Introduction to Machine Learning – Shrinkage Methods

Dr. Ab Mosca (they/them)

Slides based off slides courtesy of Jordan Crouser (https://jcrouser.github.io/)

Plan for Today

- Shrinkage Methods
  - Ridge Regression
  - The Lasso
  - Selecting the Tuning Parameter

Warm Up: Subset Selection

Form 3 groups.

With your group write out the algorithm for your assigned subset selection method.

1. Best Subset
2. Forward Selection
3. Backward Selection

What are the pros and cons of your method?

Motivation

So far, we looked at methods that determine good subsets of predictors to use when fitting linear models using least squares.

Motivation

So far, we looked at methods that determine good subsets of predictors to use when fitting linear models using least squares.

An alternative approach is to fit a model containing all p predictors, but to ***constrain*** or ***regularize*** the coefficient estimates.

Ridge Regression

Big idea: minimize RSS plus an additional penalty that rewards small (sum of) coefficient values.

Least squares fits by finding coefficients that minimize

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

Ridge Regression

Big idea: minimize RSS plus an additional penalty that rewards small (sum of) coefficient values.

Least squares fits by finding coefficients that minimize

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

Ridge Regression fits by finding coefficients that minimize

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

Shrinkage Penalty

where $\lambda \geq 0$ is a tuning parameter determined separately

What will the shrinkage penalty reward?

Ridge Regression

Big idea: minimize RSS plus an additional penalty that rewards small (sum of) coefficient values.

Least squares fits by finding coefficients that minimize

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

Ridge Regression fits by finding coefficients that minimize

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

Shrinkage Penalty

where $\lambda \geq 0$ is a tuning parameter determined separately

What does λ do in this equation? What happens when it is small (near 0)? Large (near infinity)?

Ridge Regression

Big idea: minimize RSS plus an additional penalty that rewards small (sum of) coefficient values.

Least squares fits by finding coefficients that minimize

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

Ridge Regression fits by finding coefficients that minimize

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

Shrinkage Penalty

where $\lambda \geq 0$ is a tuning parameter determined separately

RSS is scale invariant (multiplying any predictor by a constant won't change RSS). Is the shrinkage penalty?

Ridge Regression

Ridge Regression fits by finding coefficients that minimize

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

where $\lambda \geq 0$ is a tuning parameter determined separately

Why would ridge regression improve the fit over least-squares regression?

Ridge Regression

Ridge Regression fits by finding coefficients that minimize

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

where $\lambda \geq 0$ is a tuning parameter determined separately

Why would ridge regression improve the fit over least-squares regression?

- Ridge regression works best in situations where the least-squares estimates have high variance. It trades a small increase in bias for a large reduction in variance.

Ridge Regression

Drawback:

- Does not actually perform variable selection
- Our final model will include all predictors
  - If all we care about is prediction accuracy, this is not a problem
  - However, if we also care about model interpretability, it does pose a problem

The Lasso

Big idea: minimize RSS plus an additional penalty that rewards small (sum of) coefficient values.

Least squares fits by finding coefficients that minimize

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

The Lasso fits by finding coefficients that minimize

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

Shrinkage Penalty

where $\lambda \geq 0$ is a tuning parameter determined separately

The Lasso

Big idea: minimize RSS plus an additional penalty that rewards small (sum of) coefficient values.

Least squares fits by finding coefficients that minimize

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

The Lasso fits by finding coefficients that minimize

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

Shrinkage Penalty – lasso uses an ℓ_1

where $\lambda \geq 0$ is a tuning parameter determined separately

The Lasso

How does the lasso get coefficients exactly equal to 0?

For each value of λ, there exists a value for s such that:

- Ridge Regression

$$\min_{\beta}(RSS) \text{ subject to } \sum_{j=1}^{p} \beta_j^2 \leq s$$

- Lasso

$$\min_{\beta}(RSS) \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq s$$

The Lasso

How does the lasso get coefficients exactly equal to 0?

