Trevor Slaton (tms45)
Joseph Vokt (jpv52)

1 Model Selection and Validation

(a) S_{train} accuracy: 22/24. S_{test1} accuracy: 8/10. S_{test2} accuracy: 16/20.

(b) S_{train} accuracy: 24/24. S_{test1} accuracy: 8/10. S_{test2} accuracy: 15/20.

(c) Let $m = \frac{(n_{01}-n_{10})^2}{n_{01}+n_{10}}$ where n_{01} corresponds to the number of cases which were missclassified by the decision tree hypothesis and not the linear hypothesis, and n_{10} corresponds to the number of cases which were missclassified by the linear hypothesis but not the decision tree hypothesis. For S_{test1}, $n_{01} = 1$ and $n_{10} = 1$, so $m = 0$. Thus we can't say that one model generalizes any better than the other because m is much less than the critical value 3.84.

(d) For S_{test2}, $n_{01} = 3$ and $n_{10} = 2$, so $m = \frac{1}{5} = .2$. Similarly, we can't say that one model generalizes any better than the other because m is much less than the critical value 3.84.

(e) The accuracy for S_{test1} is the same for both the linear hypothesis and decision tree. The χ^2 statistic for the McNemar test for S_{test1} comparing the two hypotheses confirms that there is no significant difference in how well they generalize. The accuracy for S_{test2} is slightly higher for the linear hypothesis than the decision tree. However, the χ^2 statistic for the McNemar test for S_{test2} comparing the two hypotheses reveals that we are not justified in claiming that one hypothesis generalizes better than the other.

Although the results for the two McNemar tests were different, the conclusion is the same. Unlike for S_{test1}, the decision tree hypothesis does separately misclassify more examples in S_{test2} than does the linear hypothesis, but the difference between their misclassification rates is still shown by the McNemar test to be insignificant. In order to determine that results are significant at a confidence level of 95%, the χ^2 statistic for the McNemar test must exceed the critical value 3.84, which mathematically requires both a sizable discrepancy between the separate misclassifications n_{01} and n_{10} and a large total number of misclassifcations (e.g. $n_{01} = 40, n_{10} = 60$). In both cases for this question, we have neither.

2 Model Averaging with Decision Trees

(a)

(b)

(c)

(d)

3 Text Categorization with Decision Trees

(a)

(b)

(c)

(d)

(e)

(f)