CSCI 446 — Artificial Intelligence

Project #3
Experiments in Machine Learning

For the third programming assignment, you will have an opportunity to explore machine learning by implementing four powerful learning algorithms. These four algorithms are called NearestNeighbor, Naïve Bayes, TAN, and ID3. For this assignment, you will use five datasets that you will download from the UCI Machine Learning Repository, namely:

1. Breast Cancer — `https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29`

   This breast cancer databases was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg.

2. Glass — `https://archive.ics.uci.edu/ml/datasets/Glass+Identification/`

   The study of classification of types of glass was motivated by criminological investigation.

3. Iris — `https://archive.ics.uci.edu/ml/datasets/Iris/`

   The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant.

4. Soybean (small) — `https://archive.ics.uci.edu/ml/datasets/Soybean+%28Small%29/`

   A small subset of the original soybean database.

5. Vote — `https://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records/`

   This data set includes votes for each of the U.S. House of Representatives Congressmen on the 16 key votes identified by the Congressional Quarterly Almanac.

When using these data sets, be careful of some issues.

1. Some of the data sets have missing attribute values. When this occurs in low numbers, you may simply edit the corresponding values out of the data sets. For more occurrences, you should do some kind of "data imputation" where, basically, you generate a value of some kind. This can be purely random, or it can be sampled according to the conditional probability of the values occurring, given the underlying class for that example. The choice is yours, but be sure to document your choice.

2. Most of attributes in the various data sets are either multi-value discrete (categorical) or real-valued. You will need to deal with this in some way. For the continuous attributes, to be consistent you will need to discretize them in some way for both algorithms and then proceed as in the multi-valued categorical case. Technically, nearest-neighbor does not require discrete attributes, but we want the algorithms to be comparable. For nearest-neighbor, you will use the Value Difference Metric to compare.

For this project, the following steps are required:

- Download the five (5) data sets from the UCI Machine Learning repository. You can find this repository at `http://archive.ics.uci.edu/ml/`. All of the specific URLs are also provided above.
- Pre-process each data set as necessary to handle missing data and continuous (real-valued) data.
- Set up your test and training sets from the provided data to implement either 10-fold cross-validation or 5 × 2 cross-validation. This is entirely your choice, but make sure you do the same think for all experiments.
- Implement k-NN, Naïve Bayes, TAN, and ID3 with reduced-error pruning.
- Run your algorithms on each of the data sets. These runs should output the classifications on all of the test examples. When doing cross-validation, just output classifications for one fold each. But you do need to average over all of the folds to do your analysis for your report. Also, you will need to tune k.

- Write a very brief paper that incorporates the following elements, summarizing the results of your experiments. You should also output the summary statistics on classification accuracy.
    1. Title and author name
    2. A brief, one paragraph abstract summarizing the results of the experiments
    3. Problem statement, including hypothesis, projecting how you expect each algorithm to perform
    4. Brief description of algorithms implemented
    5. Brief 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 fully documented code for the data converter, results of the runs of each algorithm, your design document, and your paper.

Due Dates:

- Design Document – November 4, 2016
- Program Code and Sample Runs – November 18, 2016
- Project Report – November 21, 2016