Regularized regression II

Introduction

Introduction

Feature selection methods

- Wrapper
  - Search algorithms that add and/or remove predictors to optimize performance
  - e.g. forward, backward, and best subset selection
- Filter
  - Test individual predictors outside of the predictive model
  - e.g. t-tests, r, χ^2
- ℓ_1 **regularization**

Ridge & Lasso

Ridge regression

$$\hat{\beta}_{ridge} = \underset{\beta}{\text{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

- Penalty on ℓ_2 norm of β
- $\|\beta\|_2 = \sqrt{\sum_{j=1}^{p} \beta_j^2}$

Lasso (Least Absolute Shrinkage and Selection Operator)

$$\hat{\beta}_{lasso} = \underset{\beta}{\text{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

- Penalty on ℓ_1 norm of β
- $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$

Ridge & Lasso

Ridge regression

$$\hat{\beta}_{ridge} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

- Penalty on ℓ_2 norm of $\boldsymbol{\beta}$
- $\|\boldsymbol{\beta}\|_2 = \sqrt{\sum_{j=1}^{p} \beta_j^2}$

Lasso (Least Absolute Shrinkage and Selection Operator)

$$\hat{\beta}_{lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

- Penalty on ℓ_1 norm of $\boldsymbol{\beta}$
- $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$

Ridge & Lasso

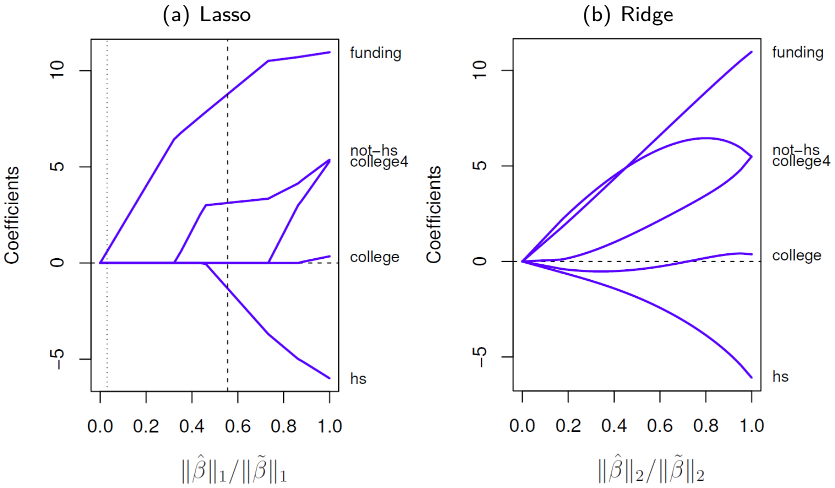
Increasing the penalty on model complexity

- $\lambda = 0$
  - Models are equivalent to OLS
- $\lambda \to \infty$
  - Ridge regression ($RSS + \lambda ||\beta||_2^2$)
    - Coefficients are shrunken towards zero
    - Shrinks coefficients of correlated predictors towards each other
  - Lasso ($RSS + \lambda ||\beta||_1$)
    - Coefficients are eventually shrunken exactly to zero (i.e. performs **variable selection**)
    - Erratic paths for correlated predictors

$\rightarrow$ The penalty λ is a tuning parameter

Ridge & Lasso

Figure: Coefficient paths



Efron & Hastie (2016)

Motivation for Extensions

Limitations of Lasso and Ridge Regression:

- Lasso can be too aggressive in setting coefficients to 0.
- With highly correlated predictors, Lasso tends to choose one and ignore others.
- On the other hand, Ridge Regression doesn't actually set predictors to 0, so limited use in variable selection.