Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Matthias Schubert
Felix Borutta

Munich, June 12, 2019

Managing Massive Multiplayer Online Games

SS 2019

Exercise Sheet 7: Knowledge Discovery and Data Mining II

The assignments are due June 19, 2019

Assignment 7-1 *Linear Regression*

The rent y_i of an apartment i depends on its size x_i. There are other influences, too, but the relation between rent and size can be simplified and represented by a linear regression model, i.e.:

$$y_i = w_0 + w_1 x_i$$

As training set the following data is available:

area in m^2	cold rent in €
30	600
60	966
100	1640
55	992
93	1790
195	2925
21	469
61	840
62	1400

(a) Calculate the regression line which minimizes the mean square error (MSE) between the predicted rent $\hat{y}_i$ and the actual rent y_i

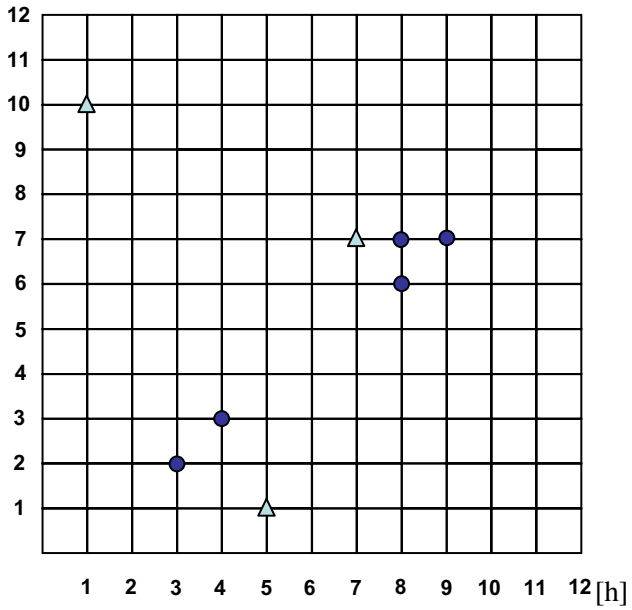
$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y_i} - y_i)^2$$

(b) Compute the square error to estimate how good the model describes the relation.

(c) Calculate the expected rent for a flat with $120m^2$ using the regression line.

Assignment 7-2 *Clustering with variance minimization*

The following data set with 8 points (e.g. two-dimensionally feature vectors) is given.



Partition of the dataset into $k = 2$ clusters. As distance function the Manhattan distance (L_1 norm) should be used.

(a) Partition the dataset into $k = 2$ clusters using the "clustering with variance minimization" procedure. The initial partitioning of the data points is given by the markers (triangles and circles). Describe every action of the algorithm.

(b) Show that the result depends on the initial partitioning.

Assignment 7-3 *Suffix Trees*

The alphabet $A = \{A, B, C, D, N\}$ is given.

(a) Insert the sequence $G_1 = \{B, A, N, A, N, A\}$ into an empty suffix tree ST

(b) Additionally insert the sequence $G_2 = \{C, A, N, A, D, A\}$ into ST.

(c) Find the subsequence $S_1 = \{N, A, N, A\}$. Which sequence contains S_1?

(d) Which is the longest common subsequence of G_1 and G_2?

(e) Which extension would be necessary to support finding the most frequent subsequence of length n (or longer)?

Assignment 7-4 *Levenshtein Distance*

Compute the Levenshtein Distance between the sequences $BANANA$ and $CANADA$.