Stockholm University

Mathematical Statistics
Stockholm University
Bachelor Thesis **2017:6**
http://www.math.su.se

A simulation study of model fitting to high dimensional data using penalized logistic regression

Ellinor Krona*

June 2017

Abstract

Preceding studies show that some common variable selection methods do not conform with high dimensional data. The purpose of this study is to introduce reduction of high dimensional data using penalized logistic regression. This study evaluates three penalization methods; ridge regression, the lasso and the elastic net for model fitting on four simulated examples of high dimensional data sets. For each example 30 data sets were simulated containing 400 predictors and 200 observations. Each example differed in correlation among predictors and relation to the binary response variable. Descriptive statistics and measures of predictive power were used to analyze the methods. The results showed that for high dimensional correlated data the elastic net and ridge regression dominate the lasso regarding the predictive power. There were significant differences (P-value $<$ 0.01) when comparing the predictive power using AUC between the methods in 2 out of 4 examples. In conclusion, the elastic net is notably useful in $p \gg n$ case. In addition, the lasso is not a satisfactory method when p is much larger than n. Ridge regression is proved to have high predictive power but is refrained from shrinking coefficients to be exactly zero.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: ellinor.krona@gmail.com. Supervisor: Jan-Olov Persson.