Machine Learning (VII): High-Dimension Model, Regularization, and Lasso

Le Wang

Motivation

Solution

Solution Method: Lasso

Motivation

Motivation

As discussed earlier, our model is inherently large and complex when the number of predictors, especially continuous variables, increases.

In the age of big data, you may have $30,000$ or so genes in the human body are directly involved in the process that leads to the development of cancer, but only data for $1,000$ patients.

Similarly, in your online experiment, you may have only 500 clients, although the information for each of them may be enormous.

Linear Regression and Its Limitation

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{i,j}\beta_j + \epsilon_i$$

OLS estimation involves minimization of the following problem:

$$\min_{\beta_0,\beta_j} \sum (y_i - \beta_0 - \sum_{j=1}^{p} x_{i,j}\beta_j)^2$$

Linear Regression and Its Limitation

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{i,j}\beta_j + \epsilon_i$$

OLS estimation involves minimization of the following problem:

$$\min_{\beta_0, \beta_j} \sum (y_i - \beta_0 - \sum_{j=1}^{p} x_{i,j}\beta_j)^2$$

Issue: This problem typicall does not have a unique solution in the case of **high-dimension** or **wide* data (i.e., when the number of predictors is way larger than the number of data points ($p \gg N$)).

Intuition for the Problem: We can think of the issue in terms of the amount of information $\frac{N}{p}$ per parameter.

Intuition for the Problem: We can think of the issue in terms of the amount of information $\frac{N}{p}$ per parameter.

If $p \gg N$, and the every predictor plays a role (Not sparse!), then number of observations, N, is too small to allow for any accurate estimation of the parameters.

Intuition for the Problem: We can think of the issue in terms of the amount of information $\frac{N}{p}$ per parameter.

If $p \gg N$, and the every predictor plays a role (Not sparse!), then number of observations, N, is too small to allow for any accurate estimation of the parameters.

Question: What should we do?

Solution

Solution Concept

Sparsity: Loosely speaking, a sparse statistical model is one in which only a relatively small number of parameters or predictors play an important role.

Solution Concept

Sparsity: Loosely speaking, a sparse statistical model is one in which only a relatively small number of parameters or predictors play an important role.

In other words, only $k < N$ parameters are actually **non-zero** in the true model.

Solution Concept

Sparsity: Loosely speaking, a sparse statistical model is one in which only a relatively small number of parameters or predictors play an important role.

In other words, only $k < N$ parameters are actually **non-zero** in the true model.

We can effectively estimate the parameters using **Lasso**, and more important, we do not need to know which k parameters are actually non-zero!

Advantage of Sparsity::

1. Interpretation of the fitted model (**Model Selection**)

Advantage of Sparsity::

1. Interpretation of the fitted model (**Model Selection**)
2. Computational convenience. (**Convex Optimization Problem**)

Advantage of Sparsity::

1. Interpretation of the fitted model (**Model Selection**)
2. Computational convenience. (**Convex Optimization Problem**)
3. *Bet on Sparsity* Principle:

   Use a procedure that does well in sparse problems, since no procedure does well in dense problems.

Solution Method

One way to achieve **sparsity** is to **regularize** the estimation process.

We impose certain **constraint** on our parameters.

Solution Method

One way to achieve **sparsity** is to **regularize** the estimation process.

We impose certain **constraint** on our parameters.

Regularized Regression

OLS estimation involves minimization of the following problem:

$$\min_{\beta_0,\beta_j} \sum (y_i - \beta_0 - \sum_{j=1}^{p} x_{i,j}\beta_j)^2$$

OLS estimation involves minimization of the following problem:

$$\min_{\beta_0, \beta_j} \sum (y_i - \beta_0 - \sum_{j=1}^{p} x_{i,j}\beta_j)^2$$

We will now estimate the following model

$$\min_{\beta_0, \beta_j} \sum (y_i - \beta_0 - \sum_{j=1}^{p} x_{i,j}\beta_j)^2$$

subject to

$$\sum_{j=1}^{p} |\beta_j|^q \leq t$$

The constraint has two things

1. t: a budget on the total magnitude of the sum of the parameters. **Shrinkage parameter** that pull the parameters toward zeros.
2. q: the choice of norm formula.

What is the impact of varying t?

1. $t = 0$?: All the coefficients will be equal to zero, then we are back to the unconditional mean.

What is the impact of varying t?

1. $t = 0$?: All the coefficients will be equal to zero, then we are back to the unconditional mean.
2. $t = \infty$: We are back to OLS estimates (unconstrained ones)

What is the choice of q in practice?

$q = 1$: **Lasso** or l_1-regularized regression

Question: Why this particular choice?

$q = 1$ is corresponding to l_1 norm, which turns out to be very special.

If t is small enough,

1. $q \leq 1$ can generate sparse solutions (many zero coefficients), while this is not the case for $q > 1$

$q = 1$ is corresponding to l_1 norm, which turns out to be very special.

If t is small enough,

1. $q \leq 1$ can generate sparse solutions (many zero coefficients), while this is not the case for $q > 1$
2. in the case of $q < 1$, the minimization problem is NOT **convex**, and therefore computationally challenging to solve.

$q = 1$ is corresponding to l_1 norm, which turns out to be very special.

If t is small enough,

1. $q \leq 1$ can generate sparse solutions (many zero coefficients), while this is not the case for $q > 1$
2. in the case of $q < 1$, the minimization problem is NOT **convex**, and therefore computationally challenging to solve.

$\Longrightarrow$ $q = 1$ is the smallest value that yields a **convex** problem and a **sparse** solution. (You can use scalable algorithms that can handle even millions of parameters as a result!)

Solution Method: Lasso

Lasso

$$\min_{\beta_0,\beta_j} \sum (y_i - \beta_0 - \sum_{j=1}^{p} x_{i,j}\beta_j)^2$$

subject to

$$\sum_{j=1}^{p} |\beta_j| \leq t$$

Lagrangian: An Equivalent form of Lasso Problem

$$\min_{\beta_j \in \mathbb{R}^p} \sum (y_i - \sum_{j=1}^{p} x_{i,j}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Note that in practice both y_i and $x_{i,j}$ are typically standardized (with mean zero and unit variance) so that the intercept term β_0 is omitted in estimation.

Ridge Regression

When we consider L_2 norm, the problem becomes the **ridge regression** estimation

$$\min_{\beta_j \in \mathbb{R}^p} \sum (y_i - \sum_{j=1}^{p} x_{i,j}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^2$$