

CS 410/510
Machine Learning
Spring, 2007

Homework 4: Assessing and Comparing Learning Algorithms

Due Thursday, May 3.

For this homework you will compare decision tree induction (via C4.5) with Naive Bayes.

Below are the steps you need to perform. Show your work in making these calculations.

I. Assessing the error of a single hypothesis.

Choose the error rates (on the test set) from one of your experiments with C4.5 and from the corresponding experiment with Naive Bayes. In this write-up I'll refer to these as $Err_{C4.5}$ and Err_{NB} , respectively.

- (a) Give 95% two-sided confidence intervals on both $Err_{C4.5}$ and Err_{NB} .
- (b) Give the 95% one-sided confidence interval on $Err_{C4.5}$ and Err_{NB} .

II. Comparing the error of two hypotheses.

- (a) Give a two-sided 95% confidence interval on the observed difference in error rate, $\hat{d} = |Err_{C4.5} - Err_{NB}|$.
- (b) Give the percentage confidence with which you can accept that the hypothesis with observed lower error rate has true lower error rate than the other hypothesis.

III. Comparing two learning algorithms

Put together a data set of spam and ham instances, with approximately equal numbers of spam and ham. You can use any of the examples of spam and ham from the previous homeworks.

Run 10-fold cross-validation on your data, for both C4.5 and your Naive Bayes algorithm.

- (a) Calculate \bar{d} as described in class. Give the data (i.e., ten sets of error rates for C4.5 and Naive Bayes) you used to calculate this value.
- (b) What is the two-sided 95% confidence interval for \bar{d} ?
- (c) Give the percentage confidence with which you can accept the hypothesis that the algorithm with lower average observed error rate has lower true error rate than the other algorithm.

IV. Summarize. Write a paragraph or two summarizing your results.

Here is what you need to turn in: **Hard copy:** Hand in a hard-copy of your (computer formatted, not handwritten) write-up as described above. That's it!