CS 6375 Machine Learning

Homework 5

Due: 04/12/2015, 11:59pm

1. Boosting. (30 pts)

Three learning algorithms in a binary classification problem are applied independently to a set of 1000 training examples to train three classifiers.

- **Algorithm A** produces **Classifier A** that correctly classifies 800 examples and incorrectly classifies 200 examples.
- **Algorithm B** produces **Classifier B** that correctly classifies 800 examples and incorrectly classifies 200 examples. All the mistakes of Classifier B are on examples that were correctly classified by Classifier A.
- **Algorithm C** produces **Classifier C** that correctly classifies 900 examples and incorrectly classifies 100 examples. All the mistakes of Classifier C are on examples that were correctly classified by both Classifier A and Classifier B.

Combine the three classifiers using the AdaBoosting technique. Use Classifier A as the first weak classifier, Classifier B as the second weak classifier, and Classifier C as the third weak classifier.

(a) Compute the weight of Classifier A. Show your work.

(b) Compute the weight of Classifier B. Show your work.

(c) Compute the weight of Classifier C. Show your work.

(d) If a test instance is classified as POSITIVE by A, POSITIVE by B, and NEGATIVE by C, how will it be classified by the combined classifier? Briefly justify your answer.

(e) If a test instance is classified as POSITIVE by A, NEGATIVE by B, and POSITIVE by C, how will it be classified by the combined classifier? Briefly justify your answer.

(f) If a test instance is classified as NEGATIVE by A, POSITIVE by B, and POSITIVE by C, how will it be classified by the combined classifier? Briefly justify your answer.

2. PAC learning. (10 pts)

Assume in a learning problem where each instance is some integer in the set X={1, 2, …, 49, 50}, and where each hypothesis $h \in H$ is an interval of the form $a \leq x \leq b$, where a and b can be any integers between 1 and 50 (inclusive) as long as $a \leq b$. A hypothesis labels instance x positive if it falls into the interval and negative outside.

(i) How many distinct hypotheses are there in H?

(ii) How many training examples suffice to assure with probability 0.95 that any learner will output a hypothesis whose true error is at most 0.1?

3. VC dimensions. (25 pts)

What's the VC-dimension for the following two hypotheses? Explain your answer.

A. Intervals in R. The target function is specified by an interval, and labels any examples positive iff it lies inside the interval.

B. Axis aligned rectangles in the plane R^2. The target function labels examples positive iff it is inside the rectangle.

4. Programming (bagging). (35 pts)

In this programming assignment, you will use bagging with some classifiers to see if there is any performance gain compared to using a single classifier. You may try different numbers of bags to evaluate the impact. Write a short report about your findings (including setup, results, discussion).

You have several choices for this problem.
(1) You can use the classifier you implemented in previous homework assignments (either decision tree or perceptron, or both).
(2) You can use any existing toolkits for the base classifiers.
(3) You can use the data sets in previous homework assignments, or data sets from elsewhere (e.g., UCI data)

What to submit for the programming assignment?

- your report
- your code and readme file for the bagging part. You don't need to submit the base classifier implementation. Please make sure your readme file clearly describes how bagging is done. It is likely that you will submit some wrapper scripts (to train multiple classifiers, and combine their system output). There should be a clear description about how you did your experiments.