Project 2 - Predictive Modeling for Stat 159

Dakota Lim & Kartikeya Gupta

November 2, 2016

Abstract

This report, which is an extension to what we have learnt in class, explores the relationship between "Balance" (average credit card debt) and a set of variables given in the Credit data set. It closely follows the analysis done in Chapter 6 - Linear Model Selection and Regularization from the book An Introduction to Statistical Machine Learning by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. The main idea of this report is to fit a predictive model that can help decribe the relationship better.

Introduction

This report explores the relation between the average credit card debt across people and a set of variables that include credit limit, age, income, ethnicity etc. While still sticking to a linear model, we do however, explore models other than the OLS. To create this predictive model, we will be testing two shrinkage reduction methods - Ridge and Lasso, and two dimension reduction methods - Partial Least Squares and Principle Component Analysis. The final model for prediction will be that which explains the most of variation in the data for the given least Mean Squared Error.

Data

For this model, we will be using the Credit.csv data set available on the Introduction to Statistical Learning website: http://www-bcf.usc.edu/~gareth/ISL. It is also the same data set used in Chapter 6 of the book.

This dataset contains "Balance" which is the average credit card debt across people and 10 explanatory variables, of which 6 are quantitative and 4 and qualitative.

The quantitative variables are: Income, Credit Limit, Credit Rating, No. of Cards, Age and Education. The qualitative variables are: Gender, Student Status, Married and Ethinicity.

Method

We start our analysis by first conducting basic Exploratory Data Analysis to understand the variables we are working with and then following on with Pre-modeling Data Processing to structure the data in such a way that allows us to fit the multiple models on this set.

For EDA, we consider the summary statistics and use some basic plots to understand the variables. For qualitative variables, we create fequency tables and chart these proportions. Given that we want to understand the relationship between variables, we plot a matrix of correlation for the quantitative variables, ANOVA between Balance and qualitative variables along with conditional boxplots for the same.

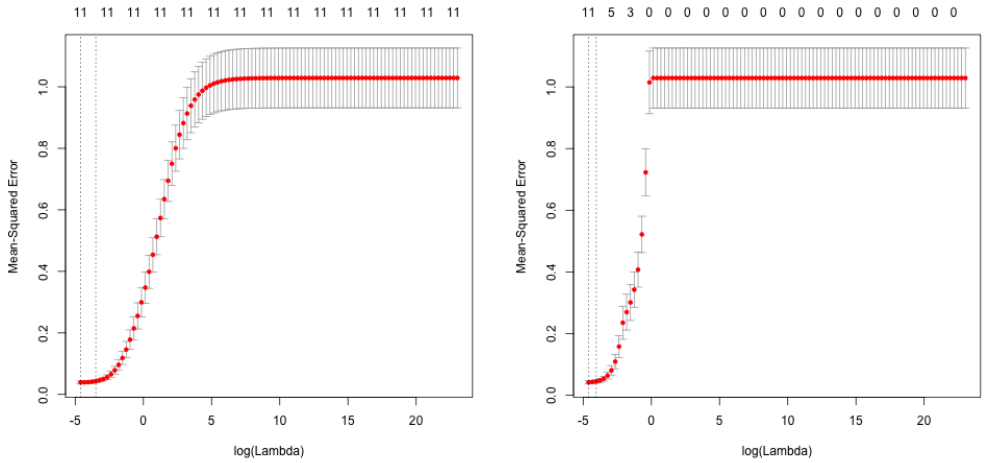
For Pre-modelling Data Processing, we follow the two major steps: 1. Coverting Factors into Dummy Variables so that they can be used in the regression. 2. Mean Centering and Standarization of Variables - This is done because the different models can only be compared when their coefficients represent the same scales. Hence, we mean center all the variables to 0 and the adjust the standard deviation to 1. Doing this given us comparable scales. Finally, we randomize this scaled data and take out a training set of size 3/4 of the data and the test set which is size 1/4 of the data.

Moving to the 5 regression models that we use:

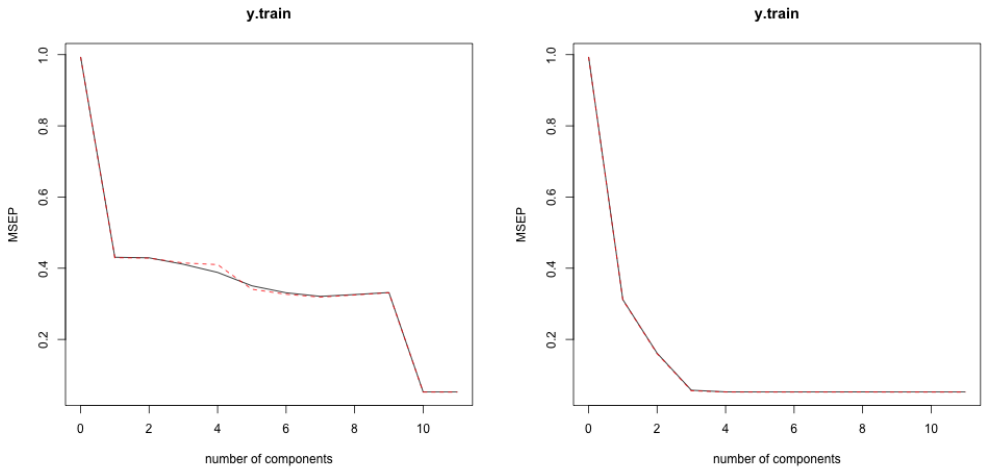
- An OLS model on the training data
- A Ridge model, using 10-fold cross validation to select lambda, on the training data
- A Lasso model, using 10-fold cross validation to select lambda, on the training data
- A PCR model, using 10-fold cross validation to select the number of components, on the training data
- A PLSR model, using 10-fold cross validation to select the number of components, on the training data

Each model was built using a scaled and centered version of the raw data. To ensure consistency we divide the data into test/training sets once and use the same sets to train and test each model.

Each model (excluding OLS) required a tuning parameter be definied, lambda values for Ridge and Lasso, and the number of components for PCR and PLS. To decide thse tuning parameters we used cross validation, below are the plots of prediction intervals for the MSE of the Ridge and Lasso methods as a function of different lambda values:



Next we consider the plots for MSEP as a function of the number of components taken in the PCR and PLS methods:



We simply take the minimal value (lambda or #-components depending on the model) of each plot and use that value as the tuning parameter for our full models.

Analysis

Results

Conclusion