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

Knowledge Discovery in Databases

WS 2009/10

Übungsblatt 4

Aufgabe 4-1 Bayes Klassifikation
Hausaufgabe

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) Berechnen Sie die *a priori* Wahrscheinlichkeiten für die beiden Klassen `Tennis spielen = ja` und `Tennis spielen = nein` (auf den Trainingsdaten)!

(b) Berechnen Sie für alle möglichen Attributswerte die Wahrscheinlichkeit, dass aufgrund des Wertes und der Trainingsdaten eine bestimmte Klasse gewählt wird!

(c) Entscheiden Sie, ob Sie bei den folgenden Wetterbedingungen Tennis spielen gehen oder nicht! Verwenden Sie dazu den naiven Bayes-Klassifikator.

Tag A: sonnig, heiß, normal, stark
Tag B: regnerisch, mild, hoch, schwach
Tag C: sonnig, kühl, hoch, stark

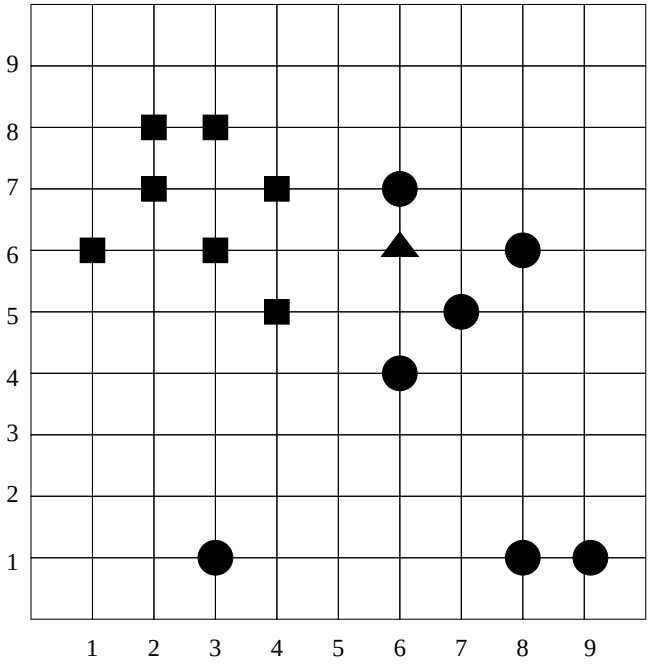
Aufgabe 4-2 Nächste-Nachbarn-Klassifikation

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 wieder 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 4-3 Entscheidungsbäume

Betrachten sie erneut dis Tennis-Daten aus der obigen Aufgabe . In der Vorlesung haben Sie Entscheidungsbäume als weitere wichtige Klassifikationmethode kennengelernt. Im folgenden sollen Sie dieses Verfahren auf das obige Beispiel anwenden. Gehen hierzu folgendermaßen vor:

(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