Predictive Analytics
Regression and Classification

Lecture 3 : Part 2

Sourish Das

Chennai Mathematical Institute

Aug-Nov, 2020

Class of Ill-Posed Problems

- A class of problem is known as ill-posed problem - if either of the following feature exits
  1. Unique solution does not exist
  2. Unique solution exists - but computationally not feasable
  3. Unique solution exists - but unreliable

1 Problem of variable selection in "large p, small n" setup considered as ill-posed problems

2 Problem of variable selection in large p is considered as ill-posed problems for model complexity.

3 Problem of multicollinearity also considered ill-posed problems.

Class of Ill-Posed Problems 1

- Unique solution does not exist

1. Problem of variable selection in "large p, small n" setup considered as ill-posed problems

- Such problems are common in medical sciences.

- For example, in a study of the efficacy of treatment; suppose the study randomly chose to observe 100 patients. It means the sample size n is 100.

- Now scientist collects 1000 of test results from each patient, from regular glucose level to genetic marker, etc. It means the number of features p is 1000.

Class of Ill-Posed Problems 1

- Unique solution does not exist

1. Problem of variable selection in "large p, small n" setup considered as ill-posed problems

- Such problems are common in medical sciences.

- In such kind of problem, you have infinitely many solutions; in fact, $\boldsymbol{\beta} = 0$ is also a possible true solution.

- It means none of the features of your study has any significant effect on your target variable $\mathbf{y}$, say efficacy. Certainly, it is not a desirable solution.

Class of Ill-Posed Problems 2

- Unique solution exists - but computationally not feasable

2. Problem of variable selection in large p is considered as ill-posed problems for model complexity.

- Suppose you are working in a credit rating group; where you are working with customer databases.

- The number of customers in the database is more than 100,000, and for each customer, you have 1000 features.

Class of Ill-Posed Problems 2

- Unique solution exists - but computationally not feasable

2 Problem of variable selection in large p is considered as ill-posed problems for model complexity.

- For such large dataset, if you apply a stepwise feature selection algorithm; then it has to fit $1 + \frac{p(p+1)}{2} = 500,501$ many models.

- It may take several days to complete the job.
- However, often time in the corporate environment you do not have several days and upper management wants the result by the end of the day.

- These are scenarios, where theoretically you have a unique and good solution. But computationally not feasible.

Class of Ill-Posed Problems 3

- Unique solution exists - but unreliable

3 Problem of multicollinearity also considered ill-posed problems.

- Multicollinearity is an interesting problem. In a sense, you have an unique solution. However, it is not reliable - because the standard error become so large that you cannot do a reliable statistical inference.

Regularization of Ill-posed Problems

- How to regularize an "Ill-posed problems"? So that we can have a solution !
- Tikhonov Regularization (1943) tries to find a solution for ill-posed problems by imposing certain restrictions, or conditions on the solution space.
- If a solution can be obtained, then we can say that process as the regularization of the ill-posed problem.

cmi

Penalizing Objective Function

- The class of functions is controlled by explicitly penalizing $RSS(f)$ with a roughness penalty

$$PL_2 = PRSS(f;\lambda) = RSS(f) + \lambda P(f)$$

- The amount of penalty is controlled by $\lambda \geq 0$.
- $\lambda = 0$ means no-penalty
- Typically λ is estimated from data.

  As we take $f(\mathbf{X}) = \mathbf{X}\beta$

$$\begin{aligned} PL_2 = PRSS(\beta;\lambda) &= RSS(\beta) + \lambda P(\beta) \\ &= (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda P(\beta) \end{aligned}$$

Penalizing Objective Function

- What about penalizing L_1-norm error? Can we penalize L_1-norm error?
- Yes we can. The model is:

$$PL_1 = ||\mathbf{y} - \mathbf{X}\beta||_1 + \lambda P(f)$$

- For now we focus on L_2-norm error.

What penalty to choose?

- For the model,

$$PL_2^2(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda P(\beta),$$

  one possible choice is L_2-norm penalty.
- That is

$$P(\beta) \quad = \quad (\beta - \beta_0)^T(\beta - \beta_0)$$

- Typical case $\beta_0 = 0$ and the penalty looks like

$$P(\beta) = \beta^T\beta$$

Analysis with L_2-penalty

- We want to minimize the L_2-penalized loss

$$PL_2^2(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda\beta^T\beta,$$

and we can obtained the Ridge solution as,

$$\hat{\beta}_{Ridge} = \text{argmin}_\beta \left[(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda\beta^T\beta\right]$$

- An equivalent way to write the ridge problem is

$$\begin{aligned} \hat{\beta}_{Ridge} &= \text{argmin}_\beta \left[(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)\right] \\ &\quad \text{subject to } \beta^T\beta \leq t, \end{aligned}$$

which makes explicit the size constraint on the parameters.

- There is a one-to-one correspondence between the parameters λ and t.