For each value of λ, there exists a value for s such that:

- Ridge Regression

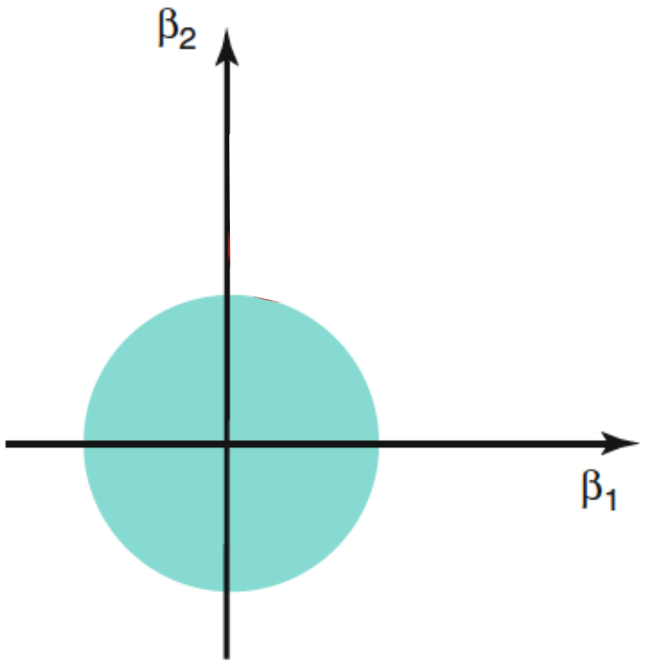
$$\min_{\beta}(RSS) \text{ subject to } \sum_{j=1}^{p} \beta_j^2 \leq s$$

- Lasso

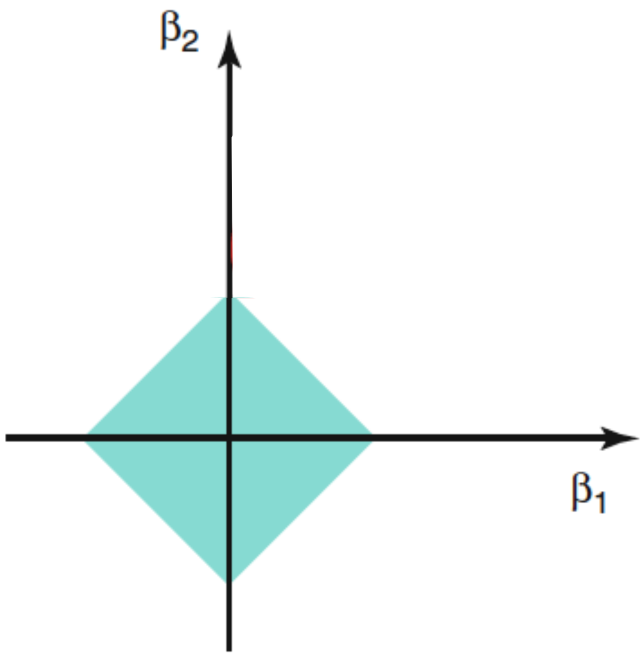
$$\min_{\beta}(RSS) \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq s$$

Consider the case where $p = 2$, what does s work out to in each case?

