CptS 580: Advanced Topics in Machine Learning

Homework #1
Due: Tuesday, September 6, 2011

Description

This homework assignment allows you to review some of the machine learning techniques you have learned and used in earlier classes and research projects. However, there are two additional skills that you will gain from the assignment. The first is familiarity with the Orange data mining environment. The second is a critical assessment of the state of the art of machine learning techniques and their limitations for real-world applications.

Part One

In the first part of this assignment, you are asked to download and install Orange and apply familiar machine learning algorithms within this environment to a standard dataset. The Orange toolkit is available from http://orange.biolab.si/ and will run on Windows, Mac OS, and Linux.

The datasets you will use for this homework are:

- Voting Records, available at http://archive.ics.uci.edu/ml/ datasets/Congressional+ Voting+Records. This is a fairly simple dataset that maps binary-valued voting decisions on various issues to the class party of the individual (democrat or republican). There are missing values in the dataset.

- Yeast, available at http://archive.ics.uci.edu/ml/datasets/Yeast. This dataset maps features of a protein sequence to the cellular localization site. The dataset has different characteristics than the voting dataset, primarily because the features are continuous valued.

As you look at the Orange documentation, you will see that widgets are available for a number of standard machine learning algorithms, visualization tools, and preprocessing tools. For this assignment, please run the following tools on the voting and yeast datasets, as appropriate.

- Discretization (any type, specify which was used and why)
- Naïve Bayes classifier
- Decision tree
- K nearest neighbor
- Majority

Analyze the results using the following measures:

- Accuracy using 10-fold cross validation
- Area Under the ROC Curve (AUC)
- Sensitivity
- Specificity

Include a summary of the performance results, a description of additional tools used if appropriate, and observations about the techniques.

Part Two

The second part of the assignment is designed to position you for insightful discussions throughout the rest of the semester. For this part, generate a list of 20 challenges for machine learning applications that are not easily handled by the machine learning algorithms described in the introductory machine learning class. These challenges can be limitations of existing methods, learning situations not addressed by current techniques (at least those described in the intro class), or open research issues. You will be graded based on the quality, diversity, and uniqueness of the challenges you mention (so you might not want to share these with others!). Draw from insights gained in finishing part one, from your own experiences applying machine learning techniques, and from issues discussed by other machine learning researchers.

Turning in the Assignment

The assignment should be mailed as a PDF file to cook@eecs.wsu.edu by 9:00am on the due date. No late assignments will be accepted.