Selection

Prof. Eric A. Suess

May 6, 2020

Introduction

Today we will discuss Chapter 4 Linear Model Selection and Regularization, from An Introduction to Statistical Learning

- Feature Engineering - feature creation and feature selection
- Best Subsets
- Forward Selection
- Backward Selection
- Ridge Regression
- Lasso
- Principal Components Analysis (PCA)
- Multidimentional Scalling (MDS)
- Big Data - Tall $n >> p$ and Wide $n << p$

Feature Engineering

When modeling data it is common to create new features from the data.

The **features** are **numeric and categorical columns** of data that can be used as input into algorithms to fit predictive models.

The features can be transformed.

There is usually a **target variable** and there are **input variables**.

Be careful not to use the target variable (or a function of the target variable) as an input variable.

Then we try to find the important features for prediction or classification. This is **feature selection**.

Linear Regression

When we **build regression models** we try to build the "best" model for predicting Y.

To build/estimate the regression model

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \epsilon$$

the parameters β_i, $i = 0, 1, ..., p$, are estimated by minimizing the Residual Sum of Squares (RSS).

$$RSS = \sum_i \left(y_i - \beta_0 - \sum_j \beta_j x_{ij} \right)^2$$

Linear Regression

The usual practice is to fit the model of interested, get the $\hat{\beta}$'s, and then **drop the non-significant features** or X's.

Drop the X's with β's that have **p-values** $>$ **.05**

Best Subsets

Thinking algorithmically we could try to fit all of the models.

That is fit all of the possible regression models for the data. That is fit the models with all p variables in the model, down to 1 variable models, and to the 0 variable model.

Then select the best model according to a criterial such as **adjusted-R squared**.

Forward Regression

In a Regression class, less computationally intensive algorithms are suggested.

Forward Regression adds variables until there are no further significant variables to add.

Backward Regression

Backward Regression removes variables until there are no further non-significant variables to remove.

And there are variations on Forward and Backward regression that are presented.

These methods are used for **feature selection**.

Moving beyond unbiased estimators

Least Squares produces **unbiased estimators** that have minimum variance. In Econometics this is BLUE.

Now, considering the **Variance/Bias Tradeoff**, it may be possible to develop biased estimators that have lower variance.

The additional bias with lower variance, may produce more understandable models and the bias and variance may be quite small with Big Data.

Ridge Regression

With large numbers of features using a more automated feature selection algorithm is helpful.

The goal would be "remove" the unimporatant variables from the model.

Ridge Regression "shrinks" the estimates of the β's toward zero.

This is done by adding a penalty to the optimization problem.

$$\sum_i \left(y_i - \beta_0 - \sum_j \beta_j x_{ij}\right)^2 + \lambda \sum \beta_j^2 = RSS + \lambda \sum \beta_j^2$$

Ridge Regression

With $\lambda = 0$ the usual linear regression is fit.

With larger values of the tuning parameter λ the lesser important β's are pushed to zero.

This is an automated feature selection algorithm.

The Lasso

While Ridge Regression is useful it leaves "all" of the features in the model, but with small estimated β's.

What would be better, **force** them to zero. That is automate the removal of the lesser important β's.

This is what **The Lasso** does.

The penalty is changed

$$\sum_i \left(y_i - \beta_0 - \sum_j \beta_j x_{ij}\right)^2 + \lambda \sum |\beta_j| = RSS + \lambda \sum |\beta_j|$$

Ridge Regression and The Lasso, difference

To see what the difference between Ridge Regression and The Lasso, see page 222 in An Introduction to Statistical Learning.

From Wikipedia

- Stepwise Regression
- [Lasso](https://en.wikipedia.org/wiki/Lasso_(statistics)
- Gradient Decent

PCA and MDS

Alternative approaches to the related problem of **feature reduction**,

Again from Wikipedia

- Principal Components Analysis (PCA)
- Multidimensional Scaling (MDS)

Mapping data from a higher dimensional space into a lower dimensional space while retaining the information in the original data.

Big Data

Two kinds of **Big Data**, high-dimensional data.

1. Tall data, $n >> p$

With **tall data** we can use ordinary linear regression, but if p is larger but still smaller than n, feature selection needs to be automated. Lasso is useful.

2. Wide data, $n << p$

With **wide data** we **cannot** use ordinary linear regression, feature selection needs to be automated. Lasso is useful.