Bayesian Method for Data Science

Instructor: Yuxiao Huang

Group 2: Krystin Sinclair, Ye-in Jeon, Christian Cleber Masdeval Braz

UCI German Dataset

This README file intends to help using the source files disposal as part of the project. It is organized in the following way:

1. Exploratory data analysis, Frequentist and Bayesian Regression Analysis
2. MCMC to estimate individual and group default influence
3. Logistic and bayesian logistic regression

1 - Exploratory Data Analysis and Frequentist and Bayesian Regression

* download the files *exploratory analysis and regression.R* and *GermanData.csv*.
  + Link to the data: <http://home.cse.ust.hk/~qyang/221/Assignments/German/>
* run the R script file *exploratory analysis and regression.R*.

2 - MCMC to estimate individual and group default influence

* For the PURPOSE categorical feature
  + download the folder mcmc\_estimating\_purpose\_default
  + set this folder as your work space
  + run the R script file credit\_loan\_driver.R
* For the JOB categorical feature
  + download the folder mcmc\_estimating\_job\_default
  + set this folder as your work space
  + run the R script file credit\_loan\_driver.R
* For the Housing categorical feature
  + - download the folder mcmc\_estimating\_housing\_default
    - set this folder as your work space
    - run the R script file credit\_loan\_driver.R
* For the SAVING categorical feature
  + download the folder mcmc\_estimating\_saving\_default
  + set this folder as your work space
  + run the R script file credit\_loan\_driver.R
* For the CREDIT HISTORY categorical feature
  + download the folder mcmc\_estimating\_credit\_history\_default
  + set this folder as your work space
  + run the R script file credit\_loan\_driver.R

3 - Logistic and bayesian logistic regression

We used MCMC to help estimate the posterior of our bayesian inference. In this case, the parameters to be estimated are the coefficients of a linear function, relating the features to the response variable, which result is the parameter of the Sigmoid function. Thus, there is a model written in the JAGS software to accomplish this task. The desired result, i.e., the list of all the coefficients of the linear function, is generated at the end of the processing in a file called *loan-default-SummaryInfo.csv* (folder mcmc\_bayesian\_logistic). We do not recommend to run this step from scratch, once it takes more than one hour to finish. Instead, you can run directly the logistic analysis doing:

* download the file *loan-default-SummaryInfo.csv*  from the folder mcmc\_bayesian\_logistic
* download the R markdown file *frequentist\_and\_bayesian\_logist.Rmd*
* set your work space
* run the file