This details three of the possible solutions to question 4 in the homework, comparing two hypothetical decision trees. In your solution, you should have compared all the trees from parts 2 and 3 with that from part 1c. All three methods were accepted, providing you clearly explained what you were computing.

We take some typical value for the performance of two decision trees:

Say T_1 make 44 errors on the test data, which has 101 examples, and T_2 makes 37 errors. Lets also say that there are 9 examples where T_1 makes an error but T_2 does not, and 2 examples where T_2 makes an error and T_1 does not.

Solution 1

We can use equation 5.1 on page 132 of Mitchell, assuming the errors are normally distributed.

$$error_D(h) = error_S(h) \pm z_n \sqrt{\frac{error_S(h)(1 - error_S(h))}{n}}$$

Using table 5.1 on the same page in the book, we see that to get a 95% confidence interval, we take $z_n = 1.96$. So on T_1:

$$error_D(h) = \frac{44}{101} \pm 1.96 \sqrt{\frac{(44/101)(1 - 44/101)}{101}} = 0.435 \pm 0.097$$

Similarly, for T_2, we get:

$$error_D(h) = \frac{37}{101} \pm 1.96 \sqrt{\frac{(37/101)(1 - 37/101)}{101}} = 0.366 \pm 0.094$$

Hence we have:

$$0.338 \leq error_D(T_1) \leq 0.532$$
$$0.272 \leq error_D(T_2) \leq 0.460$$

In this case, we see that the range for values for the error of T_1 contains $error_S(T_2)$, and similarly the range of values for the error of T_2 contains $error_S(T_1)$. If either range contains the estimate of the error for the other tree, then we know the hypotheses are not significantly different. Hence neither tree is better than the other with 95% confidence.

Solution 2

We can also perform the paired t-test, as follows:

The error $E_1 = error_S(T_1)$ is $\frac{44}{101} = 0.435$. The error $E_2 = error_S(T_2)$ is $\frac{37}{101} = 0.366$. Hence the difference between the errors is $E_1 - E_2 = 0.069$.

Since we have a large number of training examples (more than 30), we can estimate the variance on the error as follows:

$$var = \frac{E_1(1-E_1)}{n_1} + \frac{E_2(1-E_2)}{n_2} = \frac{0.435 \times 0.564}{101} + \frac{0.366 \times 0.634}{101} = 0.00473$$

Then, we get a confidence interval of the range in the difference between the true errors of:

$$(E_1 - E_2) \pm z_n\sqrt{var} = [-0.0658 \;\; 0.204]$$

The difference in errors could be positive (indicating T_2 is better than T_1) or negative (indicating T_1 is better than T_2). Hence we cannot say for sure which tree is better and thus the trees are not significantly different.

Solution 3

Finally, we can perform the Binomial Sign Test, also called McNemar's Test:

We know that there are 9 examples where T_1 makes an error but T_2 does not, which means $d_1 = 9$. Similarly, there are 2 examples where T_2 makes an error but T_1 does not, so $d_2 = 2$.

Given our hypothesis that both trees are equally accurate, the probability that T_2 wins over T_1 at least d_1 out of $n = d_1 + d_2$ times is:

$$\begin{aligned} P(D_2 \leq d_2 | p = 0.5,\ n = 11) &= \sum_{i=0}^{d_2} \frac{n!}{i!(n-i)!} p^n \\ &= \sum_{i=0}^{2} \frac{10!}{i!(10-i)!} 0.5^{10} \\ &= 0.055 \end{aligned}$$

We want a 95% confidence interval, i.e. we want the probability that the trees are equally good to be less than 5%, or equivalently that either of the trees is better than the other to be less than 2.5% (since there are two tails in the probability distribution). However, the above tells us there is a chance of 0.055 = 5.5% that T_2 appears to be better than T_1 just at random, which is greater than 2.5%. Hence we cannot say that the trees are significantly different.