

CS 410/510
Machine Learning
Winter, 2006

Homework 5, part IV: Test significance of your results.

IV. Test significance of your results.

(a) Given your results, what is the approximate probability that the hypothesis encoded by your naive Bayes classifier has a higher true accuracy than the hypothesis encoded by your decision tree? (See Section 5.5.1 in the textbook.)

(b) Perform a paired t -test to compare the naive Bayes classifier and C.4.5. (The procedure below follows Section 5.6 in the textbook.)

1. Combine all the training and test data (spam.data and spam.test), and use this large data set, D_0 , to perform k -fold cross-validation on each algorithm, as described below and in Table 5.5 in the textbook. In particular, D_0 will have 4601 examples. Let $k = 6$.

(a) To make $|D_0|$ divisible by 6, discard 5 examples at random, 2 spam and 3 non-spam. Now $|D_0| = 4596$. Partition D_0 into $k = 6$ disjoint subsets, T_1, T_2, \dots, T_6 , of 766 examples each. In each of these, maintain approximately the same proportion of spam and non-spam examples as was in the original D_0 (39.4% spam).

(b) For each algorithm L , and for $i = 1$ to k , do

- Run L using as training data all the T_j except T_i . Let h_1 be the resulting naive Bayes classifier, and h_2 be the resulting decision tree.
- Let $\delta_i = error_{T_i}(h_1) - error_{T_i}(h_2)$. Record δ_i .

2. Record

$$\bar{\delta} = \frac{1}{6} \sum_{i=1}^6 \delta_i.$$

3. This term, $\bar{\delta}$, is an estimator for the expected true difference between these learning algorithms,

$$E_{S \subset D_0}[error_D(NaiveBayes(S)) - error_D(C4.5(S))].$$

Calculate the estimate of $\bar{\delta}$'s standard deviation, $s_{\bar{\delta}}$, using

$$s_{\bar{\delta}} = \sqrt{\frac{1}{k(k-1)} \sum_{i=1}^k (\delta_i - \bar{\delta})^2}.$$

4. Now give the approximate 95% confidence interval of $\bar{\delta}$:

$$\bar{\delta} \pm t_{95,5} s_{\bar{\delta}},$$

where $t_{95,5}$ is the “Students t ” statistic that corresponds to the z_{95} that we used for the normal distribution. We use 5 instead of 6 since with 6 samples of δ_i for each algorithm, there are 5 degrees of freedom. From a table of the Student’s t distribution, $t_{95,5} = 2.015$.

5. Assuming that $\bar{\delta}$ ’s distribution is symmetric about its mean, use a method similar to that in Section 5.5.1 in the textbook to calculate the approximate probability that

$$E_{S \subset D_0}[\text{error}_D(\text{NaiveBayes}(S)) - \text{error}_D(C4.5(S))] > 0.$$

You will need a table (or on-line applet) giving the Student’s t distribution. A handy applet for calculating t is at:

<http://math.uc.edu/~brycw/classes/148/tables.htm>

6. Report the details of your work (including the δ_i and $\bar{\delta}$ values) in your writeup.