PubH 7475/8400 Homework 3 (Spring 2007)

Due on April 10, 2007

1. Apply 1) a fully grown tree; 2) an optimally pruned tree; 3) bagging; 4) random forest; 5) boosting to one of the following two data sets: (50 pts)

    - NCI (NCI60) microarray data: there are $p = 6830$ predictors (i.e. genes). By ignoring a few classes with only few samples, we only consider 5 CNS, 9 renal, 7 breast, 9 NSCLC, 8 melanoma, 6 ovarian, 6 leukemia and 7 colon tumor samples. The predictors are in a file called Data, and the class labels in Info. *Use LOOCV to evaluate a classifier.*

      As mentioned in class, this dataset is one of the three used in Dudoit et al (JASA, 2002, p.77-87) to evaluate several classification methods.

    - Spam data: there are $p = 57$ variables (in the Data file) to distinguish two classes, spam (coded as 1) and email (coded as 0) (as the last variable in the Data file). There are total 1813 spams and 2788 emails. As done in the textbook (p.262-263), we take a random subset with 3065 observations as a training set, and the remaining ones as a test set. *Use the test set to evaluate a classifier.*

      It is desirable to use the same random seed you used earlier so that you can compare the results.

    For each of the three ensemble methods, you can use the tree (or any other classifier) as the base learner; you can choose any reasonable tuning parameters.

    The data and some information on the data are available from the Data link on our course homepage.

    Please attach your computer program and relevant output.

2. **(PubH 8400)** Justification for exponential loss: (20 pts)

    (a) Prove equation (10.16) on p.307:

    $$f^*(x) = \arg\min_{f(x)} E_{Y|x}(\exp(-Yf(x))) = \frac{1}{2}\log\frac{Pr(Y=1|x)}{Pr(Y=-1|x)}.$$

    (b) Is $f^*(x)$ also the population minimizer of the binomial negative log-likelihood as discussed on p.308? Show your evidence.