Statistical Learning (MT7049) - Project 2

Instructions: *This project consists of 2 tasks that should be solved individually. You are free to use any programming package in this project.*

The solution should be submitted at the course webpage in a single .pdf file with your source code attached as appendices. Your source code should include clear comments and documentations to describe what you are evaluating.

TASK 1

This task contains some mathematical and conceptual exercises for you to get more familiar with boosting trees and random forest.

a) Referring to Table 10.2 in the course book, show that the gradient of the deviance as loss function is given by the expression in the last row of the table.

b) Complete exercise 15.1 in the course book.

c) Complete exercise 15.4 in the course book.

d) In the construction of random forest, the hyperparameter *m* (see Algorithm 15.1 in the course book) needs to be fixed. Discuss how the choice of *m* (i.e., small or big *m*) affects the bias and variance of the resulting random forest model. Please write your reasoning concisely and to-the-point.

TASK 2

This task allows you to experience and explore the use and performance of gradient boosting trees for binary classification. This task uses the “Spam” data in the course book. To download the data, go to the webpage of the course book (https://web.stanford.edu/~hastie/ElemStatLearn/), and then click “Data” on the left panel to find the data and follow the instruction to use them.

a) To begin with, pick up the correct function/package in R, Python, etc., to perform binary classification using gradient boosting trees. Describe which function and package are picked. Explain which loss function is used and why it is suitable for the task. List and explain the choices of all other parameters and settings required by the function to train the boosting trees.

b) Write a code to construct gradient boosting trees with different number of terminal nodes (i.e., $m = 2$ (stump), 5, 10, 20 and 50 nodes). Attach your source code at the end of the report.

c) Construct the cross-validation (CV) plot, i.e., CV error versus number of boosting trees M, with 10-fold CV. (Hint: You can simply evaluate the errors for M = 10, 20, 30, …) Standard error bars should be included in the plot so that you can use the one-standard-error rule to identify the optimal model.

d) Referring to the CV plot in part c), discuss the performance of the boosting trees with different values of m and M, especially on its connection to the concept of bias-variance tradeoff. Please writing your reasoning concisely and to-the-point.