Shrinkage Methods

Wednesday, September 22, 2021 6:08 PM

- Constraints (Coefficients)
  - Model (Linear)

$$Y = \beta_0 + \beta_1 x_1 + \cdots \beta_p x_p$$

$\beta_j \neq 0$

- Ridge Regression
  - Error/Loss Function:

$$RSS = \sum_{i=1}^{n} \left(y_i - \underbrace{\beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}}_{\hat{y}} \right)^2$$

- Penalized Loss Function

$$\underbrace{\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2}_{RSS} + \underbrace{\lambda \sum_{j=1}^{p} \beta_j^2}_{\text{penalty}}$$

$$= RSS + \lambda \underbrace{\sum_{j=1}^{p} \beta_j^2}_{\text{shrinkage penalty}}$$

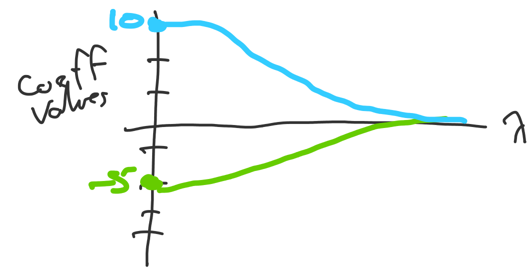
tuning param

shrinkage penalty ← Ridge

* All Betas:

$$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_p \end{bmatrix} \rightarrow L_2 \text{ Norm } \|\hat{\beta}\|_2^2$$

* Tuning Parameter → $\|\hat{\beta}_j\|_2 \rightarrow \|\hat{\beta}_{j\,\lambda}^{R}\|$



* Scale Equivalent → Feature: x_j, multiply by constant c:

$$c x_j$$

→ coeff: $\beta_j \rightarrow \frac{1}{c}\beta_j$

$x_j\beta_j$ constant!

↳ Does not apply for Ridge Regression!

Standardize Predictors!

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}}$$

brings all features to same scale!

- Pros/Cons:
  - OLS given a linear relationship between $Y \sim X_1 .. X_p$
    low bias / high variance
    ↳ especially as p increases relative to n!
  - Ridge trades bias for reduction in variance
    ↳ especially if $p > n$
  - For fixed lambda → computationally efficient solution!

- Does not take $\beta_j \Rightarrow 0$! No feature selection!

- Lasso
  - Perform feature selection $\Rightarrow \beta_j \to 0$
  - Loss/Error Function

$$\underbrace{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2}_{RSS} \; + \; \underbrace{\lambda \sum_{j=1}^{p}|\beta_j|}_{\text{penalty}}$$

$$= RSS + \underbrace{\lambda \sum_{j=1}^{p}|\beta_j|}_{\text{shrinkage penalty}} \longleftarrow \text{Lasso}$$

* All Betas

β_j is 0! $\rightarrow \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_p \end{bmatrix} \rightarrow$ L_1 Norm $\|\beta_j\|_1^1$ (Sparse Solution)

- Optimization View:
  - Ridge:

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \right\} \quad \text{obj func}$$

Constraint $\longrightarrow$ s.t. $\sum_{j=1}^{p}\beta_j^2 \le \underline{s}$

every λ corresponds

to specific value of s!

- Lasso:

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\}$$

$$\text{s.t.} \quad \sum_{j=1}^{p} |\beta_j| \le \underline{s}$$

Ex. 2-feature case: $p=2 \to \beta_1 \ \beta_2$