Stockholm University

Mathematical Statistics
Stockholm University
Bachelor Thesis **2019:7**
http://www.math.su.se

Predictive Power of Logistic Regression versus Random Forest: A simulation study

Amanda Möller*

June 2019

Abstract

In statistical and machine learning, there are many powerful techniques that can be used to make predictions using big data. A central problem in statistical learning involves choosing the best method for a given application. In this thesis two classifiers are examined, logistic regression and random forest, where these two are investigated and compared to each other in order to find the best classifier. The analysis is done both practically and theoretically and is focused on binary classification where the response variable is categorical and divided in classes. The simulation studies are performed for three different models where the number of predictors and the size of the data set vary and in each study are 25 data sets generated. AUC, F-score and misclassification rate are the measures used to analyze and evaluate the predictive power for the classifiers. It is concluded that the predictive power of logistic regression is better than random forest when the two classes are linearly separable and random forest predicts better than logistic regression when the two classes are non-linearly separable. Both classifiers perform better when the correlation between the explanatory variables increases.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: amanda.moller95@gmail.com. Supervisor: Ola Hössjer.