Regularization

Ridge (L2) and Lasso (L1)

Ramesh Srinivasan

October 22, 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

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

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

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)

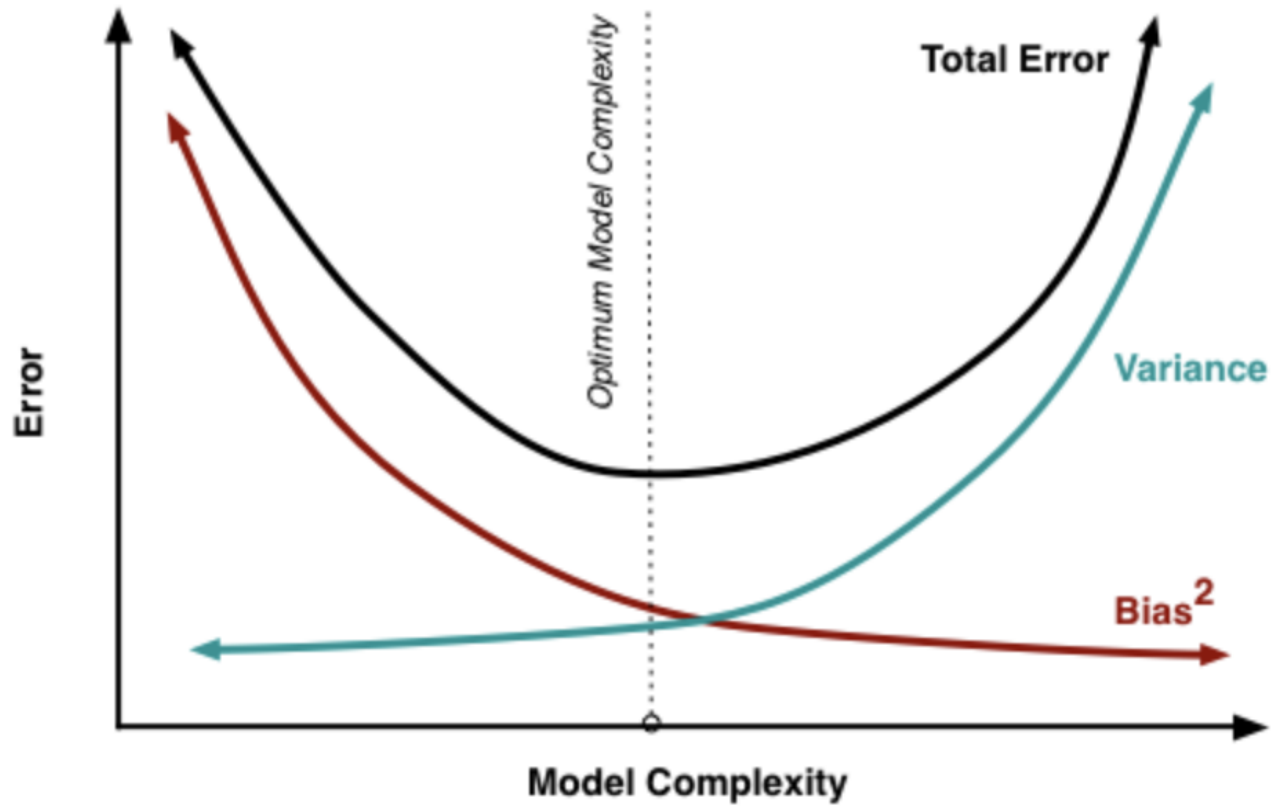
Classifiers

- 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.

