Statistics 315B - Spring 2020
Homework 1

Due 5/3/20 (11:59pm)

Submit via Gradescope

The statistical package R is to be used for the computational problems of this Homework. The R package can be downloaded from *http://cran.r-project.org*. There are many R tutorials on the web. The data sets *age_ stats315B.csv* and *housetype_ stats315B.csv* along with the corresponding documentation *age_ stats315B.txt* and *housetype_ stats315B.txt* can be found in the class web page in the Files section. The data can be read into R using the *read.csv* command with *header=TRUE*.

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. (15) Data Mining Marketing. The data set *age_ stats315B.csv* represents an extract from a commercial marketing database. The goal is to fit a regression tree to predict the age of a person 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 age 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 your age.

2. (15) Multi-Class Classification: Marketing Data. The data set *housetype_ stats315B.csv* comes from the same marketing database that was used for problem 1. Refer to the documentation *housetype_ stats315B.txt* 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. (5) What are the two main reasons why a model that accurately describes the data used to build it, may not do a good job describing future data?

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

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

6. (5) 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.

7. (10) Suppose the loss for an incorrect classification prediction is the same regardless of either the predicted value c_k or the true value c_l of the outcome y. Show that in this case misclassification risk reduces to the classification error rate. What is the Bayes rule for this case in terms of the probabilities of y realizing each of its values$\{Pr(y = c_k)\}_{k=1}^{K}$? Derive this rule from the general (unequal loss) Bayes rule, for this particular loss structure $L_{kl} = 1(k \neq l)$.

8. (5) Does a low error rate using a classification rule derived by substituting probability *estimates* $\{\widehat{\Pr}(y = c_k)\}_{k=1}^{K}$ in place of the true probabilities$\{\Pr(y = c_k)\}_{k=1}^{K}$ in the Bayes rule imply accurate estimates of those probabilities? Why?

9. (5) Explain the bias-variance trade-off.

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

11. (10) Consider the regression tree model

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

where $\{R_m\}_{m=1}^{M}$ 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

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

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

12. (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.

13. (10) 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.

14. (5) Suppose training data $T = \{y_i, \mathbf{x}_i\}_{i=1}^{N}$ and consider the data based target function estimate

$$\hat{f}(\mathbf{x}) = \arg \min_{g(\mathbf{x}) \in F} \sum_{i=1}^{N} [y_i - g(\mathbf{x}_i)]^2.$$

Here F is a restricting class of functions. By enlarging F (reducing the restriction on $g(x)$) one can obtain better a fit to the training data. That is

$$\widehat{mse} = \frac{1}{N}\sum_{i=1}^{N}[y_i - \hat{f}(\mathbf{x}_i)]^2$$

will be smaller. Is this always a good idea? Will it necessarily lead to better expected mse on future data? Why or why not? Conversely, is it always better to reduce the size of F (increasing the restriction on $g(\mathbf{x})$), thereby fitting the training data less well? Why or why not?

15. (5) The recursive partitioning strategy described in class for building decision trees uses two-way (binary) splits at each step. This is not fundamental, and one could envision multi-way splits of each non-terminal node creating several (rather than two) daughter regions with each split. What would be the relative advantages and disadvantages of a such a multi-way splitting strategy?

16. (5) As described in class, the recursive binary splitting strategy considers splits that only involve a single input variable. One could generalize the procedure by including "linear combination" splits of the form

$$\begin{array}{ll} \text{if } \sum_{j \in \text{numeric}} a_j x_j \leq s & \text{go left} \\ \text{else} & \text{go right} \end{array}$$

where the sum is over the numeric input variables only. The values of the coefficients $\{a_j\}$ and the split points are (jointly) chosen to optimize the splitting criterion, which is same as that used for single variable splits. What would be the advantages and disadvantages of including such linear combination splits in the tree building strategy?