruppen gibt für die spezielle Therapien besser funktionieren als für andere.

(g) Einteilung von Webseiten in Kategorien wie “Sport”, “Wirtschaft”, “Unterhaltung”.

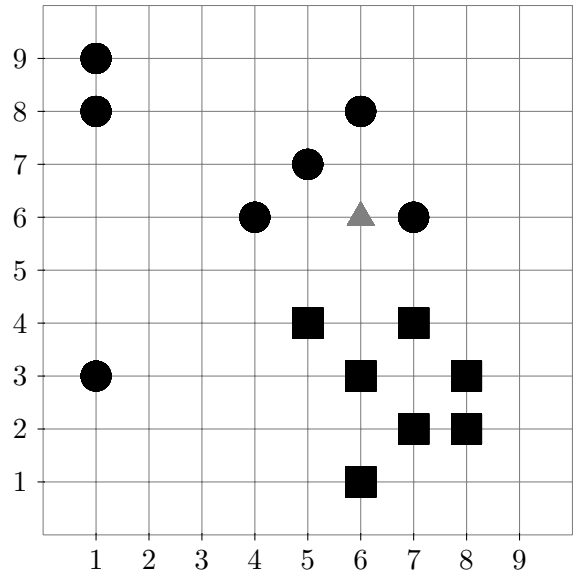
Aufgabe 11-2 Nächste-Nachbarn-Klassifikation

Die 2D Featurevektoren in der nachfolgenden Abbildung seien mit zwei unterschiedlichen Klassenlabeln (Quadrate und Kreise) versehen. Klassifizieren Sie den Punkt (6,6) — im Bild dargestellt durch ein Dreieck — mit einem k-nächsten Nachbarn Klassifikator. Distanzfunktion soll die L_1-Norm (Manhattan-Distanz) sein. Verwenden Sie dabei als Entscheidungsregel die ungewichtete Anzahl der einzelnen Klassen in der k-nächsten Nachbarn Menge, d.h. der Punkt wird der Klasse zugewiesen, die die meisten k-nächsten Nachbarn stellt. Führen Sie die Klassifikation für folgende Werte für k durch und vergleichen Sie die Ergebnisse mit ihrem eigenen intuitiven Ergebnis:

(a) $k = 4$

(b) $k = 7$

(c) $k = 10$



Aufgabe 11-3 Nächste-Nachbarn-Klassifikation

Geben Sie eine Punktmenge an, bestehend aus mindestens vier 2-dimensionalen Punkten, so dass die Nächste-Nachbarn-Klassifikation ($k = 1$) auf diesen Punkten nur Fehlklassifikationen liefert! Als Distanzfunktion sei die euklidische Distanz gegeben.