Claremont Graduate University, Mathematics M454, Statistical Learning
Lab 3 In-class Exercises
Instructor: Qidi Peng
TA: Nan Rao
April 3rd, 2018

CH 5: Linear Model Selection and Regularization

This is an independent machine learning project. Use $set.seed(1)$ before generating a random variable!

Project : Company Bill

Download the file "CompanyBill" from Canvas: Files-LABS-Lab3. Load this data set. Attention: This data set contains missing data. "0" means the corresponding data is missing. Delete the rows where there is a missing data (please check the tutorial CH5andCH6 for hint). Call the new data set "CompanyBill".
Questions.

(a) Show the dimension of the data set "CompanyBill".
(Hint: since the input data are not numeric, use (this is one example, please modify the path to your own)

$> CompanyBill = read.table("C:/ProgramFiles/R/CompanyBill.txt", header = TRUE);$

to load convert the data to numerics. Use $fix(CompanyBill)$ to show the datasheet to TA.)

(b) We would perform a model selection using best subset selection method. Use $V1, \ldots, V7$ to rename the 7 columns of CompanyBill. Let $CompanyBill\$V1$ be the response and $CompanyBill\$V2$ to $CompanyBill\$V7$ be 6 candidate predictors (the feature space dimension $p = 6$). Run the **best subset method** on the data $CompanyBill$ and show the summary of the results. **(Grab a TA to make sure you understand every single detail of the output.)**

(c) By using the **forward stepwise selection method**, show the plots of R^2, adjusted R^2, C_p and BIC VS number of predictors in the same picture (Window of size $2 * 2$). Use red color to identify the extreme values of these statistics, if they exist. **(Grab a TA to make sure you understand every single detail of the output.)**

(d) Randomly pick half of the observed data as a training set, using

$> set.seed(1);$
$> train = sample(1 : nrow(CompanyBill), nrow(CompanyBill)/2);$

Perform ridge regression on $V1 \sim V2, \ldots, V7$ on the training set. Use cross-validation to find the best lambda. Show the corresponding regression coefficients when lambda is the best one.

(e) Perform the lasso on $V1 \sim V2, \ldots, V7$ on the training set. Use cross-validation to find the best lambda. Show the corresponding regression coefficients when lambda is the best one.

(f) Pick half data for fitting and half for validation to fit $V1 \sim V2$ to $V7$. Use subset validation method to tell which model is the best among ridge regression and the lasso. (You have to determine the test MSE for each of these 2 models and make a comparison.)