The Lasso

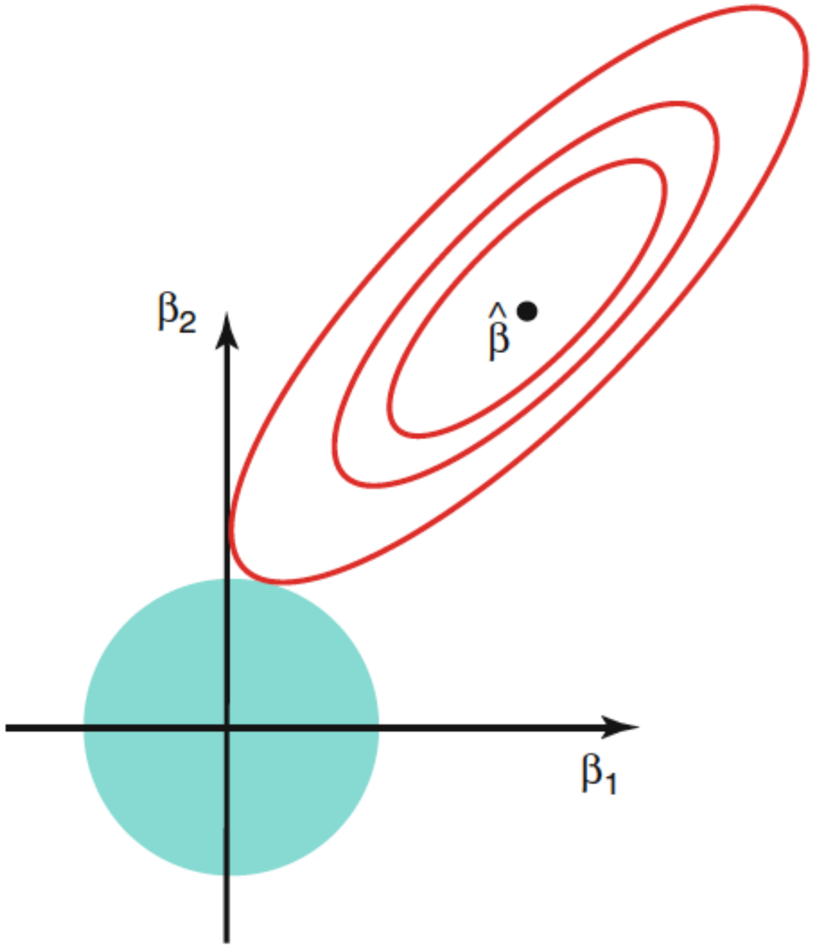


Ridge regression

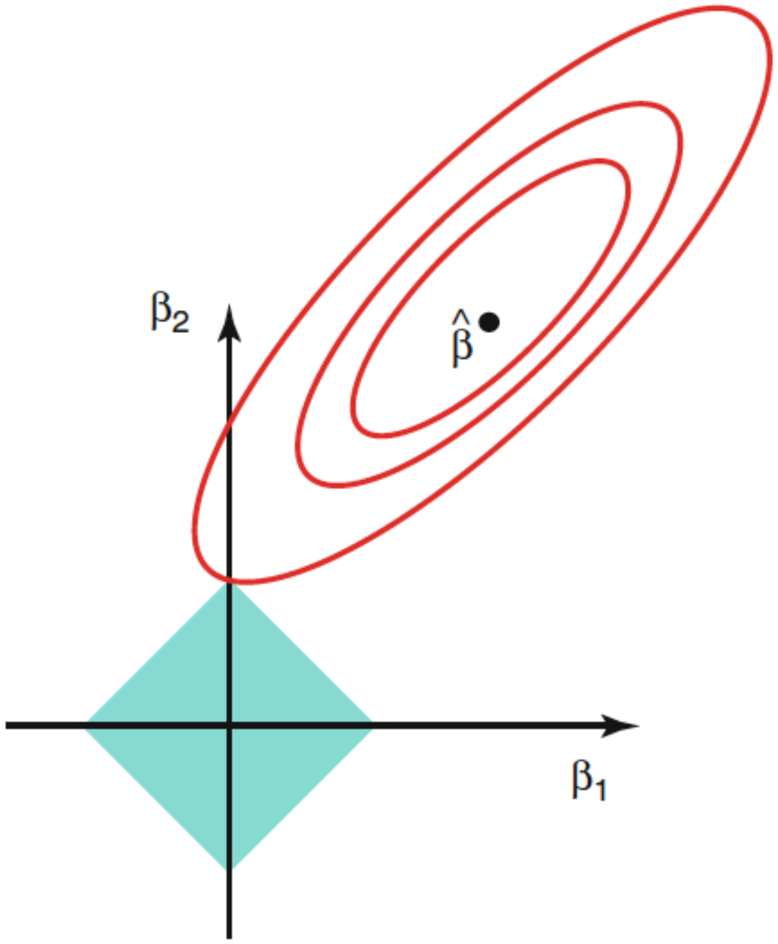


Lasso

The Lasso



Ridge regression



Lasso

Tuning Parameter Selection

- We choose the appropriate λ using cross validation
  - Choose a grid of λ
  - Use CV to compute test error for each
  - Select the λ for which CV test error is smallest

Inventory

So far, we've talked about solving the problem of $n \leq p$ in linear regression via subset selection (best and stepwise approaches), and via shrinkage methods (ridge regression and the lasso).

Let's recap and take inventory of these options. Break into 4 groups. Each group will be assigned one option.

With your group:

- If you've been assigned a subset selection approach
  - Write out the algorithm in pseudo code
  - Visualize the algorithm running on an example
  - ID the pro's and con's to your approach compared to others
- If you've been assigned a shrinkage method
  - Write out the shrinkage penalty for your approach
  - Show how the penalty works with an example
  - ID the pro's and con's to your approach compared to others
- Be prepared to share with the class