MATH/STAT 4450/8456 Homework 3

Due date: February 21, 11:59 pm

1. (3 points) The following equation from lecture notes added $\sum_{j=1}^{J} \frac{n_j}{n}$ to Equation (8.6) in the textbook. Explain the role of the additional summation over $j = 1, 2, \cdots, J$. Note that the textbook uses m to denote the mth region, but the lecture notes use j.

$$\text{Gini Index} = \sum_{j=1}^{J} \frac{n_j}{n} \sum_{k=1}^{K} \hat{p}_{jk}(1 - \hat{p}_{jk})$$

2. (12 points) Textbook exercises.

   (a) (Chapter 8.4) 4, 5, 9

3. [Additional problem only for graduate students. Undergraduate students will receive 2 extra credits if solving the additional problems correctly.] (5 points)

   The spambase data can be found from http://thinktostart.com/data/data.csv, and the column names are provided at http://thinktostart.com/data/names.csv. Consider making a linear combination of three models: rpart, C5.0, and bagging. Do 6-fold cross-validation to choose a weight vector $\mathbf{w} = \{w_{rpart}, w_{C50}, w_{bagging}\}$, from the set of vectors with entries in $\{.1, .2, .3, .4, .5, .6, .7, .8, .9\}$ summing to 1. After finalizing the weight vector, compare the ensemble model with three independent models using the validation set approach. You can use the following code to read the data:

```
spam = read.csv("http://thinktostart.com/data/data.csv",header=FALSE,sep=";")
names(spam) = read.csv("http://thinktostart.com/data/names.csv",
                        header=FALSE,sep=";",stringsAsFactors=FALSE)$V1
spam$y = factor(spam$y)
```