Shrinkage Methods: Ridge regression, LASSO

Mingjun Zhong

Department of Computing Science

University of Aberdeen

Motivation

- Remember linear regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

  - Least-square estimate (LSE):

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

- Why we are not happy with LSE?
  - ✓ ***Interpretation***: Large number of predictors, which predictors are more important?
  - ✓ For example: the <u>location</u> or the <u>size</u> of the house is more important for house prices?
- Mitigations – *subset selection* or *shrinkage*:
  - ✓ ***Interpret*** model: determine a small subset of parameters; the strongest effects; sacrifice small details

Best-subset selection

- Number of predictors p, subset of size $k \in \{0,1,2,\cdots,p\}$
- House price example:
  - p =2
  - X1: location; X2: size
  - Y: price
  - Then K=0,1, or 2
  - All possible models:
    - When k=0, $Model1$: $y = \beta_0 + \epsilon$
    - When k=1, $Model2$: $y = \beta_0 + \beta_1 x_1 + \epsilon$, and $Model3$: $y = \beta_0 + \beta_1 x_2 + \epsilon$
    - When k=2, $Model4$: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$
  - There are four models, which is the best?

Shrinkage methods

- Subset selection:
  - Pros: Interpretable
  - Cons: Discrete

- Shrinkage methods:
  - ***Ridge Regression***
  - ***The Lasso***
  - They are continuous

Linear regression

- Linear regression loss function:

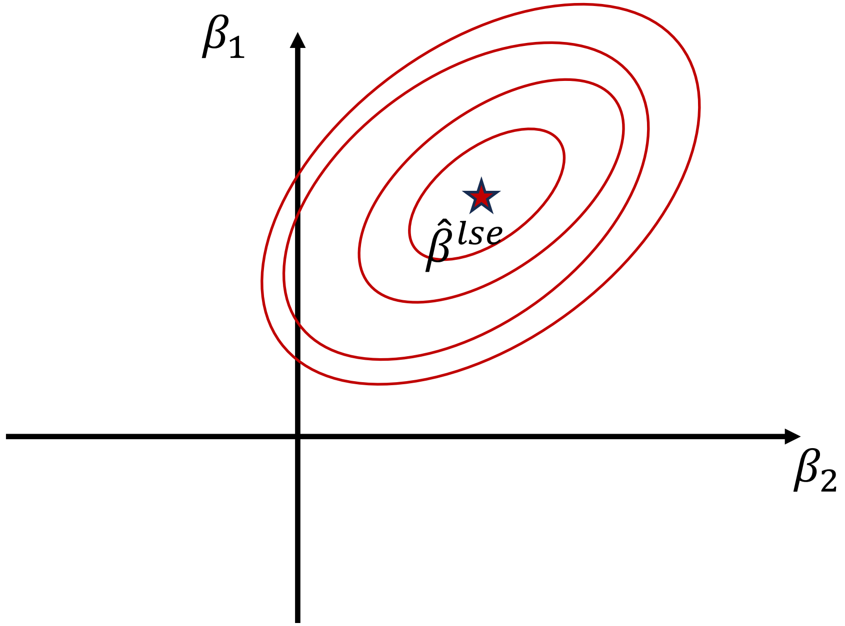
$$loss(\beta) = \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$

- The best fit is to find $\hat{\beta}^{lse}$ (*lse: least squared errors*) to minimize the loss:

$$\hat{\beta}^{lse} = argmin_{\beta}[loss(\beta)]$$

- Notation $argmin_{\beta}$ means finding the argument β to minimize the loss.
- There is a unique solution $\hat{\beta}^{lse}$.
- Roughly, if $\beta_1 > \beta_2$, then x_1 (location) is more important than x_2 (size) for interpreting y.
- Both β_1 and β_2 **are not zero** – generally true.
- But we want to select a subset of predictors for predicting house prices – we want to turn some variables to zero, and so a subset is selected.

Loss function



House price example:

x_1: location of a house

x_2: the size of a house

y: house price

Ridge regression

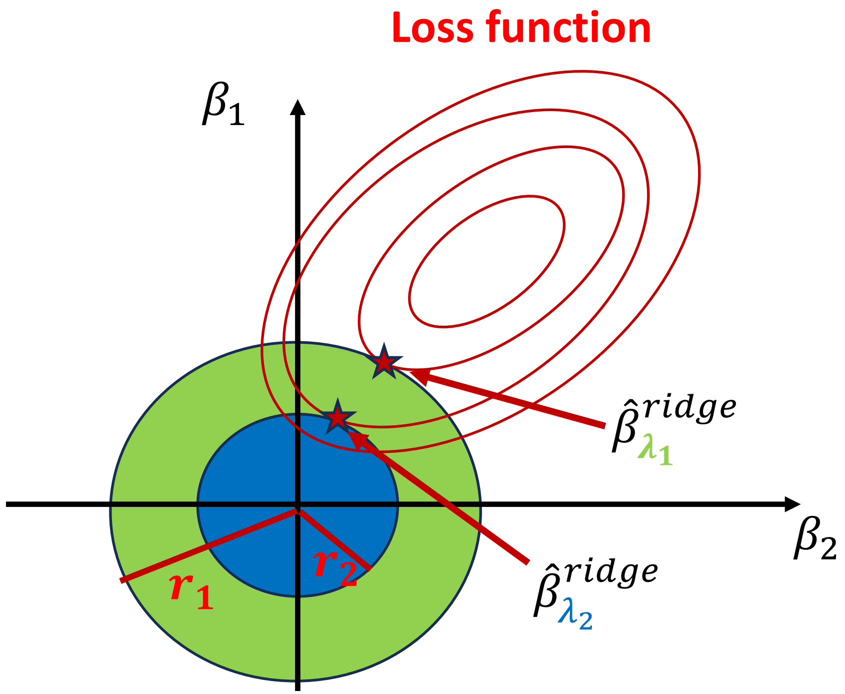
- Ridge regression loss function:

$$loss(\beta) = \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

- The best fit is to find $\hat{\beta}^{ridge}$ to minimize the loss:

$$\hat{\beta}^{ridge} = argmin_{\beta}[loss(\beta)]$$

- $\lambda \geq 0$: complexity parameter – controls the amount of shrinkage
  - Larger the value, greater the amount of shrinkage
    - Larger λ (= smaller radius r), second factor dominates, all non-negative thus shrinks
  - Also used in Deep Neural Networks – called ***weight decay***
- There is a **unique** solution $\hat{\beta}^{ridge}$ for any λ (or r).
- Both β_1 and β_2 **are not zero** – generally true. But some are very close to zero.
- But we want to select a subset of predictors for predicting house prices – we want to turn some variables to zero, and so a subset is selected.



House price example:

x_1: location of a house

x_2: the size of a house

y: house price

LASSO regression

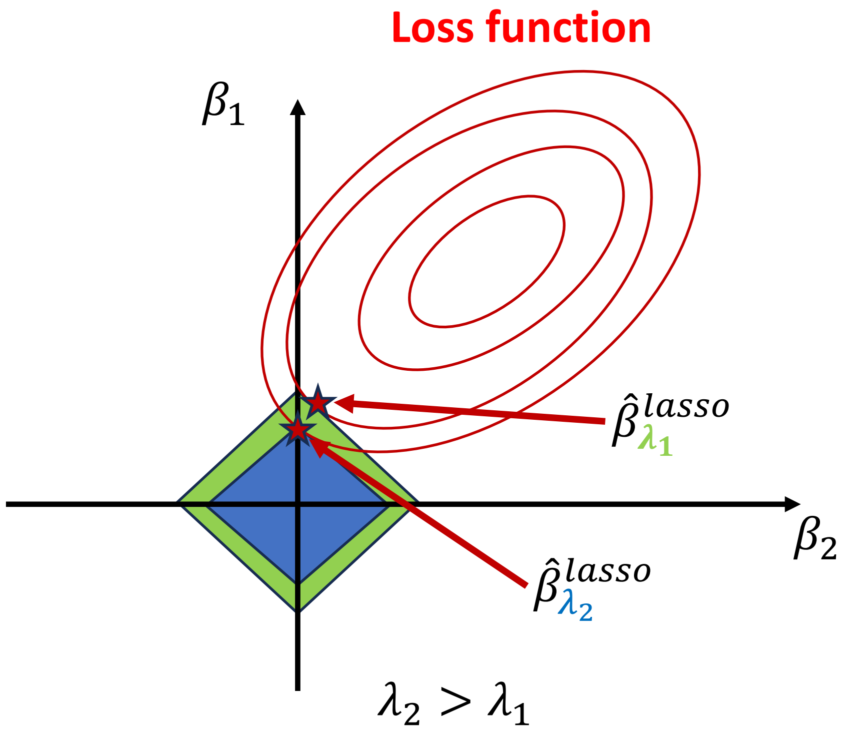
- Ridge regression loss function:

$$loss(\beta) = \sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

- The best fit is to find $\hat{\beta}^{lasso}$ to minimize the loss:

$$\hat{\beta}^{lasso} = argmin_{\beta}[loss(\beta)]$$

- $\lambda \geq 0$: complexity parameter – controls the amount of shrinkage
  - Larger the value, greater the amount of shrinkage
    - Larger λ, second factor dominates, all non-negative thus shrinks
  - Also used in Deep Neural Networks – called ***weight decay***
- **No closed solution** $\hat{\beta}^{lasso}$ for any λ (or r).
- For suitable λ, one of β_1 and β_2 **is very close to zero (nearly)** – generally true.
- When there are many variables, many of them would **be zero**.
- Need to choose the right λ.
- A subset is selected.
- Solution can be found by using sklearn.linear_model.Lasso



House price example:
x_1: location of a house
x_2: the size of a house
y: house price

How to choose λ?

- Larger λ, more chances some of the β_js are close to 0, and so those features x_js are not important for the model
- $\lambda = 0$, Lasso = ridge = LSE (least squared errors).
- We need to choose best λ:
  $$\lambda = 0, 0.0001, 0.001, 0.0011, \cdots, 0.01, \cdots, 0.1, 0.2, \cdots, 0.9, 1, \cdots$$
- Each λ corresponds to one model, we need to use model selection methods like Cross-Validation to compare those models. (Covered later)

Summary

- Subset selection:
  - Pros: Interpretable
  - Cons: Discrete
- Shrinkage methods:
  - ***Ridge Regression***
  - ***The Lasso***
  - They are continuous
  - Need to select the best complexity parameter λ
  - Interpretable: can choose important variables/predictors
  - The idea is widely used in deep learning (weight decay)
- Ridge and Lasso methods can turn off some predictors and thus subset selection