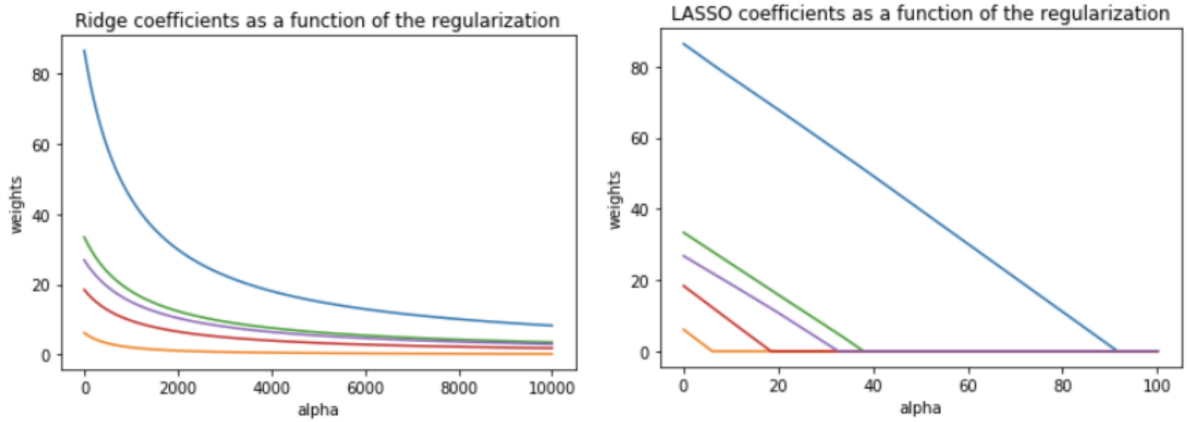
Model Selection: Bias-Variance Trade Off



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 .



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.

Regularization: Overview

- The idea of regularization revolves around modifying the loss function L
- In Linear Regression, the Loss function is the MSE
- In Logistic Regression, the Loss function is a Likelihood.
- The idea of the regularization is adding a 2nd term to the loss function that penalizes properties of the model parameters θ .

$$L_{reg}(\theta) = L(\theta) + R(\theta)$$

- In our Bias/Variance graph, we are adding Bias to the model, in the expectation that we can reduce variance by doing so.
- We will consider two forms of R which penalize the magnitude of θ in different ways.
- The L1-norm (absolute value) penalty is a LASSO penalty.
- The L2-norm (squared absolute value) penalty is a Ridge penalty.

Ridge Regularization

- Regularized estimator proposed by Hoerl and Kennard (1970).
- Imposes L2-norm penalty as

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{R}} = argmin_{\boldsymbol{\beta}} \| \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} \|^2 + \lambda \|\boldsymbol{\beta}\|^2$$

- In practice, the ridge estimator shrinks the coefficients with the amount of regularization dependent on λ

Ridge Regression versus OLS

- If we evaluate the matrix derivative with respect to beta and set it equal to zero, we obtain the solution

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{R}} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

- Recall, the Ordinary Least Squares (OLS)

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

Eigendecomposition and Matrix Inverses

- If we obtain the eigenvalues and eigenvectors of a matrix $\boldsymbol{C}$ we can write its eigendecomposition as

$$\boldsymbol{C} = \boldsymbol{M}^T \boldsymbol{\Lambda} \boldsymbol{M}$$

  where $\boldsymbol{M}$ is the model matrix whose columns are the eigenvectors and $\boldsymbol{\Lambda}$ is a diagonal matrix with the eigenvalues along the diagonal.
- With this decomposition we can compute the inverse of C as

$$\boldsymbol{C}^{-1} = \boldsymbol{M} \boldsymbol{\Lambda}^{-1} \boldsymbol{M}^T$$

- The inverse of a product of matrices

$$(\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1} \boldsymbol{A}^{-1}$$

- The inverse of a modal matrix (orthogonal, unit vectors)

$$\boldsymbol{M}^{-1} = \boldsymbol{M}^T$$

- The inverse of the eigenvalue matrix is simply

$$\boldsymbol{\Lambda}^{-1} = \begin{bmatrix} \lambda_1^{-1} & 0 & 0 & \dots \\ 0 & \lambda_2^{-1} & 0 & \dots \\ \vdots & & & \\ \dots & 0 & 0 & \lambda_p^{-1} \end{bmatrix}$$

Ridge Estimator and PseudoInverses

- The regularization in the Ridge estimator is equivalent to simply adding a constant to the eigenvalues before computing the inverses.
- This prevent very small (or even zero) eigenvalues from blowing up in the inverse.
- In domains where computing pseudo inverses are frequently necessary, such as ill-posed inverse problems (with more unknowns than measurements). this is known as Tikhanov regularization.
- The ridge regressor is a special case of Tikhanov regularization.
- Tikhanov regularization can also be used to impose smoothness constraints on solutions that are often useful in ill-posed inverse problems.

