CSCI 447 — Machine Learning: Soft Computing

Project #1

Assigned: August 28, 2017
Design Document Due: September 11, 2017
Project Due: September 18, 2017

The purpose of this assignment is to give you an introduction to machine learning by exploring a freely available environment for running machine learning experiments. Note that you will not be permitted to use this environment in future projects, but it was felt that this was a good way to get "up to speed" in learning how to design experiments and seeing immediate results.

This experiment requires you to download, install, and use the Waikato Environment for Knowledge Acquisition (WEKA). In addition to testing the various machine learning algorithms in WEKA, you are required to write a research paper describing the results of your experiments. For this project, the following steps are required:

- Download WEKA at `http://www.cs.waikato.ac.nz/ml/weka/`.
- Download five (5) data sets from the UCI Machine Learning repository. You can find this repository at `http://archive.ics.uci.edu/ml/`. These data sets should each have at least five attributes and should be data sets used for classification problems. It is strongly recommended that you select data sets with common feature types and no missing values.
- Prepare a design document that focuses on the ARFF converter as well as your overall experimental design.
- Create a utility program to convert data sets from the UCI ML Repository into the Attribute Relationship File Format (ARFF), as required by WEKA. Details on this file format can be found at `http://weka.wikispaces.com/ARFF`. Note that WEKA comes with several data sets already in ARFF format. If you decide to use any of these data sets, you must work from the original data on the UCI ML Repository and convert them to ARFF using your converter. You may be asked to demonstrate your converter on any data set of the instructor's choosing.
- Devise or select at least two (2) different evaluation measures (i.e., loss functions) that you will use to evaluate your algorithms.
- Develop a hypothesis for each data set based on expected performance of the various algorithms, emphasizing the role of their respective inductive biases in the expected performance.
- Design and execute experiments in WEKA to test your hypotheses, comparing the following five (5) machine learning algorithms, each run on the five UCI ML data sets you choose. Make sure the experiment generates sufficient data from which you can draw meaningful statistical conclusions. Be sure to tune each algorithm appropriately.
    1. K-nearest neighbor (IBk) using $k > 2$.
    2. Naïve Bayes
    3. Logistic regression
    4. Decision tree with pruning (J48)
    5. Support vector machine with a nonlinear kernel (LibSVM or SMO)
- Write a paper that incorporates the following elements, summarizing the results of your experiments. Make sure you explain the experimental setup, the tuning process, and the final parameters used for each algorithm.
    1. Title and author name
    2. A brief, one paragraph abstract summarizing the results of the experiments

3. Problem statement, including hypothesis
4. Description of algorithms implemented
5. Description of your experimental approach
6. Presentation of the results of your experiments
7. A discussion of the behavior of your algorithms, combined with any conclusions you can draw
8. Summary
9. References (you should have at least one reference related to each of the algorithms implemented, a reference to the data sources, and any other references you consider to be relevant)

- Submit your design document, fully documented code for the data converter, results of the runs of each algorithm, and your paper.