Introduction to Artificial Intelligence (236501)

Home assignment #3 – Machine learning

Due date: Monday, February 13, 2005, at 12:00 noon.
Submit your work in pairs. Only typed (i.e., not handwritten) submissions will be accepted. Warning: Absolutely no extensions will be granted! Late submissions will not be accepted.

In this assignment you will gain hands-on experience in applying ML algorithms to real-world datasets. You will not need to write any code (unless you want to), as you can use an existing ML package (Weka).

Weka

Weka is a publicly available suite of machine learning algorithms implemented in Java. You can use the algorithms either through the GUI (Explorer or Experimenter) or through the command-line interface. The former option is recommended. You can install Weka on any computer (including your home computer) by downloading its distribution package from http://www.cs.waikato.ac.nz/ml/weka (a number of tutorials and documentation pointers are available at this address as well). WekaDoc – an extensive documentation site on Weka – is available at http://weka.sourceforge.net/wekadoc/index.php/Main_Page

A very cool part of Weka is KnowledgeFlow – a graphical environment where you design your learning experiment from basic building blocks such as loading the data, cross-validation, classifier, visualizer, etc. Detailed documentation about KnowledgeFlow is available at http://weka.sourceforge.net/wekadoc/index.php/en:Knowledge_Flow_%283.5.2%29

Weka is installed on CSL1. Running instructions:
To run Weka on CSL1, use the following commands (you need an X terminal):

```
cd /usr/local/weka-3-4-7
/usr/java/jdk1.5.0_06/bin/java -jar weka.jar
```

UCI Repository of ML datasets

The ML community often uses a standard set of datasets for evaluating algorithms. These datasets are available from a repository maintained at UC Irvine, which is accessible at http://www.ics.uci.edu/~mlearn/MLRepository.html and http://www.ics.uci.edu/~mlearn/MLSummary.html. Each dataset is accompanied by a brief description.

Many datasets in the UCI repository are in the so-called C4.5 data format, which is very common in machine learning. A brief description of this format is available at http://www.cs.washington.edu/dm/vfml/appendixes/c45.htm. You can learn more about C4.5 data format at the source: J. Ross Quinlan, “C4.5: Programs for Machine Learning”, Morgan Kaufmann, 1993. Some datasets use other data formats, which are explained in the documentation files that accompany each dataset.

Weka accepts C4.5 format as well as ARFF, Weka’s own format, which is explained in detail in http://www.cs.waikato.ac.nz/~ml/weka/arff.html and http://weka.sourceforge.net/wekadoc/index.php/en:ARFF_%283.5.1%29
What you need to do

1) Select any 3 datasets from the UCI repository
2) Convert them into a format that can be understood by Weka
3) For each dataset D, perform the following experiments using 10-fold cross-validation:
   a) Classify D using ID3
   b) Classify D using KNN for K = 1, 5, 10, 25, 50
   ➔ Summarize your experimental results in tables and graphs
4) Compare the performance of ID3 and KNN based on the accuracy they yielded in the above
   experiments. Draw conclusions about their relative performance with respect to the datasets
   of different nature and with different parameter settings (values of K).
5) For ID3 and for KNN with the value of K that yielded the best performance in step (3) above,
   repeat the experiments with different levels of feature selection (called “attribute selection”
   in Weka). That is, classify each dataset D with each of the following feature selection levels:
   {5%, 10%, 25%, 50%, 100%}. To perform feature selection, use the Information Gain
   criterion, each time selecting the appropriate fraction of the highest-scoring features. Don’t
   forget to perform all experiments under 10-fold cross-validation!
   Attribute selection is available KnowledgeFlow through Filters / Supervised /
   AttributeSelection.
6) For ID3 and for KNN with the value of K that yielded the best performance in step (3) above,
   repeat the experiments to build learning curves, i.e., perform each experiment for
   increasingly large fractions of the training set, and plot the accuracy as a function of the
   cardinality of the training set. Explain your findings.
   Plotting the learning curve: suppose you’re using cross-validation with N=10 folds.
   Normally, for each of the 10 folds you use 9/10 of the data for training (TRAINi, i=1..10),
   and 1/10 for testing (TESTi, i=1..10). To build a learning curve, you also consider subsets of
   the training data. That is, for each fold you repeat the experiment by using 10%, 20%, …,
   100% of TRAINi for training, while still using the entire test set (TESTi) for testing.
   To build learning curves in Weka:
   a) In Experimenter, use Advanced Mode / Result Generator /
      weak.experiment.LearningRateResultProducer
   b) In KnowledgeFlow, select Filters / Unsupervised / RemovePercentage
   Remember, there are usually several ways to accomplish different tasks in Weka.

Notes

1. There is no need to write any code (except for a few lines of scripts to convert the data into a
   suitable format for running Weka). Thus, there is no need to submit any code.
2. The analysis of experiments constitutes the main point of this assignment. Therefore, you
   should discuss your experimental results in detail, explaining the nature of your results and
   observations. Specifically, compare ID3 to KNN (with different values of K), and discuss
   their performance on different datasets, different levels of feature selection and different
   fractions of the training set.
3. Your entire submission should be in hard copy (no electronic submission is necessary for this
   assignment).
Submit your work in pairs. In order to find a partner, you may take advantage of the “Find a partner” mechanism available on the course Web page at http://webcourse.technion.ac.il/236501 (see buttons on the left).

Good luck!

Questions? Evgeniy Gabrilovich (gabr@cs.technion.ac.il)
Flames? /dev/null