Statistics 159 Project 2 Report

Aoyi Shan, Yukun He

10/28/2016

Abstract

In this project, we explore the difference between various muitiple regression methods and conclude on what approach should be used in order to fit our data better.

Introduction

In this project, our objective is to apply model selection methods introduced in Chapter 6, Linear Model Selection and Regularization, from the book "An Introduction to Statistical Learning" by Gareth James, Deniela Witten, Trevor Hastie and Robert Tibshirani. In our analysis, we compare the fit of models generated by 5 different regression methods, regression model obtained by Ordinary Least Squares, Ridge regression, Lasso regression, Principal Components regression and Partial Least Squares regression. To evaluate how well each model fits the data, we perform 10-fold cross validation in the model construction process and select the model with minimum cross-validation errors in terms of the running parameter. The requirement for this project can be found at https://github.com/ucb-stat159/stat159-fall-2016/blob/master/projects/proj02/proj02-predictive-modeling.pdf.

Data

The data we used in this analysis can be downloaded from http://www-bcf.usc.edu/~gareth/ISL/Credit.csv, which is provided by the textbook we refer to throughout the project, An Introduction to Statistical Learning.

After downloading the original data, in order to have comparable scales, we first standardize the data by centering around means and dividing by their respective standard deviation. Then, since we will use cross validation to improve the accuracy of our fit, we divide the data into training set and test set. We have 400 observations in total, so we randomly select 300 entries to be in the train set and the remaining 100 in the test set.

Methods

Least Squares Method

OLS estimators are considered as our base case in this project. OLS estimators are obtained by minimizing the sum of squares $RSS = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}(\beta_j x_{ij}))^2$ where $\beta_0, ..., \beta_p$ are our coefficient estimates. This is the most common regression method, and it works the best when we have homoskedastic and uncorrelated errors within the data.

Shrinkage Method

Similar to the OLS method, shrinkage method also fits a model containg all independent variables but with a focus on constrains the coefficient estimates toward 0. By shrinking the coefficient estimates, the variance in estimators is significantly reduced compared to that of OLS.

Ridge Regression

Instead of minimizing RSS, Ridge regression minimizes $RSS + \lambda \sum bj^2$. This regression abandons the requirement of an unbiased estimator in order to obtain a more precise prediction intervals. The addition term is a shrinkage penalty and since it's a sum of coefficient estimators, it will be small when βs are all close to 0. Here, λ is called a tuning parameter and it controls how much RSS and shrinkage panelty affects the regression model. When $\lambda = 0$, ridge regression is the same as the least squares models. Becuase each λ corresponds to a different set of coefficient estimates, we usually try a wide range of λs and pick the best tuning parameter with the smallest MSE.

One of the main feature of ridge regression is the bias-variance trade-off. When $\lambda = 0$, which is the least square model, the variance is high the estimators are unbiased. As lambda increases, variance decreases significantly but bias only sightly increase. Since $MSE = variance + squaredbias$ along with the relative effect of the change, when λ is smaller than 10, the variance drops rapidly, with very little increase in bias, so MSE decreases.

Lasso Regression

The major difference between Lasso and Ridge is that Lasso method aims to minimize $RSS + \lambda \sum |\beta_j|$. This is to compensate for the fact that ridge regression with always generate a model with all predictors since it doesn't involve a process of variable reduction. Therefore, lasso improves on this by allowing some of the coefficient estimates to be exactly 0 if the tunning parameter is sufficiently large. Therefore, this variable selection process makes the model much easier to interpret with a reduced amount of predictors.

Comparison between Ridge and Lasso

In general, lasso is expected to outperform ridge when we have a small amount of predictors having significant coefficients, and the rest close to 0. Ridge regression will perform between when the response is affected by many regressors with equal-sized coefficients.

Dimension Reduction Methods

Dimension reduction method involves a transformation of variables before we fit in the least squares model. Instead of estimating $p + 1$ coefficients $\beta_0, ..., \beta_p$ when we have p regressors, it transform the data to $Z1, ..., Z_m, M < p$ where Z_m represent linear combinations of the original predictors, so that we only need to estimate $M + 1$ coefficients $\theta_0, \theta_1, ..., \theta_M$.

