Reading Report

- 姓名：李佳政
- 学号：201828013229075
- 2018年11月5日

Title: Regression Shrinage and Selection via the Lasso

LASSO: **L**east **A**bsolute **S**hrinkage and **S**election **O**perator

- Robert Tibshirani
- *University of Toronto, Canada*
- 1996

Abstract

This paper is publicized in 1996, and a new method was proposed for estimation the machine learning models. The main thought is that minimized the residual sum error. Different from the ridge regression, **LASSO**'s weight can produce a model that with more 0 cofficients, rather than be close to 0. In other words, we can get a more sparse weighting matrix, therefore, we can use it to discovery the patterns in training dataset, for instance, feature selection. The idea is general and can be used in lots of statistical model. In the end of the paper, there are some experiments about data mining. By the way, we still use it in 2018.

Two advantages

- Reduce over fitting. Original least squares have low bias but large variance, so it is overfitting the train set.
- Feature selection. Ridge regression is a continuous process, so its result weight matrix is more dense than lasso.

Orthonormal Formulae

The shrinkage can be gleaned from the orthonormal design case, therefore, the original equation can be easily shown to be

$$\beta_j = \mathrm{sign}(\beta^0_{\mathrm{j}})(|\beta^0_{\mathrm{j}}| - \gamma)^+$$

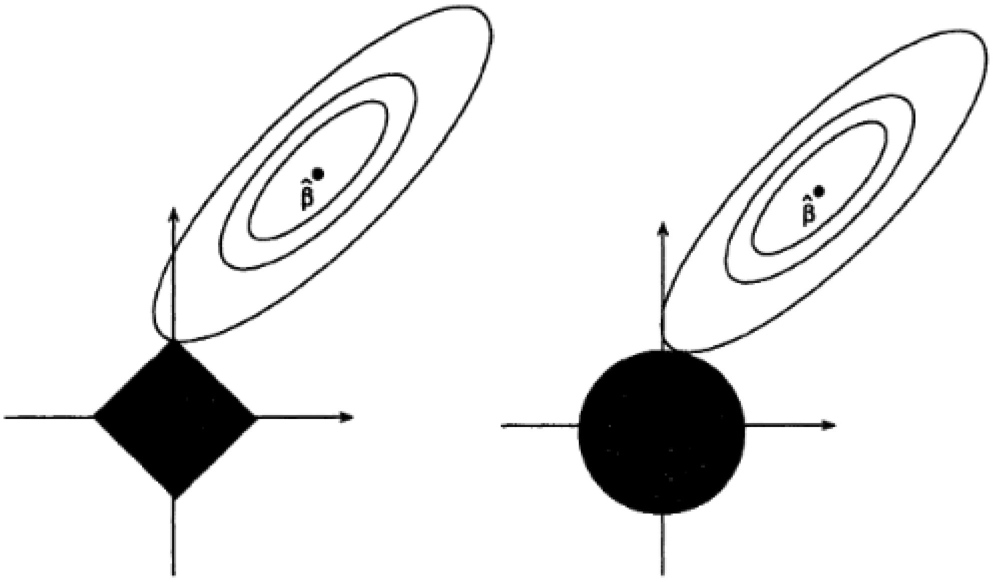
The γ is used to select feature, and it is determined by the condition $\sum |\beta_j| = t$. In other paper, it is minimum L1-nomrm penalty. And we select the k subset selection by choosing k largest cofficients and set the rest to 0. The operation can lead to the shrinkage and selection.

Geometry Analysis

How about the general(non-orthogonal) setting? We can get the equation,

$$(\beta - \beta_0)^{\mathrm{T}} \mathrm{X}^{\mathrm{T}} \mathrm{X} (\beta - \beta^0)$$

This paper used two figures to describe the zero-point distributions, the classical figures are showed as following.



We can find its intersections of lasso are on the coordinate axis, and ridge's not, so we can believe that lasso could produce more sparse cofficient matrix.

Errors

We can see that t is a hyper parameter, so we can optimize t, and fixed it to trian model. We can get the estimation,

$$\beta^* = (\mathrm{X}^{\mathrm{T}} \mathrm{X} + \lambda \mathrm{W}^{-})^{-1} \mathrm{X}^{\mathrm{T}} \mathrm{y}$$

The iteration is same as the ridge algorithm to compute the lasso estimation.

Experiments

This section is author's examples to prove the lasso's usage and some parameter tuning experience, including bigger model, various of regression model and so on. The conclusion is that the lasso is truly useful in machine learning field for shrinkage and selection for regression and generalized regression problesm.

The experiments are mainly in three different scnarios:

- *small numbers of large effects*, other subset selction > lasso > ridge.
- *small to moderate number of moderated-sized effects*, lasso > ridge > other subset selection.
- *large number of small effects*, ridge > lasso > other subset selction.

Overview from 1996 to 2018

This paper has been published for 22 years, the lasso has been a general knowledge for a machine learing researcher. The L1-norm, lasso regression is still useful for regression task, feature selection and reducing overfitting. This classical paper is worthy to read and think deeply. We can get more knowledge through reading the origin work!