Stat 151a Lecture Notes 21

Chae Yeon Lee

April 17, 2023

1 Overview

Today:

- Shrinkage
- Introduction to GLMS

Relevant Readings:

- Fox 13.2.2; Fox 13.2.3
- Fox 14.1

2 Shrinkage Methods

Problems with prior selection strategies:

- **Forward/Backward selection** is a greedy search algorithm. Thus, we don't guarantee that we pick the optimal model. (But it is efficient)
- **All subsets** approach is running 2^p models. Thus, it becomes computationally infeasible or we could have $n < p$ (high dimensional data.)

Tangentially:

- We can potentially lower MSE by changing the optimization to induce some bias into coefficient estimates while significantly lowering their variance.
- We do this by shrinking parameter estimates towards 0 using a "sliding scale" hyperparameter (λ)

First, we do some **pre-processing** since

- i) we exclude intercept from variable selection
- ii) the variables / betas of the original data are on different scales. For standard OLS, we didn't care if the covariates have different scales because the standard error accounts for their different scales. But, when using shrinkage, we have to penalizing the total magnitude of betas for all covariates. Therefore, if we don't standardize the scales, then we are penalizing the covariates with larger scales than the covariates with smaller scales.

Standardization:

- $X \rightarrow Z = \left[\frac{\vec{x_1} - \bar{x_1}\vec{1}}{S_{x_1}} \quad \ldots \quad \frac{\vec{x_p} - \bar{x_p}\vec{1}}{S_{x_p}} \right]$
- X is all mean zero and don't have intercepts.
- $\vec{y} = \vec{y} - \bar{y}\vec{1} = \vec{y}$
- Center y in order to remove intercepts.

Now: regress $\vec{y^*}$ on Z.
To "shrink" the parameters, we place a constraint on the least squares optimization.

$$\min_{\vec{\beta}} \| \vec{y*} - Z\vec{\beta} \|^2$$

s.t.

$$\| \vec{\beta} \|_2^2 \leq c$$

or

$$\sum_{i=1}^{p} |\beta_i| \leq c$$

Regressions:

- Ridge Regression: $\min_{\vec{\beta}} \| \vec{y*} - Z\vec{\beta} \|_2^2 + \lambda \| \vec{\beta} \|_2^2$
- Lasso Regression: $\min_{\vec{\beta}} \| \vec{y*} - Z\vec{\beta} \|_2^2 + \lambda \sum_{i=1}^{p} |\beta_i|$
- L0 Norm (similar to subset selection): $\min_{\vec{\beta}} \| \vec{y*} - Z\vec{\beta} \|_2^2 + \lambda \sum_{i=1}^{p} I(\beta_i \neq 0)$

L0-norm is non-convex and difficult to solve.

2.0.1 RIDGE REGRESSION

Ridge has a closed form solution, which means that we can solve for β^* as a function of our inputs.

$$f(\beta) = \min_{\vec{\beta}} \|\overrightarrow{y*} - Z\overrightarrow{\beta}\|_2^2 + \lambda\|\overrightarrow{\beta}\|_2^2$$

$$= (\bar{y}^* - z\overrightarrow{B})^T(\overrightarrow{y^*} - Z\beta) + \lambda\overrightarrow{\beta^T}\overrightarrow{B}$$

$$= \vec{y^*}^T\vec{y}^* - 2\vec{y^*}^T Z\beta + \vec{\beta}^T Z^T Z\vec{\beta} + \vec{\beta}^T\lambda\mathbb{1}_p\vec{\beta}$$

Then we take derivative.

$$\frac{\partial f}{\partial\vec{\beta}} = 0 - 2Z^T\vec{y}^* + 2(Z^T Z + \lambda\mathbb{1}_p)\beta$$

Then we set to 0 to optimize.

$$\hat{\beta_\lambda} = (Z^T Z + \lambda\mathbb{1}_p)^{-1} Z^T\vec{y}^*$$

Now, we calculate the variance using the matrix-variance property(Var(A) = AVar(X) A^T).

$$Var(\hat{\beta}_\lambda) = \sigma_\epsilon^2(Z^T Z + \lambda\mathbb{1}_p)^{-1} Z^T Z(Z^T Z + \lambda\mathbb{1}_p)^{-1}$$

We can see that larger λ leads to smaller $\hat{\beta}$ and smaller var($\hat{\beta}$)

$$Bias(\hat{\beta}_\lambda) = E[\hat{\beta}_\lambda] - \beta$$

$$= (Z^T Z + \lambda\mathbb{1}_p)^{-1} Z^T Z\vec{\beta} - \beta$$

When λ is zero, we can get the identity matrix. Then β is unbiased. However, as $lambda$ gets larger, $(Z^T Z + \lambda\mathbb{1}_p)^{-1} Z^T Z$ diverges from the identity matrix.
In fact, there is always some value of λ such that the MSE of ridge estimator is lower than that of OLS. However, this λ is an unknown parameter.

2.0.2 EVALUATION OF RIDGE REGRESSION

Pros of Ridge:

- analytical solution for each λ
- theoretically can always outperform OLS.

Cons of Ridge:

- no explicit variable selection (lasso however is.)
- No good inference approach