MATH 80619 Advanced statistical learning

Assignment #3 – chapter 4

Due March 8th, 2021

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".

Question #1

This question deals with the L2-boosting algorithm presented in slide 22 of the course notes. This algorithm is equivalent to the Gradient boosting algorithm when a L2 loss function is used.
In slide 29, this algorithm was applied to the Ames data, using the function **`gbm`**. The goal of this question is to look at the different components of the object **`gbmgc`** that was created in slide 29.

1) By referring to the algorithm presented in slide 22, explain what the value of the **`initF`** component represents

2) What does the component **`fit`** represent ? Use it to compute the MSE in the training sample

3) What does the component **`train.error`** represent ? Explain the link with the MSE value computed in 2)

4) We mentioned in slide 32 that gbm allows to estimate the error with cross validation. By default (which was the case in the analysis presented in the notes), no cross-validation is performed. Use the option **`cv.folds`** in the function to perform a 10-fold CV and use the results in the component **`cv.error`** to plot the error as a function of the iterations (trees). What can you say about the optimal number of trees and comment about our choice to use 100 trees in our analysis.

5) From the results obtained in 4), what is the final cross-validated MSE of the model ? Compare this value to the MSE computed in 2) and comment.

6) Using the cross validation results in 4), explain the difference between the component **`cv.fitted`** and the component **`fit`** from the analysis without the

cross validation. Use the **`cv.fitted`** to recompute the cross-validation error obtained in 5).

7) Draw a graph showing the 10-fold cross validation error of the model as a function of the number of trees (like you did in 4), for different values of the shrinkage parameter epsilon (use 0.01, 0.05, 0.1, 0.2 and 0.3). Comment your results and choose the optimal value of the shrinkage parameter.

Question #2

See part 2 of the assignment…

Question #3

Create 2 questions of the type "multiple choice questions" on the material seen in chapter 4.

- 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.
- **Note that your questions can be based on R outputs (in which case you need to provide the R output)**

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).

Question 4 - extra for those of you who didn't get all the points to question 1b of assignment 1 (cross validation code)

If you wish, you can redo this question and get +1 point on your assignment 1.