MATH 80619 Advanced statistical learning

Assignment #2 – chapter 3

Specific instructions

- In order to ensure reproducibility of your results, your assignment must be written using R Markdown or R Sweave. Your R code, output and comments should be included in the main document.
- You can submit the html or pdf output file on Zone Cours in the section "Remise de Travaux".
- Deadline is one week after we finish chapter 3, before the beginning of the class (10 am).

Question #1

Draw an example (of your own invention) of a partition of two dimensional feature space that could result from recursive binary splitting. Your example should contain at least six regions. Draw a decision tree corresponding to this partition. Be sure to label all aspects of your figures, including the regions R1,R2, . . ., the cutpoints, and so forth. Your result should look something like slide #6 of the notes. You can draw the figure by hand and include a picture of it in the .Rmd file.

Question #2

Suppose we produce ten bootstrapped samples from a data set containing 2 classes. We then apply a classification tree to each bootstrapped sample and, for a specific value of X, produce 10 estimates of P(Class =1 |X):

0.1, 0.15, 0.2, 0.2, 0.55, 0.6, 0.6, 0.65, 0.7, and 0.75.

There are two common ways to combine these results together into a single class prediction. One is the majority vote approach discussed in the notes. The second approach is to classify based on the average probability. In this example, what is the final classification under each of these two approaches, assuming a cutoff of 0.5 (explain your answer) ?

Question #3

For this question, you will analyze the dataset *Pima Indian Diabetes*. This data set is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the study is to predict whether or not a patient has diabetes, based on

several clinical measurements included in the data set. The original data set can be found on `mlbench` library and the description of the variables can be found here: https://rdrr.io/cran/mlbench/man/PimaIndiansDiabetes.html
In particular, all subjects are females, at least 21 years old, and of Pima Indian heritage. The data set consists 768 observations with 8 predictor variables and 1 target variable (*diabetes*).

For the following questions, use the training and test datasets provided in the files `pimadiabetes_test.csv` and `pimadiabetes_train.csv`. **Compute the global error rate, false positive rate and false negative rate on the test set.**

a) Grow a single tree using the CART algorithm based on a cutoff of 0.5 for prediction using i) the majority vote technique and ii) the average probability technique (see question #2)
b) Repeat question a) using an optimal cutoff minimizing the global error rate assuming an equal cost between false positive and false negative. In general, can you compare the error rates computed using different gain matrices (answer yes or no, and explain)?
c) Repeat questions a) and b) using a conditional tree
d) Grow a random forest with B=500 trees using the CART algorithm and a cutoff of 0.5
e) Grow a random forest with B=500 trees using the conditional tree algorithm and a cutoff of 0.5
f) Repeat questions d) and e) using the bagging approach
g) Compare the results. Compute the variable importance measures based on the best model.

Question #4

Create 2 questions of the type “multiple choice questions” on the material seen in chapter 3.
- Your questions should have 4 possible answers, 1 true answer and 3 false answers.
- The level of the question should not be too easy (i.e. wrong answers should not be obvious).
- Solutions - with explanations for the right and wrong answers- need also to be included.

Your mark for this question will be based on the correctness of your question/answers, and on the depth of your question and answers (especially the wrong ones).