Ridge and Lasso Regression

Sravani Vadlamani

Agenda

- Shrinkage Methods
  - Ridge Regression
  - Lasso Regression

Shrinkage Methods

- Subset selection uses least squares to fit a linear model using a subset of predictors
- Shrinkage methods
  - Fit a model using all *p* predictors using a technique that *constrains* or *regularizes* the coefficient estimates i.e., they *shrink* the coefficient estimates towards zero
  - Shrinking the coefficients significantly reduces their variance
  - Two popular shrinkage methods are ridge regression and lasso regression

Ridge Regression

- Similar to least squares except that coefficient estimation is done by minimizing a different quantity from RSS
- Least squares estimate $\beta_0, \beta_1, \ldots. \beta_p$ by minimizing residual sum of squares given by

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

- Ridge regression coefficient estimates minimize

$$RSS + \lambda \sum_{j=1}^{p} \beta_j^2$$

$\lambda \geq 0$ is a tuning parameter

Ridge Regression

- Ridge regression estimates the coefficients by minimizing the RSS
- The second term $\lambda \sum_{j=1}^{p} \beta_j^2$ is called shrinkage penalty. It is small when the coefficients β_1, …. β_p are close to zero.
- The tuning parameter λ influences the effect of two terms on the regression coefficient estimates.
  - When λ = 0, shrinkage penalty has no effect and ridge regression estimates equal the least squares estimates.
  - When λ approaches infinity, the impact of the shrinkage penalty grows and the pushes/shrinks the ridge regression coefficients closer to zero

Ridge Regression

- Ridge regression produces different set of coefficient estimates $\hat{\beta}_{\lambda}^{R}$ for each value of λ.
- Selection of appropriate λ is critical and is done using cross-validation
- The shrinkage penalty is only applied to variable coefficients β_1, …. β_p and not to the intercept β_0. The goal is to shrink the impact of each variable on the response variable but not the intercept which is the mean value of the response when none of the predictor variables are present

l_2-norm (pronounced “ell 2”)

- $\|\beta\|_2 = \sqrt{\sum_{j=1}^{p} {\beta_j}^2}$
- The l_2 norm measures the distance of β from zero.
- As λ increases, the l_2 norm of $\hat{\beta}_\lambda^R$ will always decrease as the coefficient estimates shrink closer to zero.

Ridge vs Least Squares

- $Least\ square\ regression\ coefficients\ are\ scale\ equivalent\ but\ ridge\ regression\ coefficients\ are\ not.$
- In least square, multiplying X by a constant C scales the least square coefficient estimates by a factor of 1/C. Regardless of the scaling of the jth predictor, the value of $\beta_j\ X_j$ remains the same.
- However, ridge regression coefficient estimates change dramatically when the scale of a given predictor is changed. $X_j\ \hat{\beta}_\lambda^R$ may depend on the scaling of other predictors as well. Hence, it is best to apply ridge regression after standardizing the predictors

Scaling or Standardizing

- Use the following equation to bring all the predictors to the same scale as a result of which all the predictors will have a standard deviation of one. The final fit will not depend on the scale of the predictors

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

Improvement of Ridge over Least Squares

- Advantage of ridge regression over least squares is rooted in bias-variance trade off.
  - As λ increases, the flexibility of the ridge regression fit decreases, leading to decrease in variance but increase in bias.
  - For least squares (λ =0), the variance is high but there is no bias.
  - Ridge is best employed where least square estimates have high variance.
  - The computational requirements to calculate ridge regression estimates simultaneously for all values of λ are almost identical to those for fitting a model using least squares.

Ridge vs Best Subset Selection

- Best subset selection searches through 2^p models which may be infeasible for moderate values of p
- In contrast, ridge regression fits only one model for any value of λ and the model fitting procedure is faster.
- Also, ridge regression always uses all the p predictors. It will shrink the coefficient estimates toward zero but will never set them to exactly zero. The extra predictors do not impact the prediction accuracy but make the interpretability difficult especially when p is large.

The LASSO

- Least Absolute Shrinkage and Selection Operator
- Recent alternative to ridge regression that allows for exclusion of some variables
- Lasso regression coefficient estimates minimize

$$RSS + \lambda \sum_{j=1}^{p} |\beta_j|$$

- The main difference between ridge and lasso is the penalty term. Instead of β_j^2, lasso uses the l_1 norm of coefficient vector as the penalty which is given by

$$\|\beta\|_1 = \sum |\beta_j|$$

The LASSO

- l_1 can force some of the coefficient estimates to be exactly zero when the tuning parameter λ is sufficiently large. This implies that LASSO can perform variable selection like the best subsets method.
- The models generated from LASSO are easier to interpret than the ridge regression models. Hence, lasso models are sometimes called sparse models as they include only a subset of variables.

LASSO vs Ridge

- As λ increases, variance increases and bias decreases. Neither lasso nor ridge will universally dominate the other
- Lasso will perform better when not all the predictors affect the response or some variables are weakly associated with the response variable.
- Ridge performs better when the response variable is a function of many predictors, all with coefficients of roughly equal size.
- It is not easy to know the number of variables related to the response *a priori* and hence cross-validation is used to determine which approach is better.

Cross-validation to select λ

- General algorithm to select λ:
  - Select a range of values for λ
  - Compute the cross-validation error for each value of λ
  - Select the value of λ for which the cross-validation error is the smallest
  - Refit the model using all the available observations and the chosen/selected λ