BUSINESS INTELLIGENCE LABORATORY

Weka API

Salvatore Ruggieri

Computer Science Department, University of Pisa

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/

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

Option handling

- Either with get/set methods

```
Remove r = new Remove(); // unsupervised attribute filter
r. setAttributeIndices("1"));
```

- Or with the setOptions(String []) method

```
Remove r = new Remove();
String opt = "-R 1"; // options as shown in Weka Explorer
r.setOptions(opt.split(" "));
```

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?