605.449 — Introduction to Machine Learning

Programming Project #1

Due: September 9, 2018

The purpose of this assignment is to give you an introduction to machine learning by implementing two fairly simple learning algorithms. These two algorithms are called WINNOW-2 (introduced in Module 01) and NAÏVE BAYES (introduced in Module 02). For this assignment, you will use three of the following 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. Not all of these data sets correspond to 2-class classification problems. For Naïve Bayes, that is not really problem. But for Winnow-2, you will need to use one classifier for each class. To be fair in comparing the two, you may want to generate Boolean classifiers for Naïve Bayes, even though this is not strictly required.

2. 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.

3. 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 multi-value situation, once again Naïve Bayes should be fine, but Winnow-2 will have a problem. In that case, you can apply what is called "one-hot coding" where you create a separate Boolean attribute for each value. Again, I recommend you go ahead and use this for Naïve Bayes, even though it is not really necessary. For the continuous attributes, you will need to discretize them in some way for both algorithms and then proceed as in the multi-valued categorical case.

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 non-Boolean data (both classes and attributes).

- Set up your test and training sets from the provided data. Specifically, split the data into two groups randomly where 2/3 of the data will be used for training and 1/3 will be used for testing. If you are more ambitious, you may set up a cross-validation experiment. In that case, I recommend either 10-fold cross-validation or 5 × 2 cross-validation. If you don't know what this means, don't worry about it for now.
- Implement both Naïve Bayes and Winnow-2.
- Run your algorithms on three of the five the data sets. These runs should output the learned models in a way that can be interpreted by a human, and they should output the classifications on all of the test examples. If you are doing cross-validation, just output classifications for one fold each.
- For extra credit (for up to 10 additional points), run your algorithms on all five data sets.
- 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, the outputs from running your programs, and your paper. Your grade will be broken down as follows:
    - Code structure – 10%
    - Code documentation/commenting – 10%
    - Proper functioning of your code, as illustrated by a 5 minute video – 30%
    - Summary paper – 50%