Part I Assignment for SDC Biostatistics

April 26, 2021

Introduction:
Checking tumors' response to certain drugs has great applications in practice, including screening for certain genes that are the most predictive ones of drug response. Typically, this analysis will be based on the drug response and gene expression data. Here we take the drug (Doxorubicin_IC_50) as an example to analyze the response of tumor cells and screen some predictive genes.

In data1, the rows represent tumors, while the columns represent the genes (or the explanatory variables). The entries in the matrix represent gene expression level in corresponding tumors. There are 641 tumors and 100 genes in total. In data2, the rows are also the tumors and the first column is the response of the tumors to the drug. The 641 tumors have been divided into two categories according to their response to the drug (Doxorubicin_IC_50). The second column is the class label (0 represents class 1, 1 represents class 2).

Tasks:
1. Check the data preprocessing and analyze whether the data needs some preprocessing (e.g., normalization).

2. Try to select 10 genes that are the most explanatory ones to the drug (Doxorubicin_IC_50) response. What methods can you choose to achieve that (at least try three ways) and try to describe your analysis with detailed description and some figures.

3. Try to use the forward selection procedure with P-value as a criterion to show the process of screening 10 genes step-by-step.

4. Try to plot and analyze the changing trend of variable coefficients as the penalty parameter (lambda) increasing in lasso and ridge regression. Try to describe the properties of the coefficient changing for them.

5. Please randomly divide the 641 tumor samples into 2 parts (with ratio is 5:1) as training data and testing data respectively. Try to train a classification model (e.g., SVM or Logistic Regression) on your training data and test your model on the testing data. Please describe your results and test some possibility to improve your model.

6. Try to find the influence of different kernel functions and penalty super-parameter (e.g., C) on your training error and testing error, and describe your findings with experiments.