

AI course Project 1

2012.04.26

Eun-Sol Kim



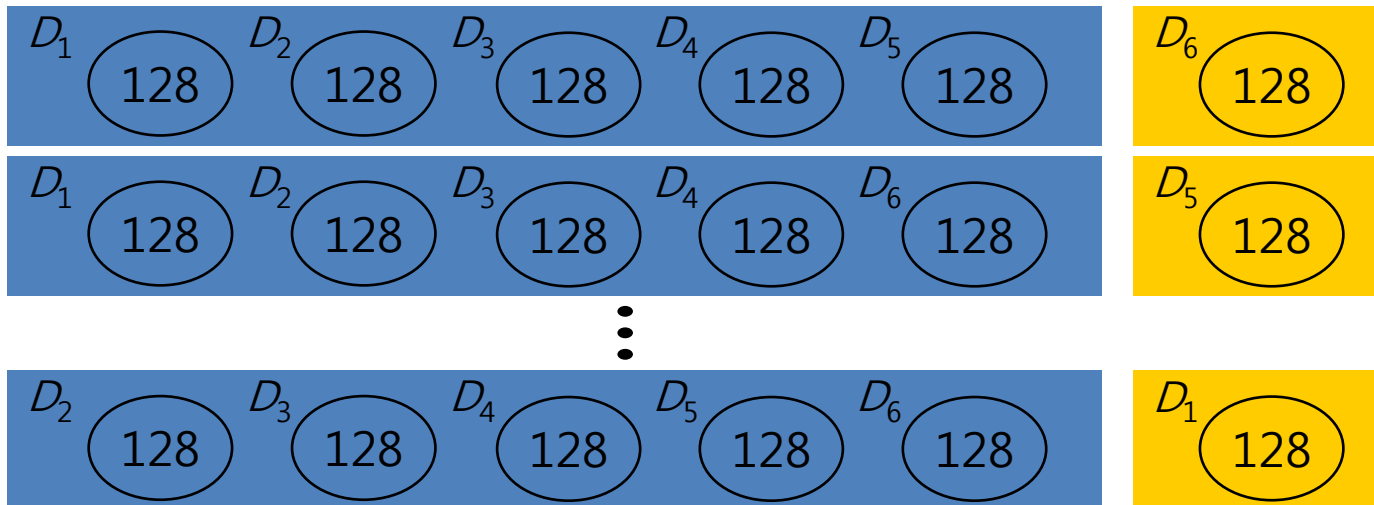
Evaluation Method - Cross Validation

- **K-fold Cross Validation**

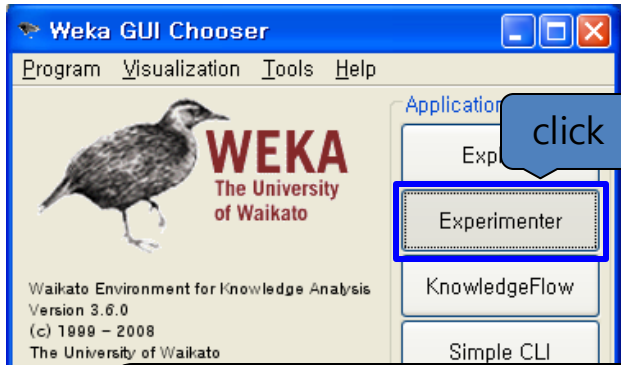
- The data set is randomly divided into k subsets.
- One of the k subsets is used as the 'test set' and the other $k-1$ subsets are put together to form a 'training set'.

6-fold cross validation

$$Error = \frac{1}{k} \sum_{i=1}^k Error_i$$



Using Experimenter in Weka



Weka GUI Chooser

Program Visualization Tools Help

WEKA
The University of Waikato

Waikato Environment for Knowledge Analysis
Version 3.6.0
(c) 1999 - 2008
The University of Waikato
Hamilton, New Zealand

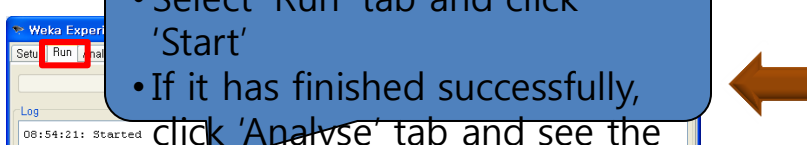
Application

Experimenter

KnowledgeFlow

Simple CLI

click



Weka Experiment Environment

Setup Run Analyse

Experiment Configuration Mode: Simple

Open... Save... New

Results Destination

ARFF file Filename: Browse...

Experiment Type

Cross-validation

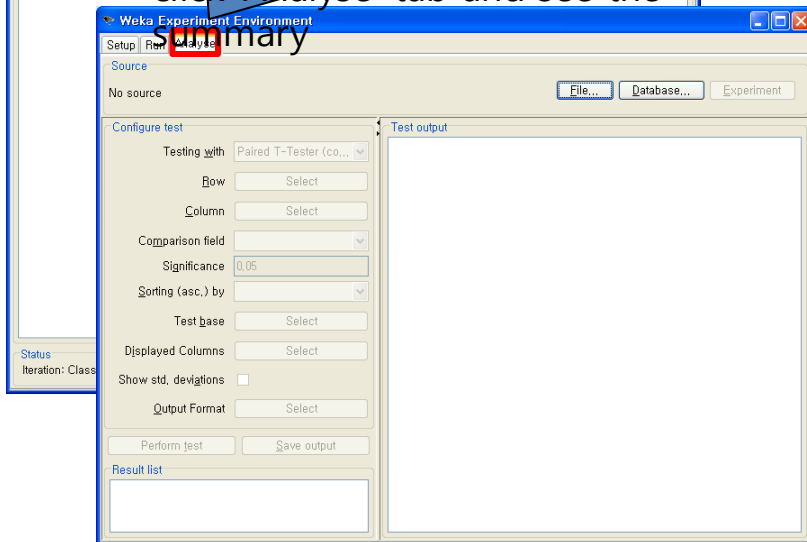
Number of folds: 10

Classification Regression

Datasets

Add new... Edit select... Delete select...

