Discussion 2

王欣奕，wangxy6@shanghaitech.edu.cn

Review

- Linear regression models
- The Gauss-Markov theorem
- Subsets selection
- Shrinkage Methods: Ridge Regression and the Lasso

Linear regression models

- A linear regression model assumes that the regression function E(Y| X) is linear in the inputs.

1. Simple linear regression:

$$f(x) = \beta_0 + \beta x$$

$$\hat{\beta}_0, \hat{\beta} = argmin \sum_{i=1}^{n} (y_i - \beta_0 - \beta x_i)^2$$

2. Multiple linear regression:

$$f(x) = \beta_0 + \sum_{j=1}^{p} x_j \beta_j$$

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{i,j} \beta_j)^2 = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)$$

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

3. Multiple Output regression

The Gauss-Markov Theorem

The least squares estimator has the lowest sampling variance within the class of linear unbiased estimators.

Remarks

- Among the unbiased linear methods, least squares has the lowest MSE
  - MSE = Var + Bias^2
- A biased methods probably has lower MSE
  - Var-Bias trade-off
  - A small increase in Bias might gives rise to a large reduction in Var ← Model selection

Two limitations of least squares

- prediction accuracy
  - low bias and high variance

    → sacrifice a little bias to reduce the variance
- interpretation
  - hard to interpret a large number of input features

    → find a subset of features exhibiting strong effects

We need Model Selection !

Subset selection

- Best-subset selection
  - For each $s \in \{0,1,\ldots,p\}$, find the subset in size of s that gives lowest $\mathrm{RSS}(\beta) = \|\mathbf{y} - \mathbf{X}^{(s)}\beta\|_2^2$

We always choose the smallest model that minimizes an estimate of the expected prediction error.

Subset selection

- Forward-stepwise
  - starts with intercept
  - sequentially adds the best predictor
- Greedy algorithm
  - sub-optimal
- Advantages
  - Computational
    - even $p \gg N$
  - Statistical
    - constrained search
    - lower variance, more bias

- Backward-stepwise
  - starts with the full model
  - sequentially deletes the worst predictor
- Greedy algorithm
- Only useful when $N > p$
  - linear regression

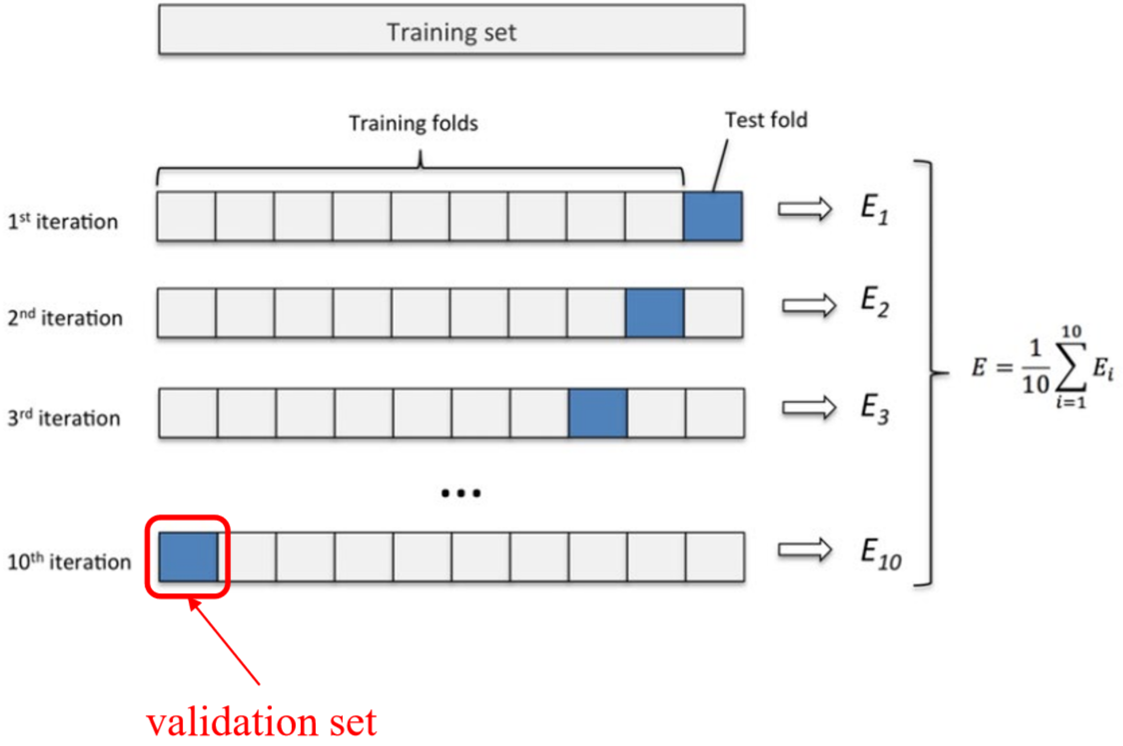
- Smart stepwise
  - group of variables
  - add or drop whole groups at a time

K-Fold Cross-Validation

- Each has a complexity parameter λ
  - the subset size in subset selection
  - the neighborhood size in k-NN
  - The coefficient of regularization
- K-fold cross validation
  - divide the training data into K roughly equal parts ($K = 5$ or 10)
  - for $k = 1, \ldots, K$,
    - fit the model with $K - 1$ parts
    - compute the error E_k on the rest part
  - The K-fold cross validation error

$$E(\lambda) = \frac{1}{K} \sum_{k=1}^{K} E_k(\lambda)$$

Repeat this for many values of λ, and choose the best value that makes $E(\lambda)$ lowest.



