Regularization

Ridge (L2) and Lasso (L1)

Ramesh Srinivasan

October 15, 2024

Regularization

- There are 3 motivations for Regularization.
- Deal with instability issues arising from ill-posed problems.
- Avoid overfitting and improve prediction performance.
- Improve the interpretation of model prediction

Linear Regression Model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- $\boldsymbol{X}$ is an n $\times$ (p+1) matrix
- $\boldsymbol{\beta}$ is an p+1-dimensional vector
- $\boldsymbol{Y}$ is an n-dimensional vector
- $\boldsymbol{\epsilon}$ is an n-dimensional vector
- where n is the number of data points and p is the number of predictors

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Ordinary Least Squares Solution

If we minimize the mean-square error

$$MSE(\boldsymbol{\beta}) = \frac{1}{n}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|$$

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

The Gram matrix $\boldsymbol{X}^T\boldsymbol{X}$ is an extended covariance matrix

$$var(\hat{\beta}) = \sigma_Y^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

Eigenvalues of the Covariance Matrix

- The eigenvalues (λ) of $\boldsymbol{X}^T\boldsymbol{X}$ can inform us about the (potential) instability of our model.
- The condition number κ
$$\kappa = \frac{\lambda_{max}}{\lambda_{min}}$$
- A matrix with a small condition number is well conditoned and will provide stable solutions.
- A matrix with some eigenvalues closes to zero will be poorly conditioned and exhibit numerical instability.
- A matrix $\boldsymbol{X}$ with fewer observations (n) than predictors (p) will always have a covariance matrix with only n non-zero eigenvalues. Hence, the condition number is infinite and the matrix is not invertible.

Regularization

- We want less complex models to avoid overfitting and increase interpretability.
- We want to be able to solve problems where p = n or p > n, and still generalize reasonably well.
- We want to reduce instability (increase min eigenvalue/reduce condition number) in our estimators.
- We want to be able to handle colinear predictors
- In a nutshell, we want to avoid ill-posed problems (no solutions / solutions not unique / unstable solutions)

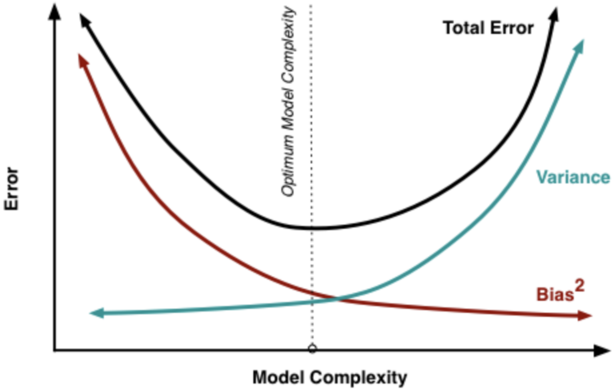
Classifiers

- The Bayes classifier based on LDA assigns the observation to the class k for which

  $$\delta_k(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu_k} - \frac{1}{2} \boldsymbol{\mu_k}^T \boldsymbol{\Sigma} \boldsymbol{\mu_k} + ln(\hat{p}_{Y=k})$$

  is maximum
- In the case of Logistic Regression its slightly more complicated to express since I can optimize the likelihood in a number of ways, but in all cases it requires a full-rank covariance matrix and is sensitive to conditioning.

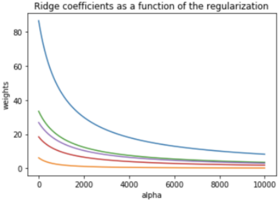
Model Selection: Bias-Variance Trade Off

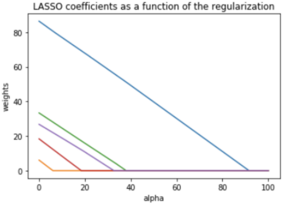


Reduce Variance by Increasing Bias

Ridge Regression

Ridge Regression is one such form of regularization. In practice, the ridge estimator reduces the complexity of the model by shrinking the coefficients, but it doesn't nullify them. This is also known as L2-norm regularization .


Ridge coefficients as a function of the regularization


LASSO coefficients as a function of the regularization

Lasso Regression

Lasso Regression is another form of regularization. In Lasso Regression, the penalty term is proportional to the L1-norm of the coefficients.

Differences between Ridge and Lasso Regression

1. Since Lasso regression tend to produce zero estimates for a number of model parameters - we say that Lasso solutions are **sparse** - we consider to be a method for variable selection. 2. In Ridge Regression, the penalty term is proportional to the L2-norm of the coefficients whereas in Lasso Regression, the penalty term is proportional to the L1-norm of the coefficients. 3. Ridge Regression has a closed form solution! Lasso Regression does not. We solve this iteratively.