Comments on Quiz 3 for COMP 135, Spring 2017   - Average = 24.8/40

1) (3.8/10) When using k-fold cross-validation with C4.5 and Reduced Error Pruning, k decision
trees are built each using a different part of the training set to evaluate which nodes of the
tree to prune. How does C4.5 find the consensus tree using the whole training set?

It’s not possible to average the k trees, since they may have different structures. It’s not justified to use
the use the tree with the lowest error, for the same reason that we need to use the Bonferroni correction
when doing multiple statistical tests. The right thing to do is to use k-fold cross-validation to set some
parameter(s), then run C4.5 using that result. With Reduced Error Pruning the most likely parameter is
the number of nodes in the tree, so stop growing the tree when that number of nodes is reached.

2) (4.9/10) In the previous scenario, explain whether it would be necessary to hold out a
separate set of labeled examples for validation.

Since the whole training set was used for building the tree including setting a parameter (in this case the
number of nodes), a separate validation set is needed to get an accurate estimate of the error.

3) (7.4/10) What is stratified cross-validation? Explain in terms of ten-fold cross-validation of a
training set with 200 examples, 120 of which are positive and 80 are negative. There are
two attributes: attribute A has values “T” (70 examples) and “F” (130 examples), and
attribute B was values “Y” (50 examples), “M” (80 examples), and “N” (70 examples).

Stratified cross-validation runs the algorithms k times, each time holding out a disjoint 1/k of the data set
for validation and training on the remaining data. Stratification means ensuring that the class distribution
on all k validation sets are the same. In this case, that would mean that each fold has 20 examples, 12 of
which are positive and 8 of which are negative. The validation sets are not stratified by attribute values.

4) (8.7/10) Why is Laplace smoothing (or a variant) used to estimate the parameters of Naive
Bayes?

If Laplace smoothing is not used, and some attribute doesn’t occur in a class, then the estimate of the
conditional probability of that attribute in that class will be zero. This means that any product using that
conditional probability will be zero, invalidating many of the estimates.