Shrinkage Methods

Ridge Regression

$$\hat{\beta}^{ridge} = \mathrm{argmin}_{\beta} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_2^2$$

- Can solve the problem of overfitting
- Has closed form solution: $\hat{\beta}^{ridge} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p\right)^{-1}\mathbf{X}^T\mathbf{y}$
- Can't get sparse model(close to 0 but not equal to 0)
- MAP with a prior $\Pr(\beta) = \mathcal{N}(\beta|0, \frac{1}{\lambda}\mathbf{I}_p)$ Gaussian distribution

(least absolute shrinkage and selection operator, 最小绝对值收敛和选择算子)

The Lasso

$$\hat{\beta}^{lasso} = \mathrm{argmin}_{\beta} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_1$$

- Can solve the problem of overfitting
- No closed form solution, needs PGD to solve it.
- Can get sparse model(can do feature selection)
- MAP with a prior $\Pr(\beta) = \frac{\lambda}{2}e^{-\lambda\|\beta\|_1}$ Laplacian distribution

Shrinkage Methods

Generalization of Ridge and Lasso

- Consider the criterion ($q \geq 0$)

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

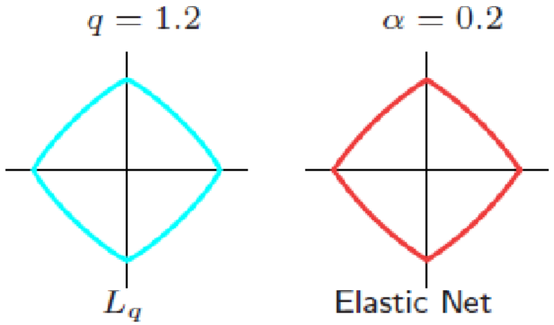
- $q = 0$, best subset
- $q = 1$, lasso
- $q = 2$, ridge regression

  - $q \in (1,2)$: a compromise between lasso and ridge regression
    - $|\beta_j|^q$ is differentiable at 0 $\rightarrow$ hard to set $\beta_j = 0, \forall j$

- Elastic-net

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

  - ℓ_2 shrinks the coefficients of correlated predictors
  - ℓ_1 selects groups of correlated predictors



Exercise

Ex. 3.30 Consider the elastic-net optimization problem:

$$\min_{\beta} ||\mathbf{y} - \mathbf{X}\beta||^2 + \lambda\big[\alpha||\beta||_2^2 + (1-\alpha)||\beta||_1\big]. \tag{3.91}$$

Show how one can turn this into a lasso problem, using an augmented version of $\mathbf{X}$ and $\mathbf{y}$.

Solution

Let the elastic-net problem be equation (1)

The lasso in matrix form: $\hat{\beta}^{lasso} = \underset{\beta}{\arg\min} \|y_1 - X_1\beta\|_2^2 + \lambda_1\|\beta\|_1$ (2)

$\therefore$ We need to change (1) into (2)

$$\because \begin{cases} \lambda_1\|\beta\|_1 = \lambda(1-\alpha)\|\beta\|_1 \\ \|y_1 - X_1\beta\|_2^2 = \|y - X\beta\|_2^2 + \lambda\alpha\|\beta\|_2^2 \end{cases} \Rightarrow \lambda_1 = \lambda(1-\alpha)$$

Then We need to use argumented version of X and y

Assume $X_1 = \begin{bmatrix} X \\ A \end{bmatrix}$ $y_1 = \begin{bmatrix} y \\ c \end{bmatrix}$

$$\therefore \|y_1 - X_1\beta\|_2^2 = \left\|\begin{bmatrix} y - X\beta \\ c - A\beta \end{bmatrix}\right\|_2^2 = \|y - X\beta\|_2^2 + \|c - A\beta\|_2^2$$

$$\therefore \|y - X\beta\|_2^2 + \lambda\alpha\|\beta\|_2^2 = \|y - X\beta\|_2^2 + \|c - A\beta\|_2^2 \Rightarrow c = 0 \quad A = \sqrt{\lambda\alpha}\, I$$

In short, if we let $y_1 = \begin{bmatrix} y \\ 0 \end{bmatrix}$, adding p zeros, p is the number of features,

$X_1 = \begin{bmatrix} X \\ \sqrt{\lambda\alpha} I \end{bmatrix}$, adding $\sqrt{\lambda\alpha} I$, in which I is a $P \times P$ identity matrix

$\lambda_1 = \lambda(1-a)$

then we can change elastic-net problem into a lasso problem.