LABORATORY OF DATA SCIENCE

Weka API

Data Science & Business Informatics Degree

Why API?

- Weka Explorer does not keep track of experimental settings
  - Every action overwrites the previous ones
- Weka Knowledge Flow documents the process, but …
  - it is time-consuming to experiment with many variants
    - (algs, params, inputs, …)
- In any case:
  - Models are typically re-built on a regular basis
    - A scheduling of the automated process must be planned
  - Models are deployed within larger applications
    - E.g., selection of customers in marketing campaigns can be suggested to the marketer by a decision-support system which exploits data mining models

Resources for the developer

- Weka documentation main page
  - http://www.cs.waikato.ac.nz/ml/weka/documentation.html
- Weka manual
  - Chapter 17: Using the API
- Weka API (developer version) javadoc
  - http://weka.sourceforge.net/doc.dev/
- Python-weka-wrapper3 package
  - easy run Weka algorithms and filters from Python
  - offers access to Weka API using thin wrappers around JNI calls using the **javabridge** package
  - https://pypi.org/project/python-weka-wrapper3

Main packages and classes

- weka.core
  - Instances – holds a complete dataset
  - Instance – encapsulates a single row
  - Attribute – holds the metadata of a column
- weka.core.converters
- weka.filter
- weka.classifiers
  - Evaluation
- weka.classifiers.trees
- weka.associations

Python-weka-wrapper3

- **Documentation:**
  - https://github.com/fracpete/python-weka-wrapper3
- **Installation:**
  - Download: https://pypi.org/project/python-weka-wrapper3/
  - And then from the cmd: `python setup.py install`
- **Requirements:**
  - JDK 1.8+

JVM

The use of the library requires to manage the **Java Virtual Machine** (JVM)

```
import weka.core.jvm as jvm

jvm.start()
```

Demo session

Practice

- Question:
  - does accuracy increase with percentage of training set?
- Starting from census.arff
  - Split into x% training and (100-x)% test
    - Stratified sampling, where x range in [20-80]
  - For which x accuracy is maximized?