M. Strohmaier, P. Singer, F. Lemmerich, E. Vasilev
GESIS - Leibniz Institute for the Social Sciences – Computational Social Science Department
Universität Koblenz-Landau

16.12.2016

2. Home Assignment "Machine Learning and Data Mining"

Winter semester 2016 / 2017

You should work on the assignment in groups of 2–3 participants. Send your solution (as .pdf) by email to sent to `vasilev@uni-koblenz.de`, `florian.lemmerich@gesis.org`, and `philipp.singer@gesis.org` until **Tuesday, 22nd of November, 23:59h**. Use `[ML-Assignment]` as the email subject and do not forget to denote the name of all contributing students in the pdf as well as in the email. The deadline is strict!

1 Clustering vs Classification

1. What is the main difference between *clustering* and *classification*?
2. What is easier to evaluate, clustering or classification?

2 k-Nearest Neighbor

Given are data points $x_1, \ldots, x_{11}$:

$$x_1 = \begin{pmatrix} 3 \\ 2 \end{pmatrix}, x_2 = \begin{pmatrix} 4 \\ 3 \end{pmatrix}, x_3 = \begin{pmatrix} 5 \\ 2 \end{pmatrix}, x_4 = \begin{pmatrix} 3 \\ 6 \end{pmatrix}, x_5 = \begin{pmatrix} 8 \\ 5 \end{pmatrix}, x_6 = \begin{pmatrix} 8 \\ 6 \end{pmatrix},$$
$$x_7 = \begin{pmatrix} 5 \\ 5 \end{pmatrix}, x_8 = \begin{pmatrix} 7 \\ 3 \end{pmatrix}, x_9 = \begin{pmatrix} 4 \\ 6 \end{pmatrix}, x_{10} = \begin{pmatrix} 3 \\ 7 \end{pmatrix}, \text{ and } x_{11} = \begin{pmatrix} 6 \\ 4 \end{pmatrix}$$

The data points $x_1, \ldots, x_6$ belong to class c_1, the data points $x_7, \ldots, x_{10}$ belong to class c_2. x_{11} has not yet been classified.

1. Classify the data point x_{11} with the k-nearest-neighbor approach and $k = 1$. Use the Euclidean distance as the distance metric.

2. Classify the data point x_{11} with the k-nearest-neighbor approach and $k = 5$. Use the Euclidean distance as the distance metric.

3. Classify the data point x_{11} with the k-nearest-neighbor approach and $k = 5$. Use the Euclidean distance as distance metric. As weights, use the inverted distance to the point to be classified, i.e., $\frac{1}{d(x_n,x_c)}$, with x_n is the neighbor, and x_c is the point to be classified.

3 Naive Bayes

Given is the following dataset:

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D_1	Sunny	Hot	High	Weak	No
D_2	Sunny	Hot	High	Strong	No
D_3	Overcast	Hot	High	Weak	Yes
D_4	Rain	Mild	High	Weak	Yes
D_5	Rain	Cool	Normal	Weak	Yes
D_6	Rain	Cool	Normal	Strong	No
D_7	Overcast	Cool	Normal	Strong	Yes
D_8	Sunny	Mild	High	Weak	No
D_9	Sunny	Cool	Normal	Weak	Yes
D_{10}	Rain	Mild	Normal	Weak	Yes
D_{11}	Sunny	Mild	Normal	Strong	Yes
D_{12}	Overcast	Mild	High	Strong	Yes
D_{13}	Overcast	Hot	Normal	Weak	Yes
D_{14}	Rain	Mild	High	Strong	No

1. Train a naive-Bayes classificator with this data and use it to classify the new data point D_{15} = *(O=Overcast, T=Cool, H=High, W=Strong)*!

2. Use your classifier to classify the data point D_{16} = *(T=Cool, H=High, W=Strong)*!
   The value for the attribute *Outlook* was not recorded for this data point.

3. Compare (*short* explanation!) the results for subtasks 1 and 2!

4 Decision Trees

A hospital wants to provide support for his physicians. For that purpose they collected data on healthy and ill patients. Decision makers read in a magazine that decision trees allow to create a model that can be used to simulate decision of a doctor.

Patient Nr.	Heart Rate	Blood Pressure	Class
1	irregular	Normal	Ill
2	regular	Normal	Healthy
3	irregular	Abnormal	Ill
4	irregular	Normal	Ill
5	regular	Normal	Healthy
6	regular	Abnormal	Ill
7	regular	Normal	Healthy
8	regular	Normal	Healthy

1. Compute a decision tree for this dataset (using information gain as a split criterion, without using the patient id) and sketch it graphically. If you want to do compute this with pen and paper only (no computer or calculator), you can use the following approximations for the logarithms:

x	$\frac{1}{5}$	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{4}{5}$
$\log_2(x)$	$-\frac{23}{10}$	$-\frac{8}{5}$	$-\frac{3}{5}$	$-\frac{3}{10}$

2. What would be the information gain for the split with the attribute *Patient Nr.* in the overall dataset? Compare the result with the information gain of other attributes. What problems occur when a decision tree uses the attribute *Patient Nr.*?
3. State a simple, but valid strategy to parallelize decision tree learning?
4. Give an example dataset such that the approach of greedily learning decision trees with information gain does *not* lead to an optimum tree with respect to the tree complexity (measured as number of nodes in the tree). Assume that the decision tree should correctly classify all training examples (no pruning).