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

München, 13.06.2016

Managing Massive Multiplayer Online Games

SS 2016

Übungsblatt 7: Knowledge Discovery and Data Mining II

Besprechung: 16.06.2016

Aufgabe 7-1 *Lineare Regression*

Der Mietpreis y_i einer Wohnung i hängt stark von der Quadratmeterzahl x_i ab. Andere Faktoren haben natürlich auch Einfluss, aber vereinfachend kann der Zusammenhang zwischen Mietpreis und Größe über ein lineares Regressionsmodell dargestellt werden, d.h.

$$y_i = w_0 + w_1 x_i$$

Als Trainingsmenge stehen folgende Daten zur Verfügung.

Größe in m^2	Kaltmiete in Euro
30	600
60	966
100	1640
55	992
93	1790
195	2925
21	469
61	840
62	1400

(a) Berechnen Sie die Regressionsgerade, die den mittleren quadratischen Fehler MSE zwischen den Mietpreis-Vorhersagen $\hat{y}_i$ und den tatsächlichen Preisen y_i minimiert

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y_i} - y_i)^2$$

(b) Bestimmen Sie den quadratischen Fehler, um abzuschätzen wie gut das Modell den Zusammenhang beschreibt.

(c) Berechnen Sie mithilfe der Regressionsgeraden den vorraussichtlichen Preis für eine Wohnung mit 120 Quadratmetern Fläche.

Aufgabe 7-2 *Frequent Itemset Mining*

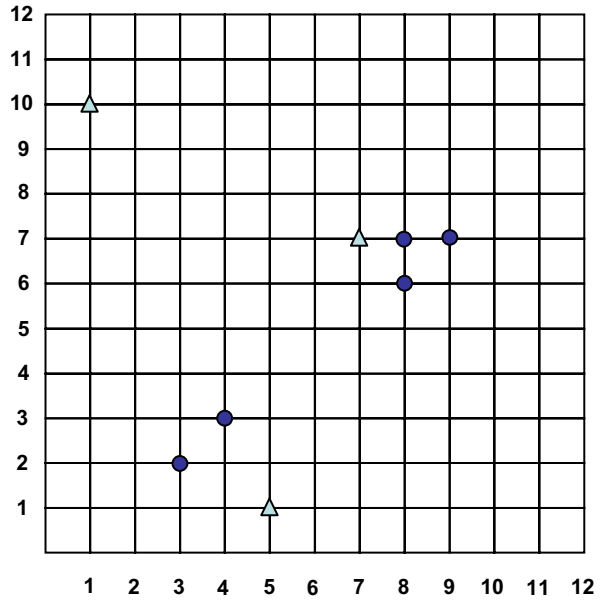
Gegeben sei ein Multiplayer Battle Arena Game, bei dem die Spieler jeweils eine von neun Heldenklassen wählen können. Es gebe folgende Klassen: Druide (D), Hexer (H), Jäger (J), Krieger (K), Mönch (M), Priester (P), Ritter (R), Schurke (S), Zauberer (Z). Ein Team besteht dabei aus 5 Spielern. Gegeben sei nun folgende Datenbank aus Teamzusammensetzungen:

Gruppe	Klassen
1	D H K P Z
2	D J K P S
3	D J M P R
4	H K M P Z
5	D J M R Z
6	D M P S Z
7	D H J M P
8	D J M P Z
9	K M P S Z
10	D J K P Z
11	J K M P Z
12	D K M P R

Bestimmen Sie alle Teilteams die mindestens 5 mal vorkommen. Verwenden Sie hierzu den Apriori Algorithmus zur Bestimmung häufiger Itemsets.

Aufgabe 7-3 *Clustering mit Varianzminimierung*

Gegeben sei erneut folgender Datensatz mit 8 Punkten (z.B. zweidimensionale Featurevektoren).



Im folgenden sollen vollständige Partitionierungen des Datensatzes in $k = 2$ Cluster berechnet werden. Als Distanzfunktion zwischen den Punkten soll dabei die Manhattan-Distanz (L_1-Norm) verwendet werden.

(a) Erzeugen Sie eine Partitionierung in $k = 2$ Cluster mit dem einfachen Verfahren "Clustering durch Varianz Minimierung". Die initiale Partitionierung der Daten ist durch die Dreiecke und Punkte gegeben (die Dreiecke bilden einen initialen Cluster, ebenso die Punkte). Beschreiben Sie jede Aktion des Algorithmus. Zeichnen Sie nach jedem Schritt die Zentroiden ein und markieren Sie die Punkte anhand ihrer Clusterzugehörigkeit. Denken Sie daran, bei der Zuordnung zu den Zentroiden die L_1-Norm zu verwenden.

(b) Zeigen Sie, dass dieses Verfahren von der Wahl des initialen Partitionierung abhängt.