ECE7252
Statistical Learning for Data Processing

Lecture 6: Machine Learning Tools and Web Resources

Chin-Hui Lee
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332, USA
chl@ece.gatech.edu

Machine Learning Datasets

- UCI machine learning repository
  - http://archive.ics.uci.edu/ml/
  - 160 data sets for general purpose machine learning
  - **Iris data set** and **glass identification data set** will be used for demonstration
- Language, speech and video corpus
  - http://www.ldc.upenn.edu/ **(Linguistic Data Consortium)**
- Machine learning & data mining: face, objects, etc.
  - http://cervisia.org/machine_learning_data.php
- Open Directory Project:
  - http://www.dmoz.org/Computers/Artificial_Intelligence/Machine_Learning/Datasets/
- Datasets for knowledge discovery
  - http://www.kdnuggets.com/datasets/
- BBC datasets: news and sports
  - http://mlg.ucd.ie/content/view/21/

Machine Learning Toolkits

- Netlab : neural network and Gaussian process (matlab code)
  - http://www.ncrg.aston.ac.uk/netlab/over.php
- HTK and GMTK: speech modeling kits
  - http://htk.eng.cam.ac.uk/ (HTK)
  - **http://ssli.ee.washington.edu/~bilmes/gmtk/ (GMTK)**
  - **http://www.cs.ubc.ca/~murphyk/Software/BNT/bnt.html (Bayes Net Toolbox)**
- CMU AI Repository
  - http://www.cs.cmu.edu/afs/cs/project/ai-repository/ai/areas/learning/systems/0.html
- JMLR machine learning open source software
  - http://jmlr.csail.mit.edu/mloss/
- R: http://www.r-project.org/
  - A free alternative to S-Plus developed at Bell Labs
  - If you know C, you will be right at home with R
- Weka: data mining tool in Java
  - http://www.cs.waikato.ac.nz/ml/weka/

Machine Learning Review

- Data preprocessing
  - Normalization, discretion, standardization, etc.
- Predictive learning (supervised)
  - Regression (continuous output variable)
  - Classification (nominal output variable)
- Clustering (unsupervised)
  - VQ, EM, etc.
- Association rule learning (market basket analysis)
- Attribute selection
  - Search algorithm
  - Evaluation measure
- Performance evaluation
  - Error rate, accuracy, recall, precision, F measure, etc.

Weka Tutorial

- Weka package is open source data mining and machine learning software written in Java.
- Weka can be used from the GUI, the command line or called by your own java code.
- Weka provides a variety of tools for data preprocessing, performance evaluation and significance testing.

Input File Format

- ARFF (Attribute-Relation File Format)

```
% 1. Title: Iris Plants Database
% 2. Sources:
% (a) Creator: R.A. Fisher

@RELATION iris

@ATTRIBUTE sepallength NUMERIC
@ATTRIBUTE sepalwidth NUMERIC
@ATTRIBUTE petallength NUMERIC
@ATTRIBUTE petalwidth NUMERIC
@ATTRIBUTE class {Iris-setosa,Iris-versicolor,Iris-virginica}

@DATA

5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
```

Data Preprocessing

- Normalization (0 – 1, or any other intervals)
  - weka.filters.unsupervised.attribute.Normalize
  - Demo:
- Standardization (zero-mean unit-variance)
  - weka.filters.unsupervised.attribute.Standardize
  - Demo:

Classification

- Zero Rule (lower bounds) (rule-based)
  - weka.classifiers.rules.ZeroR
- Decision Tree (tree-based)
  - weka.classifiers.trees.J48 (C4.5 in Weka)
- Neural Network (function-based)
  - weka.classifiers.functions.MultilayerPerceptron
- K-Nearest Neighbours (instance-based)
  - weka.classifiers.lazy.IBk (KNN in Weka)
- Naïve Bayes (Bayes classifier)
  - weka.classifiers.bayes.NaiveBayes

Weka Explorer

- Neural network learning on iris dataset
  - Demo:

Classifier output interpretation

- Prediction on testing instance
  - Demo:
  - inst#, actual, predicted, error, probability distribution

```
 1    1:Iris-set 1:Iris-set        *0.774  0.209  0.017
 2    1:Iris-set 1:Iris-set        *0.769  0.213  0.018
…
51    2:Iris-ver 3:Iris-vir    +    0.087  0.415 *0.497
52    2:Iris-ver 3:Iris-vir    +    0.106  0.423 *0.472
...
```

Performance Measure

- Summary
  - Correctly Classified Instances 124 82.6667 %
  - Incorrectly Classified Instances 26 17.3333 %
  - Kappa statistic 0.74
  - Mean absolute error 0.2703
  - Root mean squared error 0.3262
  - Relative absolute error 60.8122 %
  - Root relative squared error 69.1965 %
  - Total Number of Instances 150
- === Detailed Accuracy By Class ===
  - TP Rate FP Rate Precision Recall F-Measure ROC Area Class
  - 1 0.03 0.943 1 0.971 1 Iris-setosa
  - 0.48 0 1 0.48 0.649 0.989 Iris-versicolor
  - 1 0.23 0.685 1 0.813 0.996 Iris-virginica
- === Confusion Matrix ===
  - a b c <-- classified as
  - 50 0 0 | a = Iris-setosa
  - 3 24 23 | b = Iris-versicolor
  - 0 0 50 | c = Iris-virginica

Weka Experimenter

- Demo: 4 schemes on 2 datasets
  - ZeroR
  - Decision Tree
  - Naïve Bayes
  - KNN

Significance Test

- Tester: weka.experiment.PairedCorrectedTTester
- Analysing: Percent_correct
- Datasets: 2
- Resultsets: 4
- Confidence: 0.05 (two tailed)
- Sorted by: -
- Date: 1/27/08 5:29 PM

Dataset	(1) rules.Ze |	(2) bayes	(3) lazy.	(4) trees
Glass	(100) 35.51 |	49.45 v	70.02 v	67.63 v
iris	(100) 33.33 |	95.53 v	95.20 v	94.73 v
	(v/ /*) |	(2/0/0)	(2/0/0)	(2/0/0)

The accuracy for each of the 4 schemes is shown in each dataset row.

The annotation “v” or “*” indicates that a specific result is statistically better (v) or worse (*) than the baseline scheme at the significance level specified.

Summary

- Today's Class
  - Class project discussion
- Next Class
  - Overview on supervised learning
- Reading Assignments
  - HTF, Chapters 1, 2 & 3
  - *HAL's Legacy*, Chapters 6, 7 & 8