Key Property: Shrink the coefficients

- The Ridge estimator can be seen as a modification of the OLS estimator:

$$\hat{\beta}_{\boldsymbol{R}} = (\boldsymbol{I} + \lambda \boldsymbol{X}^T \boldsymbol{X}^{-1})^{-1} \hat{\beta}$$

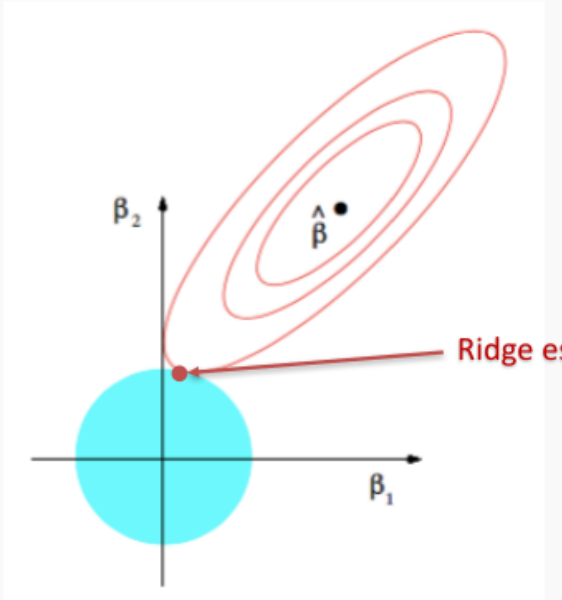
- In a univariate case, with standardized $\boldsymbol{X}$,

$$\boldsymbol{X}^T \boldsymbol{X} = 1$$

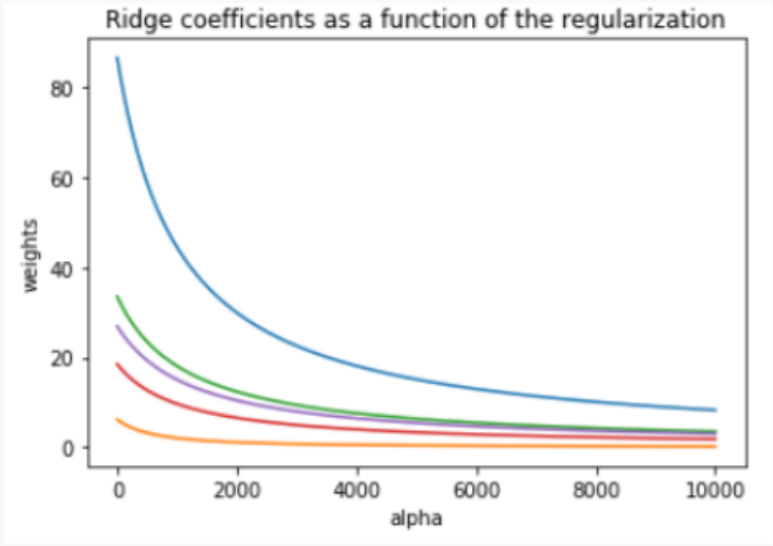
$$\hat{\beta}_{\boldsymbol{R}} = \frac{\hat{\beta}}{1 + \lambda}$$

- Ridge regression shrinks the OLS predictors, but does not nullify them.

Ridge a joint minimum of 2 Loss functions



The ridge estimator is where the constraint and the loss intersect.



The values of the coefficients decrease as lambda increases, but they are not nullified.