Use relative pat...



Weka Experiment Environment

Setup Run Analyse

Experiment Configuration Mode: Simple

Open... Save... New

Results Destination

ARFF file Filename: Browse...

Experiment Type

Cross-validation

Number of folds: 10

Classification Regression

Datasets

Add new... Edit select... Delete select...

Use relative pat...

E:\WPProgram Files\Weka-3-6\data\Wiris.arff

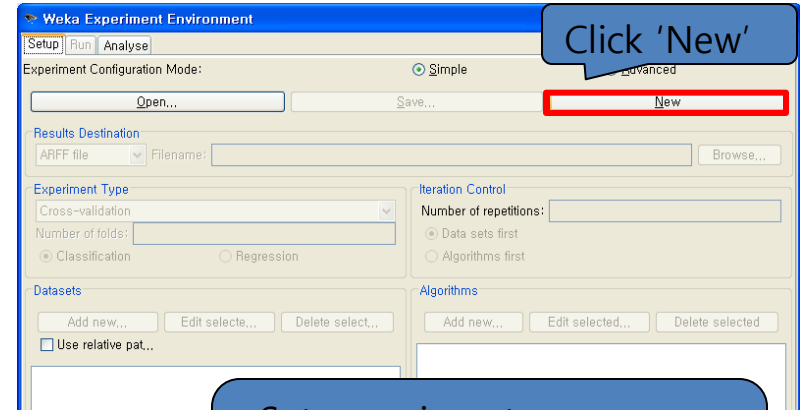
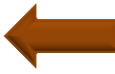
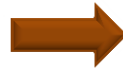
Up Down Load options... Save options... Up Down

Notes

summary

- Select 'Run' tab and click 'Start'
- If it has finished successfully, click 'Analyse' tab and see the summary

Experiments



Weka Experiment Environment

Setup Run Analyse

Experiment Configuration Mode: Simple

Open... Save... New

Results Destination

ARFF file Filename: Browse...

Experiment Type

Cross-validation

Number of folds: 10

Classification Regression

Datasets

Add new... Edit select... Delete select...

Use relative pat...

Iteration Control

Number of repetitions: 10

Data sets first Algorithms first

Algorithms

Add new... Edit selected... Delete selected

Use relative pat...

MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a
MultilayerPerceptron -L 0.3 -M 0.2 -N 100 -V 0 -S 0 -E 20 -H a
MultilayerPerceptron -L 0.1 -M 0.2 -N 100 -V 0 -S 0 -E 20 -H a

Up Down Load options... Save options... Up Down

Notes

Click 'New'

- Set experiment type/iteration control
- Set datasets / algorithms

Project 1

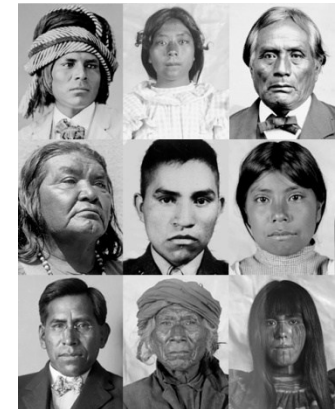
Project 1

- Classification problem with Weka
 - Data set
 - 3 different data sets
 - MNIST, Prima Indians Diabetes+1
 - Classification methods
 - MLP: iters, learning rate, momentum, # of hidden nodes, cross-validation

- Contents in the report
 - You should
 - compare the results of various parameter settings for MLPs
 - find optimal parameter setting for MLP and report the classification performance on that setting on all data sets
 - Include discussions
 - At most A4 four pages
 - Due date: 7th May 2012, pm 11:59 (via email)

Dataset #1: Pima Indians Diabetes

- Description
 - Pima Indians have the highest prevalence of **diabetes** in the world
 - We will build **classification models** that diagnose if the patient shows signs of diabetes
 - <http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes>
- Configuration of the data set
 - 768 instances
 - 8 attributes
 - age, number of times pregnant, results of medical tests/analysis
 - all numeric (integer or real-valued)
 - Also, a discretized set will be provided
 - Class value = 1 (Positive example)
 - Interpreted as "tested positive for diabetes"
 - 500 instances
 - Class value = 0 (Negative example)
 - 268 instances



Dataset #2: Handwritten Digits (MNIST)

- Description
 - The MNIST database of handwritten digits contains digits written by office workers and students
 - We will build a **recognition** model based on classifiers with the reduced set of MNIST
 - <http://yann.lecun.com/exdb/mnist/>
- Configuration of the data set
 - Attributes
 - pixel values in gray level in a 28x28 image
 - 784 attributes (all 0~255 integer)
 - Full MNIST set
 - Training set: 60,000 examples
 - Test set: 10,000 examples
 - For our practice, a reduced set with 800 examples is used
 - Class value: 0~9, which represent digits from 0 to 9

3 6 8 1 7
6 7 5 7 8
2 1 7 9 7
4 8 1 9 0
7 6 1 8 6