

Homework Assignment 3

Due date: Thu 11/3 (in class)

1 Introduction

In this assignment we review and apply some of the measures for evaluating hypotheses.

2 ROC Curves

In this part we will slightly modify the code from assignment 2 in order to draw ROC curves. The idea is to plot the performance of the learned hypothesis if we apply different thresholds (instead of zero) after learning is completed.

In particular, training of GD and EG should be done exactly as in assignment 2 but during testing we need more information. When testing on example x we need to record its true label as well as the score given by the learned model. The score for GD is $\vec{w} \cdot \vec{x}$ and the score for EG is $w^{\vec{P}} \cdot \vec{x} - w^{\vec{N}} \cdot \vec{x}$. Namely, we are simply taking the values calculated before applying the sign function.

Now, we can sort the examples according to the scores they get. If we vary the threshold from $-\infty$ to ∞ we get different true positive and false positive rates as potential behaviors of the learned model. Each such threshold provides one point on the ROC curve. Of course we only have a finite number of interesting boundaries since the test sample is finite.

Extend your code to provide such a testing facility; one way to do this is to add a second “test” procedure that prints out the numbers for the ROC curve instead of the accuracy.

Note: please consult the article by Provost, Fawcett and Kohavi, especially the discussion in section 3.2. We are not drawing average curves but simply focusing on one test sample.

Run your code for GD and EG on datasets DT.800.train, ST.800.train, sport.500.train (and appropriate test files) and *draw the ROC curves for each dataset*. Use fixed values of $\eta = 0.1$ and 3 iterations in these experiments. Do we get any dominating models?

Please submit the relevant code and the ROC curves for each dataset with a brief discussion.

3 Critic of Assignment 1

In assignment 1 you were asked to pick a method for generating attributes for text data and for picking the best N attributes according to this method. We then used this method to generate training files for Weka and reported back average accuracy in cross validation. Two popular methods were to (A) pick words by frequency of occurrence or to (B) pick words by their correlation with the label.

Professor Skeptic claims that using method B in this framework is likely to give over-optimistic values for the accuracy, but method A does not suffer from this problem.

(1) Explain why this claim is correct. (2) Propose a method to use B with cross validation in a way that avoids this problem.

4 Confidence Intervals

- You are given a hypothesis for a particular domain, test it on 100 independent examples and observe 93% accuracy. Calculate a $N = 0.99$ confidence interval for the hypothesis. Use the two solutions given in class to the problem of unknown variance to check the effect on the size of the intervals.

- Imagine you ran your favorite algorithm on a new dataset using a 10 fold cross validation scheme and got the following accuracies in the folds: 0.88 0.75 0.79 0.80 0.91 0.83 0.77 0.79 0.82 0.85. Use the formulation of the T confidence interval to calculate two intervals for the average accuracy, using confidence of 0.99 and 0.9 respectively.

5 Submitting Your Assignment

Please Submit hard copies of all the above in class.