MATH 80619 Advanced statistical learning

Assignment #1 – chapter 2

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 2, before the beginning of the class (10 am).

Question #1

We have seen in the notes that the package glmnet doesn't have a function to automatically select the parameter α for elastic net regression.

a) Use the R library caret to simultaneously optimize the α and λ parameters by cross validation in elastic net regression. Compute the MSE and MAE of the Ames data using elastic net regression with optimal parameters.
b) Write your own code in order to simultaneously optimize the α and λ parameters by cross validation. Compute the MAE and MSE on the Ames data and compare your answer with a).
c) Compare the results in a) and b) with the other methods (see slide 83)

Question #2

In this exercise, you are asked to apply all the variable selection methods seen in class to the dataset "Music Origin" and compare results based on MSE and MAE:

- Ordinary Least square (without variable selection, used as a benchmark method)
- Stepwise, forward, backward algorithms based on AIC or BIC
- Ridge regression
- Lasso regression
- Elastic net with α=0.5, 0.2 and 0.8
- Elastic net with optimal α and λ parameters (use your code from Question #1)
- Relax Lasso

The data set is from UCI Machine Learning Repository. The dataset contains 1059 music tracks from 33 countries. For each track, there exists 68 columns of audio features and

the latitude information of the country of music origin. The audio features are named as (x1, x2, …, x68) and the latitude is denoted as y. The goal is to predict the music origin (variable y) based on the audio features. The train and test set data are provided in the files `music_origin_lat_test_set.csv` and `music_origin_lat_train_set.csv`. You are not expected to pre-process the data. Don't forget to comment the results you obtain.

Question #3

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

- 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).