Principal Components Regression

In principal components regression, we first perform privipal components analysis on the original data which constructs M principal components, $Z_1, ..., Z_m$. Then, we use these components as the predictors and fit the least squares model to obtain coefficient estimates. The key assumption we hold in this regression model is that the directions in which $X_1, ..., X_p$ show the most variation are the directions that are associated with Y. If the assumption holds, then fitting a least squares model to $Z_1, ..., Z_m$ will lead to better results than fitting a least squares model to $X_1, ..., X_p$, since most or all of the information in the data that relates to the

response is contained in $Z_1, ..., Z_m$. With fewer predictors, we can also reduce the risk of overfitting. PCR performs the best when the first few pricipal components are sufficient to capture most of the variation in the predictors and their relationship with the response.

Partial Least Squares

In comparison to PCR regression, which uses unsupervised method to dentify the principal components, partial least squares method takes the advantage of a supervised learning process. It uses the response variable Y to identify new vectors that are not only similar to existing regressors, but also select the ones that are related to the response. Therefore, as indicated in the textbook, the PLS approach attempts to find directions that help explain both the response and the predictors.

Analysis

	ols	ridge	lasso	pcr	plsr
(Intercept)	0.00000	0.00000	0.00000	0.00000	0.00000
Income	-0.59817	-0.56871	-0.55166	0.25505	-0.59801
Limit	0.95844	0.71866	0.92505	0.27196	0.67458
Rating	0.38248	0.59306	0.36787	0.26946	0.66699
Cards	0.05286	0.04425	0.04500	-0.05334	0.04808
Age	-0.02303	-0.02538	-0.01666	0.05327	-0.02592
Education	-0.00747	-0.00588	0.00000	0.05542	-0.00908
GenderFemale	-0.01159	-0.01068	0.00000	0.05408	-0.01671
StudentYes	0.27815	0.27318	0.26681	0.10534	0.27394
MarriedYes	-0.00905	-0.01103	0.00000	-0.02681	-0.00800
EthnicityAsian	0.01595	0.01638	0.00000	-0.00120	0.02257
EthnicityCaucasian	0.01101	0.01101	0.00000	-0.01697	0.01101

Table 1: Regression Coefficients for 5 Regression Methods

Comparing the regression coefficients of five regression methods, we can see that the coefficients of Income, Cards, Age, Education, GenderFemale, StudentYes, MarriedYes, EthinicityAsian and EthnicityCaucasian are relatively close for each regression method while the coefficients of Limit and Rating are relatively differed from each other. We can also see that the coefficients of Education, GenderFemale, MarriedYes, EthnicityAsian, and EthnicityCaucasion are very small and close to zero.

First, comparing the ols regression coefficients with the ridge regression coefficients, we can see that the ridge regression coefficients are generally smaller than the ols regression coefficients. This is because ridge regression method shrinks the coefficients of predictor variables and makes the coefficients closer to the true ones.

Second, comparing the lasso regression coefficients with the ridge regression coefficients, we can see that the lasso regression has several coefficients as zero. This is because lasso performs variable selection and yields model that involves only a subset of the variables. Lasso zeros out the coefficients of collinear variables. The advantage of lasso regression model over ridge regression model is that lasso regression method does both the parameter shrinkage and the variable selection and ridge regression will include all predictors in the final model and create a challenge in model interpretation.

```
par(mfrow = c(1,1))
plot(reg.coef.mat[,1], xlab = "Coefficients", ylab = "Value", main = "Trend Lines of the Coefficients")
lines(reg.coef.mat[,1])
points(reg.coef.mat[,2],col= "blue")
lines(reg.coef.mat[,2],col = "blue")
```

	MSE
ols	0.05179
ridge	0.05259
lasso	0.05154
pcr	0.41615
plsr	0.05192

Table 2: MSE of 5 Regression Methods

```
points(reg.coef.mat[,3], col = "red")
lines(reg.coef.mat[,3],col = "red")
points(reg.coef.mat[,4], col = "green")
lines(reg.coef.mat[,4],col = "green")
points(reg.coef.mat[,5], col = "yellow")
lines(reg.coef.mat[,5],col = "yellow")
legend(10,0.9,c('OLS','Ridge','Lasso','PCR','PLSR'), lty = c(1,1,1,1,1), lwd = c(2.5,2.5,2.5,2.5,2.5),
abline(h = 0, lty = 3)
```