Statistics 315B – Spring 2016
Homework 1

Due 4/28/16 (11:59pm)

Submit via Gradescope

We recommend using R for the computational problems of this Homework. R for Unix is installed on the Leland systems. The Windows and Mac users can download the R package from *http://cran.r-project.org*. In the class web page there are posted some R/Splus Tutorials. Splus is a wrapper around R and most of the Splus commands work in R. The data sets Income_Data, Income_Big and Housetype_Data along with the documentation Income_Info and Housetype_Info can be found in the class web page in the Data section.

For this homework users should type > *library(rpart)* at the R prompt. Help files are available. The graphical help system can be invoked with > *help.start()*. Alternately a specific help file can be directly invoked: e.g > *help(rpart)* will display the rpart help file.

To plot tree classifiers in a nice way, use the function *post.rpart*. It creates a postscript filename.ps and places it into your working directory (like hw1, for example). You can then view the tree using ghostview (filename.ps) or by simply printing out the file filename.ps. Alternatives are the *plot.rpart* and *text.rpart* functions which will produce a tree diagram in the graphics window

Please present the tree in the output plots with a decent number of terminal nodes (say at most 10), if the optimal pruned tree is too big to properly fit on one page. To set control parameters for the function rpart, use the function *rpart.control*. Look up the help window for information on how to use these functions and what exactly they are doing.

Hint: the function *printcp* lets you see the full sequence of pruned trees together with their complexity parameters (cp), training errors (rel error) and cross-validation estimates of errors (xerror). Note that for easier reading, the error columns have been scaled so that the first node has an error of 1. The actual error of the first node is also given in the output of printcp.

(1) Data Mining Marketing. The data set Income_Data represents an extract from a commercial marketing database. The goal is to fit a regression tree to predict the annual income of a household from 13 demographic attributes and interpret the results. Note that some of the variables are categorical: be sure to mark them as such using the R function *as.factor*, before running *rpart*. Use the RPART implementation of the decision tree algorithm to fulfill this task. Write a short report about the relation between the annual income and the other demographic predictors as obtained from the RPART output and answer the following questions:

(a) Were surrogate splits used in the construction of the optimal tree you obtained? What does a surrogate split mean? Give an example of a surrogate split from your optimal decision tree. Which variable is the split on? Which variable(s) is the surrogate split on?

(b) Using your optimal decision tree, predict the annual household income of your household.

(2) Multi-Class Classification: Marketing Data. The data set Housetype_Data comes from the same marketing database that was used for problems (1) and 2. Refer to the documentation Housetype_Info for attributes names and order. From the original pool of 9409 questionnaires, those with non-missing answers to the question "What is your type of home?" were selected. There are 9013 such questionnaires.

The goal in this problem is to construct a classification tree to predict the type of home from the other 13 demographics attributes. Give an estimate of the misclassification error of an optimal tree. Plot the optimal tree if possible (otherwise plot a smaller tree) and interpret the results.

(3) What is the definition of the target function for a given problem. Is it always an accurate function for prediction. Why/why not.

(4) Is the empirical risk evaluated on the training data always the best surrogate for the actual (population) prediction risk. Why/why not. In what settings would it be expected to be good.

(5) Why can't the prediction function be chosen from the class of all possible functions.

(6) Explain the bias-variance trade-off.

(7) When would you expect that the categorical variable splitting trick not to provide the optimal split into two subsets.

(8) 7. Why not choose surrogate splits to best predict the outcome variable y rather than the primary split.

(9) Consider the regression tree model

$$F(\mathbf{x}) = \sum_{m=1}^{M} c_m \, I(\mathbf{x} \in R_m)$$

where $\{R_m\}_1^M$ represent disjoint subregions of the space of all predictor variable $\mathbf{x}$–values. Show that the values of c_m that minimize the squared–error risk score criterion

$$\sum_{i=1}^{N} [y_i - F(\mathbf{x}_i)]^2 \tag{1}$$

are given by

$$\hat{c}_m = \sum_{i=1}^{N} y_i \, I(\mathbf{x}_i \in R_m) \Bigg/ \sum_{i=1}^{N} I(\mathbf{x}_i \in R_m),$$

that is the mean on the training data outcome values for the observations within each region.

(10) Show that the improvement in squared–error risk (1) when one of the regions R_m is split into two daughter regions, $R_m \to R_l \cup R_r$ can be expressed as

$$\frac{n_l \, n_r}{n} (\bar{y}_l - \bar{y}_r)^2$$

where n is the number of observations in the parent R_m, n_l, n_r the numbers respectively in the left and right daughters, and $\bar{y}_l$, $\bar{y}_r$ are the means of the outcome variable y for observations in the respective daughter regions.

(11) Derive an updating formula for calculating the change in the improvement in prediction risk as the result of a split when the split is modified by one observation changing sides.

(12) One strategy for predicting an observation missing the value of a splitting variable is to treat the current (internal) node as terminal to make the prediction. Another is to send the observation to the daughter node which contained the most training data. Compare these two strategies with the surrogate splitting method in terms of expected benefit.

(13) Consider a missing value strategy for decision trees different from that based on surrogate splits. With this alternative strategy, the value x_j = "missing" is simply considered to be an additional categorical value assumable by each predictor variable x_j. For categorical variables this results in no change to the splitting strategy. For variables with orderable values three way splits must be considered. First missing values are split from the non missing ones, then the optimal split is made on the non missing values. These three–way splits compete with the splits on the other variables in the selection process at each node. Discuss the relative advantages/disadvantages of his strategy relative to the surrogate splitting strategy. Does this strategy allow a surrogate effect by encouraging variables within highly correlated sets to substitute for each other in the prediction process? If so how (by what mechanism)? Can this strategy be used if there are no missing values in the training data? If not, what can be done so as to enable the tree to predict with missing values in future data.