Introduction to Machine Learning - Assignment 1

Instructor: Dan Lizotte

Due at the beginning of class on Monday, 30 April 2007

This assignment covers decision trees, PAC learning, and VC dimension.
It is marked out of 50 and is worth 15% of your final mark.
For this assignment, submit a hard copy of all of your answers and of your code for Question 1.

---

1. [*20 points*] **Decision Stump**

   "Entia non sunt multiplicanda praeter necessitatem." - William of Ockham
   "K.I.S.S. - Keep it simple, stupid." - Mr. Miller, my Grade 10 math teacher.

   (*Alpaydin, Ch. 9*) Sometimes, a simple explanation works extremely well. A *decision stump* is, as you might expect, a really short decision tree. The decision tree has one split, and each leaf is labeled with its majority class. In the language of your choice, write a program that reads in the following data file and computes and displays the information gain (in bits) resulting from splitting on each attribute. Then display the training set error of each of the resulting decision stumps. The output should be something like:

   ```
   Attribute:      1
   Info gain:      foo
   Training error: blah
   ```

   or similar. The relevant files are here:

   `ftp://ftp.ics.uci.edu/pub/machine-learning-databases/voting-records`

   There is an index of the directory, a `.names` file that describes the data, and a `.data` file that contains the actual records. This data belongs to the UCI Machine Learning Repository at

   `http://www.ics.uci.edu/~mlearn/MLRepository.html`

   You'll notice that there are many missing feature values, which are marked as '?'. In fact, nearly half the records have at least one missing feature. This problem is pervasive in machine learning. For the purposes of this assignment, simply consider '?' to be another possible value the feature can take – that is, each feature is *ternary* and takes values in $\{y, n, ?\}$.

   Here are some hints:

   - You may find it easier to pre-process the file by running it through the following command:
     ```
     cat house-votes-84.data
     | sed 's/democrat/1/g;s/republican/-1/g;s/n/-1/g;s/?/0/g;s/y/1/g'
     > house-votes-84-numlabels.csv
     ```
     This will replace the labels with +1 if democrat, -1 if republican. Feature values are changed from {y,?,n} to {1,0,-1}. On the other hand, you may want to just leave things as they are, or make other label substitutions depending on the language you choose to use.
   - $\log_2(x) = \log_b(x)/\log_b(2)$
   - Information gain is always positive.

2. [*10 points*] **PAC Learning Decision Stumps**

   (a) What is the size of the hypothesis space of decision stumps over one binary class and m ternary attributes? (Note: “attribute” is just another word for “feature”.)

   (b) Give a bound on the number of training examples needed to PAC learn a decision stump over m attributes with 95% accuracy 19 times out of 20.

3. [*10 points*] **VC dimension of triangles** *(Similar to Question 10 on p.38 of Alpaydin.)* Show that the VC dimension of the “triangles” hypothesis class in $\mathbb{R}^2$ is *at least* 7. This class is the set of all triangles in $\mathbb{R}^2$; each hypothesis is defined by three points describing the vertices of a triangle: $(x_1, y_1), (x_2, y_2), (x_3, y_3)$. In this class, a point (x, y) is labeled “+” if it is inside the triangle, and “−” otherwise.

   Use pictures and explain your answer. You don’t have to draw all 128 triangles, but make your argument convincing.

4. [*10 points*] **Performance bounds using VC dimension**

   (a) Suppose our 2-D triangle classifier achieves a training error rate of 0.05 on a dataset of size 100. Give a bound on the test error of this classifier that will hold 95% of the time. (Assume the VC dimension of our hypothesis space is exactly 7.)

   (b) Give a similar bound assuming we got the above results using a linear classifier.