Mittuniversitetet
MID SWEDEN UNIVERSITY

DT044A Datamining

Intro to lab 2: Weka

Mittuniversitetet
MID SWEDEN UNIVERSITY

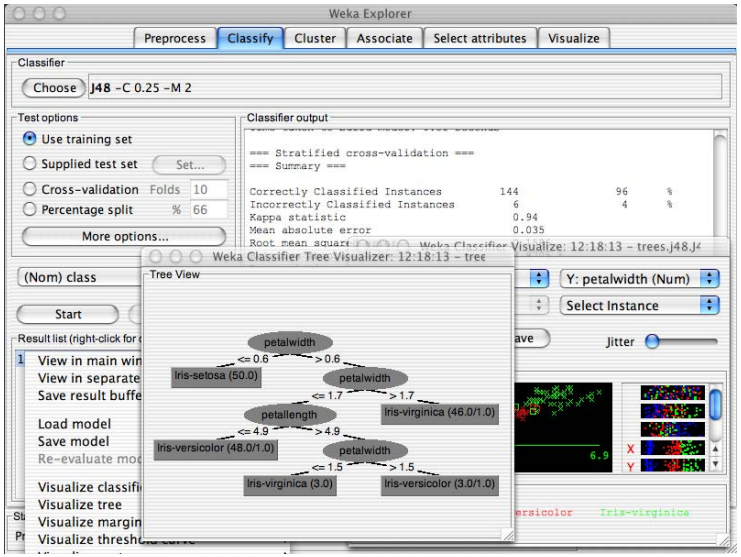
Widespread open source software tools for datamining

- R scripting language
- WEKA (program library in Java which also has a graphical user interface)
- Rapidminer (based on graphical programming – open source code until version 5.3)
- Pandas and Orange (libraries for the Python language)

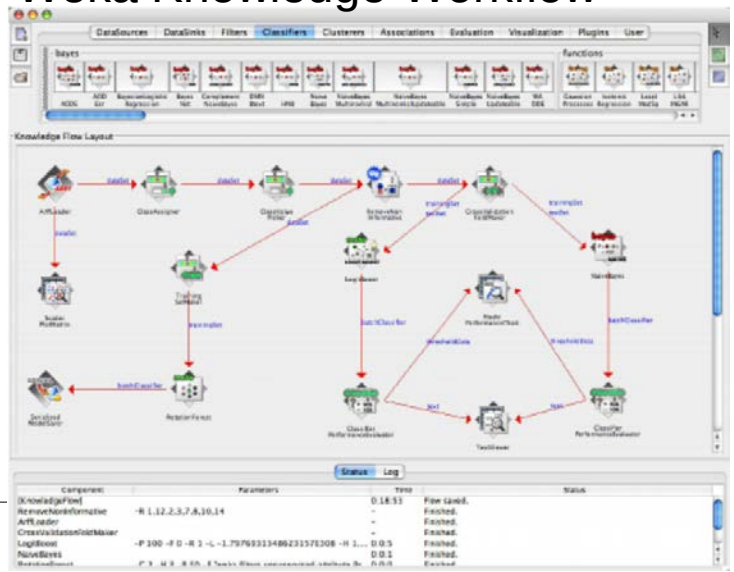
Weka - Waikato Environment for Knowledge Analysis

- Weka is a collection of open source machine learning algorithms for data mining tasks.
- Development started in 1993.
- Written as an API/library for Java
- Also has its own GUIs, called
  - **Weka Explorer**
  - **Weka Knowledge Workflow** (visual drag-and-drop programming of workflow)
- Datasets often on Attribute-Relation File Format (ARFF).

Weka Explorer



Weka Knowledge Workflow



Mittuniversitetet
MID SWEDEN UNIVERSITY

Lab 2 instructions (also on Moodle)

- Your task is to try data mining classification methods, for example in the tool Weka. This in order to give you a deeper knowledge about various tools and algorithms.
- Your task will be to use the dataset named "breast-cancer.arff" and apply the J48 tree and Naive Bayes data mining methods on the dataset. You should use two different approaches, both in the normal Weka Explorer GUI and by programming. J48 tree may be made in the Weka Explorer and Naive Bayes using Weka's Java API (or by visual programming using the Weka Knowledge Flow GUI).
- Weka GUI and Java API can be downloaded from: http://www.cs.waikato.ac.nz/ml/weka/
- You can download the dataset from here: Breast cancer recurrance dataset

Mittuniversitetet
MID SWEDEN UNIVERSITY

Preparation

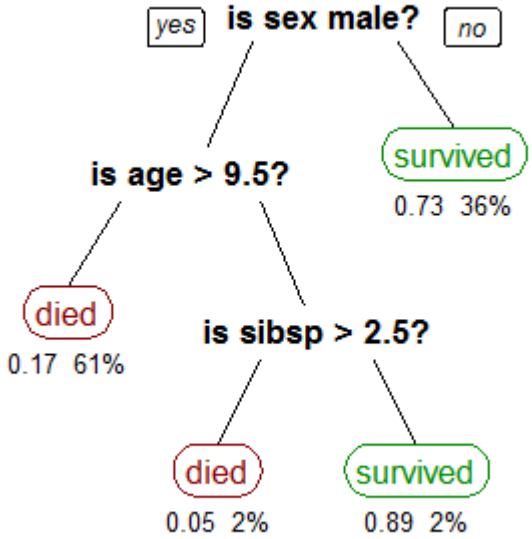
Read the reference material before the lab:

- References
  http://www.cs.waikato.ac.nz/~ml/weka/documentation.html
- Weka Help
  http://www.cs.waikato.ac.nz/~ml/weka/documentation.html

- Weka Java API Help
  http://weka.wikispaces.com/Use+WEKA+in+your+Java+code
- Breast cancer recurrence dataset Help
  https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer/breast-cancer.names

Decision tree learning

- Prediction method, often giving comprehensible results
- J48 Decision Tree algorithm
  - A Java implementation of the C4.5 algorithm, which is a successor of the ID3 algorithm
- In Weka, choose Weka Explorer -> Classify tab -> Trees
- In R, similar algorithms are available in the library rpart



Titanic example. Sibsp = number of siblings and spouses onboard. The figures under the leaves show the probability of survival and the percentage of observations in the leaf.

Mittuniversitetet
MID SWEDEN UNIVERSITY

Naive Bayes classifier

- Used in anti-spam filters, text analysis, document classification, etc.
- Spam filter example: The probabilities of specific words in spam and non-spam messages are stored in a database.
- Working under the assumption that all variables are mutually independent, except with the target variable.
- The attributes are binominal (false or true)
- The unkown datapoint will be assigned the class that has highest posteriori probability according to Bayes Theorem:

$$\hat{y} = \underset{k \in \{1,\dots,K\}}{\operatorname{argmax}} \; p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k).$$

Presenting the result

- Submission: To be graded, you must hand in the following documents:
  - A laboration report
  - A short demo/oral presenation of your result and preliminary report.
  - Present a step by step instruction for how to use J48 trees in the Weka GUI
  - The commented source code for your Java program or your Knowledge Flow chart

- The report must contain the following details:
  - An analysis on the results and the tree, when you have applied the J48 tree method
  - An analysis on the results when you have applied the Naive Bayes method
  - Explain the result, especially the details of the decision tree, and the confusion matris
  - Shortly present the purpose of each step in the Weka GUI
  - Shortly present the purpose of each line of code and in your Java program