Instruction for smoothing prediction project

February 21st, 2024

Task: Build a smoothing model with the data in *dtrain*, and make predictions on the data in *dtest*. Both *dtrain* and *dtest* can be loaded into R by command ”load(‘smooth.Rdata’)”. The description of the variables can be found in "variable_description.txt". The prediction error is evaluated by **RMLSE** (Root-Mean-Squared-Logarithmic-Error), which is the square root of the mean squared error on the natural log scale.

Prediction: For fairness, we should restrict ourselves to predicting with the smoothing methods learnt in class, that is, **local regression** and **splines** (cubic, B, natural cubic, smoothing). More specifically, the R packages listed on pages 3, 4 and 9 of the slides 08_Smoothing_Methods are allowed for prediction.

Preprocessing: Simple imputation of missing values is encouraged. R packages for imputation are allowed, but please restrict to simple imputation methods (mean, median, linear model). Advanced imputation methods that use random forest, boosting etc. are NOT allowed.

If you are unsure whether a package or method can be used, please ask the instructor.

Deadline March 7, Thursday, 11 pm

1. Submit an R source file to Learn Dropbox folder "Project_Smoothing_Predict". The file should be named as "UW_ID.R", where UW_ID should be replaced by your UW ID. Suppose your UW ID is 20654321, the source file should be named as "20654321.R". This file should include a function called "*SmoothingModel*" used to predict the response on the test set. An example R source file with instructions can be found in "20654321.R". Fill in details in the first few lines of the source file and follow the instructions within. You can submit the source file many times before the deadline, and only the most recent one is kept.

   The source file should only contain the *SmoothingModel* function. The model fitting should be a single call to a smoothing function. No composite smoothing model allowed, such as the average of a spline and a local regression. All model fitting and prediction should be inside this function, and there should be no R code outside the *SmoothingModel* function except the library() function. Any other R commands outside the *SmoothingModel* function, such as **setwd**() and **load**(), are subject to penalty. If you need to define other helper functions, they can reside in the *SmoothingModel* function as well. When we source your file,

only the *SmoothingModel* function should be introduced into the R environment.

Your source file will be sourced by an evaluation R file as "evaluation.R" to compute RMLSE. Make sure you can run through "evaluation.R" with your source file without errors to avoid penalty. If errors occur and they are non-trivial, we may contact you through email to fix the errors. In such case, please respond within 24 hours to avoid penalty escalation.

The source file should only contain the final model and necessary preprocessing steps. Model selection, model checking/diagnostic and plotting should be conducted in the R Markdown file and not in the source file.

2. To give you some feedback, you should submit your prediction result to **Kaggle** site. The link can be found in the project folder along with other files. You are allowed to submit at most ***5*** predictions per day, and your prediction accuracy in the public fold will be made available to you immediately to guide your model building. A sample result file can be found as "solution_sample.csv".

Deadline March 8, Friday, 11 pm

1. Submit an R **Markdown** file to Learn Dropbox folder "Project_ Smoothing_Markdown", and submit the generated pdf to Crowdmark. The Crowdmark link will be sent to you. They should detail your preprocessing and model building process. It should include your ID, name, Kaggle submission details, and start with a summary paragraph summarizing your model building process. The file should be named in the format of "UW_ID.Rmd". An example Rmd file can be found in "20654321.Rmd", please use it as a template and fill out details.

Important

1. This project is an individual project. Discussions among students are allowed, but no sharing of code (R and Markdown) or report are allowed.
2. You need to submit **all three files** (the R source file, the R Markdown file and its generated pdf file) for your project to be graded.
3. Please use the latest R to ensure consistency of results across different computers, R version should be > 4.3.