AMS 597: Statistical Computing

Pei-Fen Kuan (c)

Applied Math and Stats, Stony Brook University

Project Description

- This is a group project (~ 4 members, 1 group with 5 members) except for those who opted to work individually.
- Group assignment is available on Brightspace.
- You are allowed to swap groups (it will be a one-to-one swap, i.e., you need to find another member in your new group to agree on swapping)
- Deadline for group swapping is Friday 03/22/2024 at 5PM EST.
- Email the instructor your new group, cc'ing the student who is willing to swap with you.

Project Description

- Your project involves creating an R package that performs the following tasks.
- Your R package will take as input a response variable `y` and matrix of candidate predictors/independent variables `X`, where each column is a predictor.
- Your package will work for both binary `y` and continuous `y` (for continuous case, it can be assumed to be normally distributed).
- The predictors `X` can be combinations of continuous, discrete and binary predictors.
- The number of predictors p can be very large (i.e., you should also consider the case where $p >> n$, n is the sample size).

Project Description

- Your package will implement the following models:
  - linear or logistic regression
  - ridge regression (for binary and continuous `y`)
  - lasso regression (for binary and continuous `y`)
  - elastic net (for binary and continuous `y`)
  - another machine learning model, you can choose between support vector machine, random forest or boosted trees (for binary and continuous `y`)
- You may import R packages glmnet, e1071 (for SVM), randomForest and xgboost (for boosted trees) and use the functions in these packages. You cannot use other special R packages.

Project Description

- If $p >> n$, your package will also have the option that allows users to pre screening for top K most "informative" predictors to be included in the model.
- Describe in your package tutorial page how your package chooses these informative predictors. You cannot use special R packages for this task. You can utilize functions in the base R package.

Project Description

- Furthermore, to improve robustness of model fit, your package will also have the option to perform "bagging" for linear, logistic, ridge, lasso and elastic net models.
- That is, for a chosen model, your package will perform sampling with replacement for the samples R times.
- Your package will return the final predicted values as the averages of these bagged models. Propose how you will average the bagged models and describe this in your package tutorial page.
- Your will write your own function implementing the bagging approach. Do not import special R packages.
- Additionally, your package will also return a "naive" variable importance score which counts the number of times each variable is selected in the bagging process.

Project Description

- Finally, your package will also have the option that allows users to choose if they want "ensemble" learning, that is fitting more than one models on the same dataset.
- For example, glmnet and random forest.
- If yes, the users will specify the types of models to fit, and your package will return a "combined" final result. Propose how you will combine the results and describe this in your package tutorial page.
- Your will write your own function implementing the ensemble.
- You will then wrap these up as an R package called `simpleEnsembleGroupX` where `X` is your assigned group number.

Project Description

- The R package has to be complete and contains a vignette describing how to use the R package.
- The R package is due May 03, 2024 at 5:00 PM.
- Submit your package as original source package (i.e., .tar.gz file) on Brightspace>Assignments>Project. Name your package `simpleEnsembleGroupX_version.tar.gz`, where `X` is your assigned group number, e.g., `simpleEnsembleGroup20_1.0-0.tar.gz`
- Version is generated automatically after you build your package successfully.

Project Description

- All students will submit the R package to Brightspace (i.e., although members in the same group will submit the same R package, I still require each student to submit the R package to Brightspace). Points will be deducted for students who do not submit the package to Brightspace before due time.
- One of the member will email the package to the instructor (peifen.kuan@stonybrook.edu) and cc the rest of the group. In the Subject header of the email, type AMS 597 Spring 2024 Group X Project", where X is your group number.

Project Description

- Some of the grading criteria include:
  - Can the R package be installed successfully?
  - Is the R package implementing the required method correctly?
  - Has it considered all possible scenarios?
  - Is the R package user friendly (vignette, help files, warning messages, sample data, sample code)?
  - What is the computational speed?

Project Description

- Some useful links:
- https://tinyheero.github.io/jekyll/update/2015/07/26/making-your-first-R-package.html
- https://hilaryparker.com/2014/04/29/writing-an-r-package-from-scratch/
- https://combine-australia.github.io/r-pkg-dev/
- http://kbroman.org/pkg_primer/
- http://kbroman.org/Tools4RR/assets/lectures/08_rpack_withnotes.pdf
- https://cran.r-project.org/doc/contrib/Leisch-CreatingPackages.pdf
- https://ourcodingclub.github.io/tutorials/writing-r-package/

Project Description

- Some useful links for incorporating existing R package into your R package
- https://kbroman.org/pkg_primer/pages/depends.html
- https://r-pkgs.org/description.html
- Or google keywords `import R package''`,depends R package''
- Some useful links (for Windows):
- https://www.biostat.wisc.edu/~kbroman/Rintro/Rwinpack.html