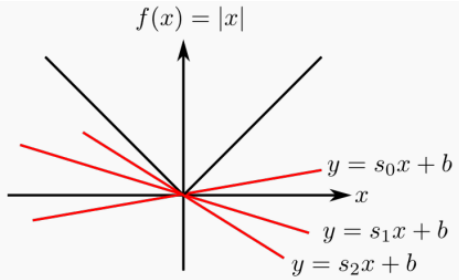
LASSO

- Least Absolute Shrinkage and Selection Operator
- Originally introduced in geophysics paper from 1986 but popularized by Robert Tibshirani (1996)
- Idea: L1 penalization on the coefficients.

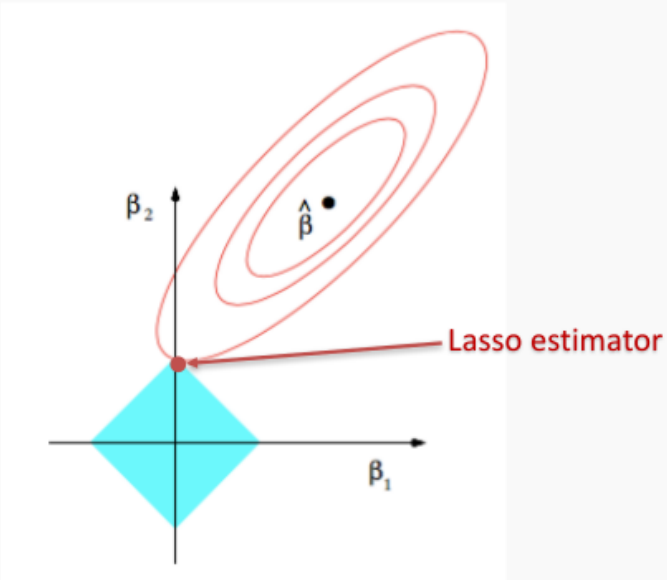
$$\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} = argmin_{\boldsymbol{\beta}} || \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||^2 + \lambda||\boldsymbol{\beta}||$$

- The key point is to penalize the absolute value of the coefficients instead of the sum squared.
- No closed form solution because L1 norm not have a derivative at zero

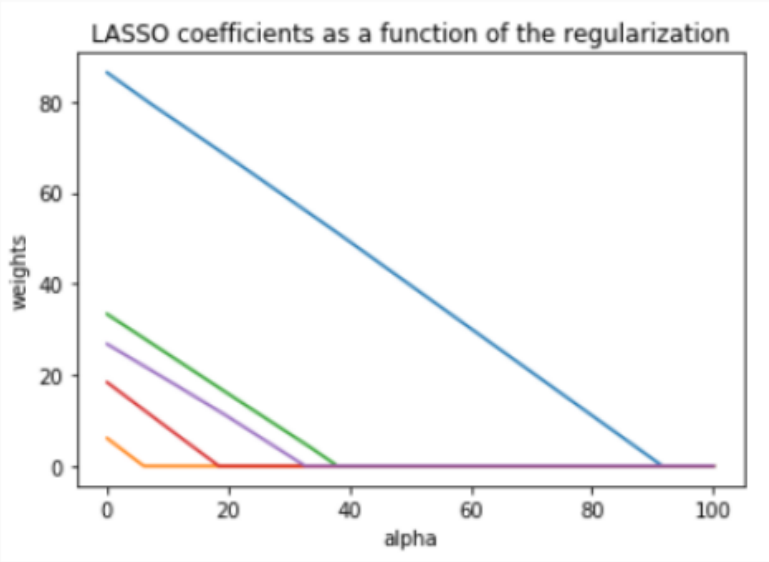
$$\partial(|\cdot|)(x) = \begin{cases} -1 & x < 0 \\ [-1,1] & x = 0 \\ 1 & x > 0 \end{cases}$$



Visualizing LASSO



The Lasso estimator tends to zero out parameters as the OLS loss can easily intersect with the constraint on one of the axis.



The values of the coefficients decrease as lambda increases, and are nullified fast.

Hyperparameter Selection λ

- We select λ by cross-validation.
- Find the minimum of the ridge/Lasso regression cost function and record the R^2 on the cross-validation.
- For classifier problems often you see accuracy as the criterion, I prefer AUC.
- Find the l that gives the largest validation statistic.
- For that selected value of λ examine performance on test data.