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

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

Let $w = (w_0, w_1)^T$ be the weight vector, y be the ground truth vector for the training data (rent) and x be the input vector (area):

$$\begin{pmatrix} 1 & 30 \\ 1 & 60 \\ \vdots & \\ 1 & 62 \end{pmatrix},$$

i.e., a 9×2 shaped matrix with the ones being used for multiplication with w_0 such that we get $y = 1 \cdot w_0 + x \cdot w_1$. This way we can reformulate the error function as: $f(w) = (y - Xw)^T(y - Xw)$. The derivative of this is:

$$\frac{\partial f(w)}{w} = -2X^T(y - Xw)$$

Note: $\frac{\partial Ax}{x} = A^T$. If we resolve for w:

$$w = (X^T X)^{-1} X^T y$$

Plugging in the training data retrieves

$$\begin{aligned} w &= \left(\begin{pmatrix} 1 & 1 & \cdots & 1 \\ 30 & 60 & \cdots & 62 \end{pmatrix} \cdot \begin{pmatrix} 1 & 30 \\ 1 & 60 \\ \vdots & \\ 1 & 62 \end{pmatrix} \right)^{-1} \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 30 & 60 & \cdots & 62 \end{pmatrix} \cdot \begin{pmatrix} 600 \\ 966 \\ \vdots \\ 1400 \end{pmatrix} \\ &= \begin{pmatrix} 3.77018108e-01 & -3.53495269e-03 \\ -3.53495269e-03 & 4.69934627e-05 \end{pmatrix} \begin{pmatrix} 11612 \\ 1177304 \end{pmatrix} \\ &= \begin{pmatrix} 216.22032624 \\ 14.27772092 \end{pmatrix}, \end{aligned}$$

i.e., we finally get $w_0 = 216.22$ and $w_1 = 14.28$.

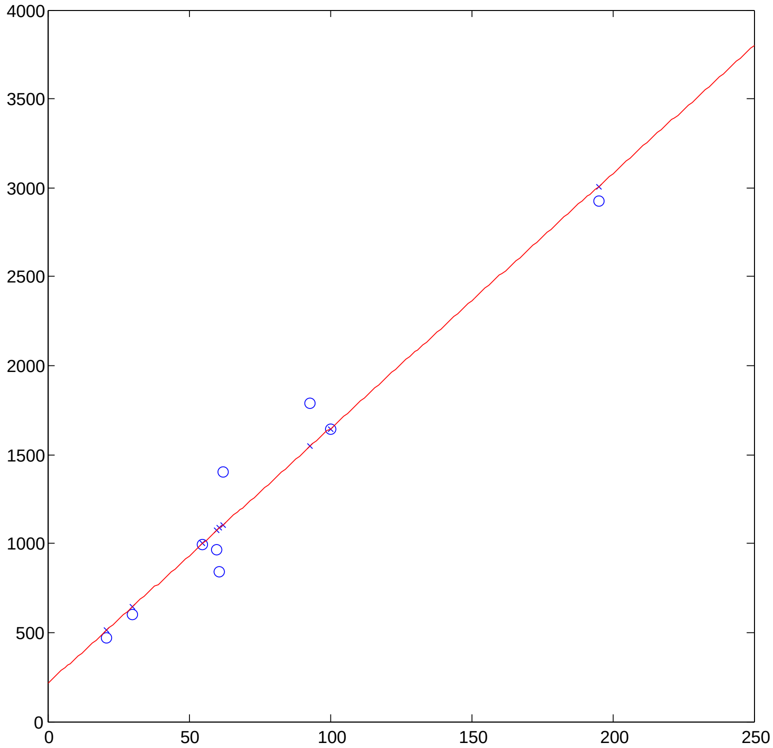
You may want to check this documentation for more details: `https://ml-cheatsheet.readthedocs.io/en/latest/linear_regression.html`

(b) Compute the square error to estimate how good the model describes the relation.
We can calculate the training error by first calculating the values that our model would predict, i.e., $\hat{\mathbf{y}}_i = Xw$, and subsequently computing the sum of the squared errors as

$$\sum_{i=0}^{9} (\hat{y}_i - y_i)^2 = 233737.62.$$

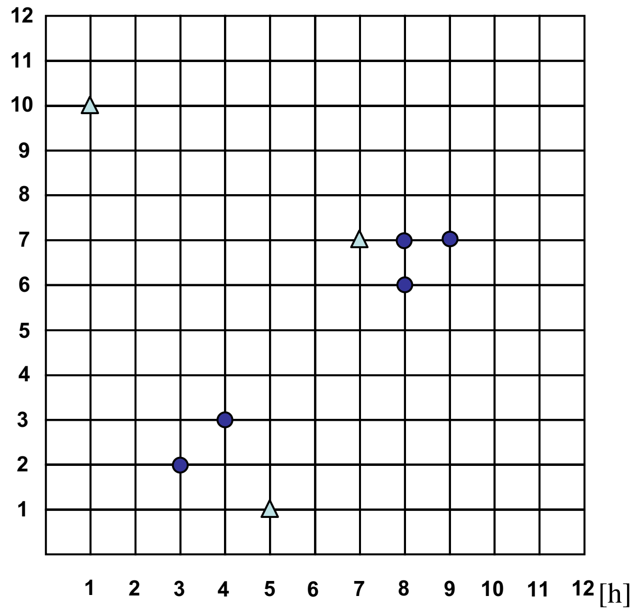
The MSE (dividing the squared error by the number of training instances) is: 25970.85. We could also calculate other errors (which might be more interpretable) like the MAE: 121.0.

