Machine Learning with Weka

Sujatha Das Gollapalli
Cornelia Caragea

August 19, 2014

Thanks to Eibe Frank for some of the slides
WEKA: the software

- Machine learning/data mining software written in Java (distributed under the GNU Public License)
- Used for research, education, and applications
- Main features:
  - Comprehensive set of data pre-processing tools, learning algorithms and evaluation methods
  - Graphical user interfaces (incl. data visualization)
  - Environment for comparing learning algorithms
- WEKA website:
WEKA: resources

- API Documentation, Tutorials, Source code.
- WEKA mailing list
- *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*
- Weka-related Projects:
  - Weka-Parallel - parallel processing for Weka
  - RWeka - linking R and Weka
  - YALE - Yet Another Learning Environment
  - Many others…
WEKA: launching

- `java -jar weka.jar`
Data Preparation and Loading
@relation heart-disease-simplified

@attribute age numeric
@attribute sex { female, male}
@attribute chest_pain_type { typ_angina, asympt, non_anginal, atyp_angina}
@attribute cholesterol numeric
@attribute exercise_induced_angina { no, yes}
@attribute class { present, not_present}

@data
63,male,typ_angina,233,no,not_present
67,male,asympt,286,yes,present
67,male,asympt,229,yes,present
38,female,non_anginal,?,no,not_present
...

Data Preparation:
WEKA only deals with “flat” files

Flat file in ARFF format
@relation heart-disease-simplified

@attribute age numeric
@attribute sex { female, male}
@attribute chest_pain_type { typ_angina, asympt, non_anginal, atyp_angina}
@attribute cholesterol numeric
@attribute exercise_induced_angina { no, yes}
@attribute class { present, not_present}

@data
63,male,typ_angina,233,no,not_present
67,male,asympt,286,yes,present
67,male,asympt,229,yes,present
38,female,non_anginal,?,no,not_present
...

WEKA only deals with “flat” files

numeric attribute
nominal attribute
Explorer: pre-processing the data

- Data can be imported from a file in various formats: ARFF, CSV, C4.5, binary
- Data can also be read from a URL or from an SQL database (using JDBC)
- Pre-processing tools in WEKA are called “filters”
- WEKA contains filters for:
  - Discretization, normalization, resampling, attribute selection, transforming and combining attributes, …
Building Classifiers
Explorer: building “classifiers”

- Classifiers in WEKA are models for predicting nominal or numeric quantities
- Implemented learning schemes include:
  - Decision trees and lists, instance-based classifiers, support vector machines, multi-layer perceptrons, logistic regression, Bayes’ nets, …
- “Meta”-classifiers include:
  - Bagging, boosting, stacking, etc.
Classifiers

Choose: J48 -C 0.25 -M 2

Test options

- Use training set
- Supplied test set
- Cross-validation Folds: 10
- Percentage split %: 66

Classifier output

Status
OK
Classifier output

--- Run information ---
Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: weather
Instances: 14
Attributes: 5
outlook
temperature
humidity
windy
play

Test mode: split 66% train, remainder test

--- Classifier model (full training set) ---
J48 pruned tree

outlook = sunny
  | humidity <= 75: yes (2.0)
  | humidity > 75: no (3.0)
outlook = overcast: yes (4.0)
outlook = rainy
  | windy = TRUE: no (2.0)
  | windy = FALSE: yes (3.0)

Number of Leaves : 5
Size of the tree : 8

Time taken to build model: 0.11 seconds
Classifier:
Choose J48 -C 0.25 -M 2

Test options:
- Use training set
- Supplied test set (Set...)
- Cross-validation (Folds 10)
- Percentage split (% 66)

Classifier output:
Number of Leaves : 5
Size of the tree : 8
Time taken to build model: 0.11 seconds

Evaluation on test split:
Correctly Classified Instances 2 40%
Incorrectly Classified Instances 3 60%
Kappa statistic -0.3636
Mean absolute error 0.6
Mean squared error 0.7746
Relative absolute error 126.9231%
Root mean squared error 157.6801%
Total Number of Instances 5

Detailed Accuracy By Class:
<table>
<thead>
<tr>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.667</td>
<td>1</td>
<td>0.5</td>
<td>0.667</td>
<td>0.571</td>
<td>yes</td>
</tr>
<tr>
<td>0</td>
<td>0.333</td>
<td></td>
<td>0</td>
<td></td>
<td>no</td>
</tr>
</tbody>
</table>

Confusion Matrix:

a b <-- classified as
2 1 | a = yes
2 0 | b = no
Weka Explorer

Classifier
Choose J48 -C 0.25 -M 2

Test options
- Use training set
- Supplied test set
- Cross-validation Folds 10
- Percentage split % 66

More options...

(Nom) play
Start Stop

Result list (right-click for options)

Status OK


Tree View

- outlook
  - sunny
  - overcast
  - rainy
    - humidity
      - <= 75
        - yes (2.0)
      - > 75
        - no (3.0)
    - windy
      - = TRUE
        - no (2.0)
      - = FALSE
        - yes (3.0)
Classifier output

Number of Leaves : 5
Size of the tree : 8

Time taken to build model: 0.11 seconds

Evaluation on test split

Correctly Classified Instances 2 40 %
Incorrectly Classified Instances 3 60 %
Kappa statistic -0.3636
Mean absolute error 0.6
Root mean squared error 0.7746
Relative absolute error 126.9231 %
Root relative squared error 157.6801 %
Total Number of Instances 5

Detailed Accuracy By Class

<table>
<thead>
<tr>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.667</td>
<td>1</td>
<td>0.5</td>
<td>0.667</td>
<td>0.571</td>
<td>yes</td>
</tr>
<tr>
<td>0</td>
<td>0.333</td>
<td></td>
<td></td>
<td></td>
<td>no</td>
</tr>
</tbody>
</table>

Confusion Matrix

2 1 a = yes
2 0 b = no
Weka Explorer

Classifier
Choose NaiveBayes

Test options
- Cross-validation Folds 10
- Percentage split % 66

Classifier output

--- Run information ---
Scheme: weka.classifiers.bayes.NaiveBayes
Relation: weather
Instances: 14
Attributes: 5
- outlook
- temperature
- humidity
- windy
- play

Test mode: 10-fold cross-validation

--- Classifier model (full training set) ---

Naive Bayes Classifier
Class yes: Prior probability = 0.63
outlook: Discrete Estimator. Counts = 3 5 4 (Total = 12)
temperature: Normal Distribution. Mean = 72.9697 StandardDev = 5.2304
humidity: Normal Distribution. Mean = 78.8395 StandardDev = 9.8023
windy: Discrete Estimator. Counts = 4 7 (Total = 11)

--- Classifier model summary ---

Log
Try Decision Tree, Naïve Bayes, and Logistic Regression and Support Vector Machines classifiers on a CiteSeerX dataset

- The dataset contains titles and abstracts of papers from Computer Science that are available in the CiteSeer digital library;
- The class for each example in the dataset is the topic of the paper. There are six possible classes.
- The dataset is available in arff format.

Use various model parameters