# Homework 3

A well-known artificial dataset, wave dataset (Breiman et al., 1984), is studied. Originally, in wave data set, there are three classes with 21 variables and in each class, the variables are generated based on a random convex combination of two of three waveforms with noise add. Rakotomalala (2005) expanded the dataset by adding noise variables and generating more subjects. He also modified this dataset as a binary response dataset by only keeping the subjects of the first two classes.

Currently in this dataset, there are 33,334 subjects and 10,000 subjected are collected as the training set. It is a balanced classification dataset because the numbers of the subjects for both classes are almost the same. In addition to the original 21 predictors, 100 noise variables are added and they are completely independent from the corresponding classification problem. Rakotomalala (2005) analyzed this dataset via a free data mining software, "TANAGRA", and he did not only used logistic regression for the classification problem but also performed forward selection according to SCORE test based on the whole training set. Overall, the classification error rate for testing set is 7.87%. For the variable selection results, there is no noise variables selected into the classification model, and 15 variables from the original 21 active variables are identified as active variables by setting the significant level as 1% in the SCORE test. For more details, please refer to the attached reference.

作業要求:
針對這 binary classification problem, 所提的參考資料中, 是一個兩階段地分析方法. 在這作業中, 請同時執行分類及變數選擇。

方法建議:
採用羅吉斯迴歸模型來作 binary classification。而在其參數估計時可採用 lasso 的方法用以選重要變數。在 R code 中，可以用 glmnet package.
https://cran.r-project.org/web/packages/glmnet/index.html
在所提供的參考文獻中 Hsu et al. (2019)，有論述 Grafting technique，這方法亦可稱為 Chebychev's Greedy Algorithm，也是一種選擇。
或是有其他作法亦可以採用，例如以線性迴歸模型來處理二元分類，則在線性模型下相關的變數選擇亦可採用。

作業內容規範

此作業視為是一個小型的 project. 報告約四頁. 須包含資料簡介, 分析概念說明及分析結果等。