(c) Calculate the expected rent for a flat with $120m^2$ using the regression line.
To make a prediction with our model, we just need to plug in the observed value into our regression function. $\hat{y} = w_0 + w_1 120 = 1929.55$.



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.

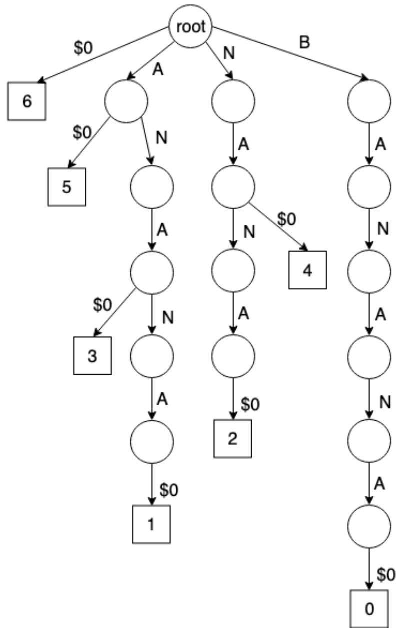
(c) Find the subsequence $S_1 = \{N, A, N, A\}$. Which sequence contains S_1?

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

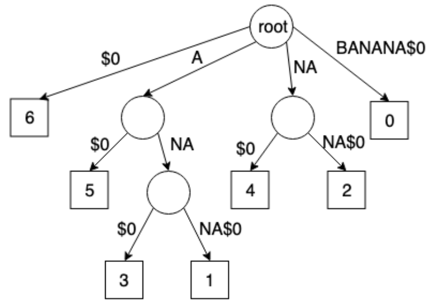
(e) Which extension would be necessary to support finding the most frequent subsequence of length n (or longer)?

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

uncompressed

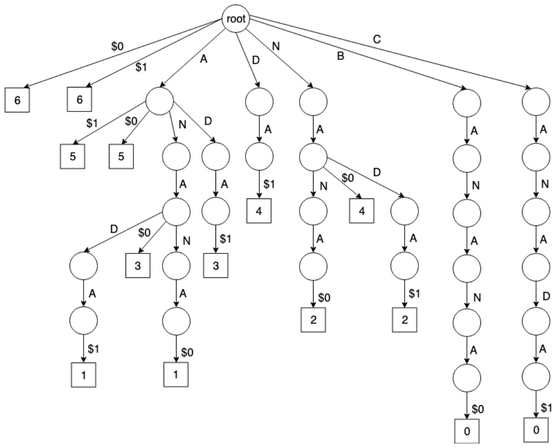


compressed

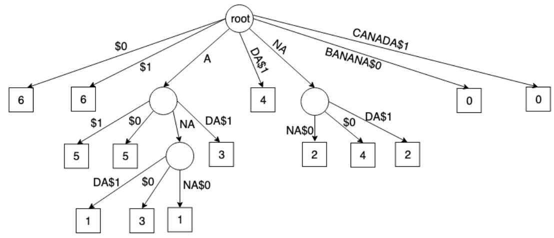


(b) Additionally insert the sequence $G_2 = \{C, A, N, A, D, A\}$ into ST.

uncompressed



compressed



(c) Find the subsequence $S_1 = \{N, A, N, A\}$. Which sequence contains S_1?
Start from the root and go down the branch which corresponds to the given sequence S_1 until you reach the last literal/object o (in this case $o = \text{'}A'$) of the sequence. Then simply check to which sequences the leaves that are in the subtree rooted in o belong.

(d) Which is the longest common subsequence of G_1 and G_2?
Find the lowest inner node n which has leaves of G_1 and G_2 in his subtree. The longest subsequence is the sequence corresponding to the path from the root node to node n. In our case, the longest subsequence is 'ANA' since the node having the edge corresponding to the latter 'A' as incoming edge is the lowest node with leaves of G_1 and G_2 in its subtree.

(e) Which extension would be necessary to support finding the most frequent subsequence of length n (or longer)?
We'd need to store the number of leaves for each inner node of the tree. This way we could find the most frequent subsequence of length n or longer by doing look-ups on the inner nodes on level n or lower.

Assignment 7-4 *Levenshtein Distance*

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

Mathematically, the Levenshtein distance between two sequences a, b (of length $|a|$ and $|b|$ respectively) is given by $lev_{a,b}(|a|, |b|)$ where

$$lev_{a,b}(i,j) = \begin{cases} max(i,j) & \text{if } min(i,j) = 0\text{, i.e., for the first column resp. row} \\ min \begin{cases} lev_{a,b}(i-1,j)+1 \\ lev_{a,b}(i,j-1)+1 \\ lev_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \end{cases} & otherwise. \end{cases}$$

where $1_{(a_i \neq b_j)}$ is the indicator function equal to 0 when $a_i = b_j$ and equal to 1 otherwise, and $lev_{a,b}(i,j)$ is the distance between the first i characters of a and the first j characters of b.

	-	B	A	N	A	N	A
-	0	1	2	3	4	5	6
C	1	1	2	3	4	5	6
A	2	2	1	2	3	4	5
N	3	3	2	1	2	3	4
A	4	4	3	2	1	2	3
D	5	5	4	3	2	2	3
A	6	6	5	4	3	3	2

Given that $a =$ BANANA with $|a| = 6$ and $b =$ CANADA with $|b| = 6$, the Levenshtein distance between the sequences 'BANANA' and 'CANADA' is $lev_{\text{BANANA,CANADA}}(6,6) = 2$.