Ridge Regression

- Solving the following minimization problem,

$$\hat{\beta}_{Ridge} = \text{argmin}_{\beta} \left[(\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) + \lambda \beta^T \beta \right],$$

  we have the Ridge solution as

$$\hat{\beta}_{Ridge} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y},$$

  where $\mathbf{I}$ is the $p \times p$ identity matrix.

- Ridge solution is a special case of Tikohonov solution.

LASSO Regression

- Least Absolute Shrinkage and Selection Operator (LASSO)
- The lasso is a shrinkage method like ridge, with subtle but important differences.
- The lasso estimate is defined as

$$\hat{\beta}_{lasso} = \text{argmin}_{\beta} \left[(\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) + \lambda ||\beta||_1 \right]$$

- Equivalently can be expressed as

$$\begin{aligned} \hat{\beta}_{lasso} &= \text{argmin}_{\beta} (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) \\ &\text{subject to} \sum_{j=1}^{p} |\beta_j| \leq t \end{aligned}$$

Remark

- LASSO doenot have closed form solution like Ridge.
- Computing the lasso solution is a quadratic programming problem.
- Efficient algorithms are available for computing the entire path of solutions as λ is varied, with the same computational cost as for ridge regression.

Remark

- Because of the nature of the constraint, making t sufficiently small will cause some of the coefficients to be exactly zero.
- Thus the lasso does a kind of **continuous subset selection**.
- Ridge takes care of multicollinearity kind of issues.
- compromise between ridge and lasso was give Zou and Hastie (2005), known as Elastic Net penalty

$$P_{EN}(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} (\alpha \beta_j^2 + (1-\alpha)|\beta_j|)$$

LASSO, Ridge and Elastic Net

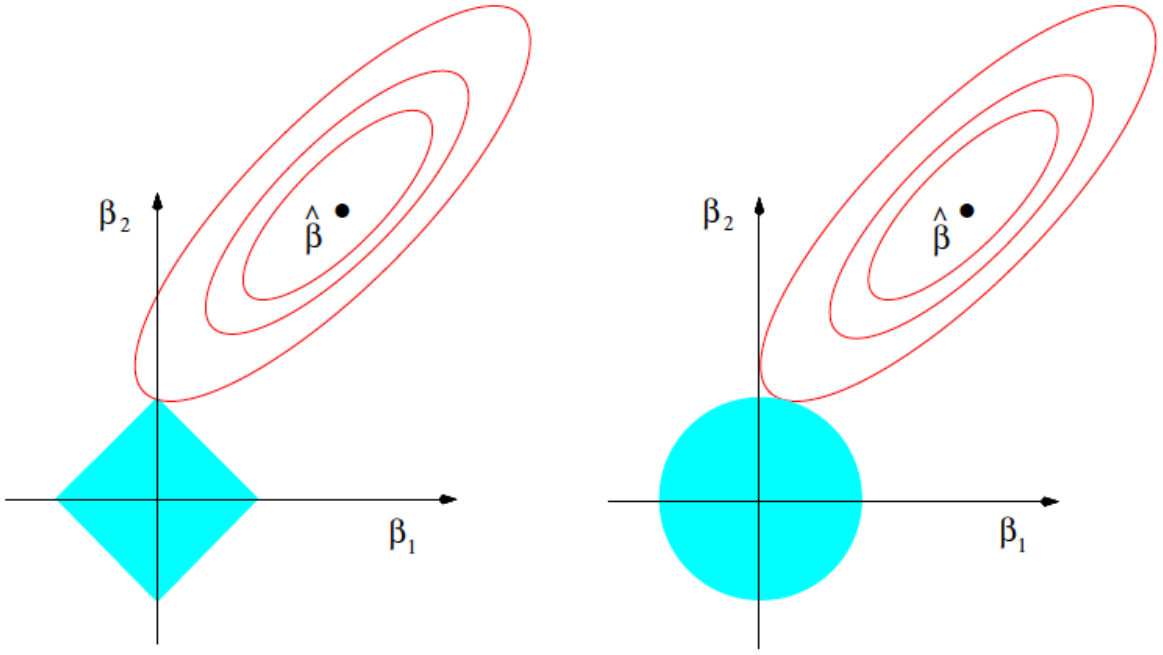
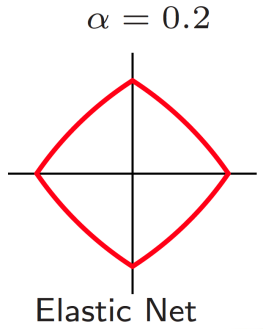


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$*, respectively, while the red ellipses are the contours of the least squares error function.*



Source: figure from "Elements of Statistical Learning" by Hastie and Tibshirani

Tikohonov Regularization for multicollinearity and feature selection

- Ridge Regression takes care of multicollinearity (Hoerl and Kennard (1970))
- LASSO Regression takes care of feature selection (Tibshirani, 1996)
- ElasticNet Regression takes care of feature selection (Zou and Hastie, 2006)

In the next part..

- We will resume some hands-on.

Thank You

sourish@cmi.ac.in