Extended Models in Linear Regression

고태훈 (taehoonko@dm.snu.ac.kr)

Extended Models in Linear Regression

- Stepwise linear regression
- Ridge regression
- LASSO
- ~~ElasticNet~~

How to determine a set of predictors

- 모델에서 이용하는 입력 변수의 집합이 달라지면, 모델의 성능이 달라진다.
  - 어떤 입력 변수 집합이 가장 좋은 성능을 보일 것인가?
  - 이를 feature subset selection이라고 한다.

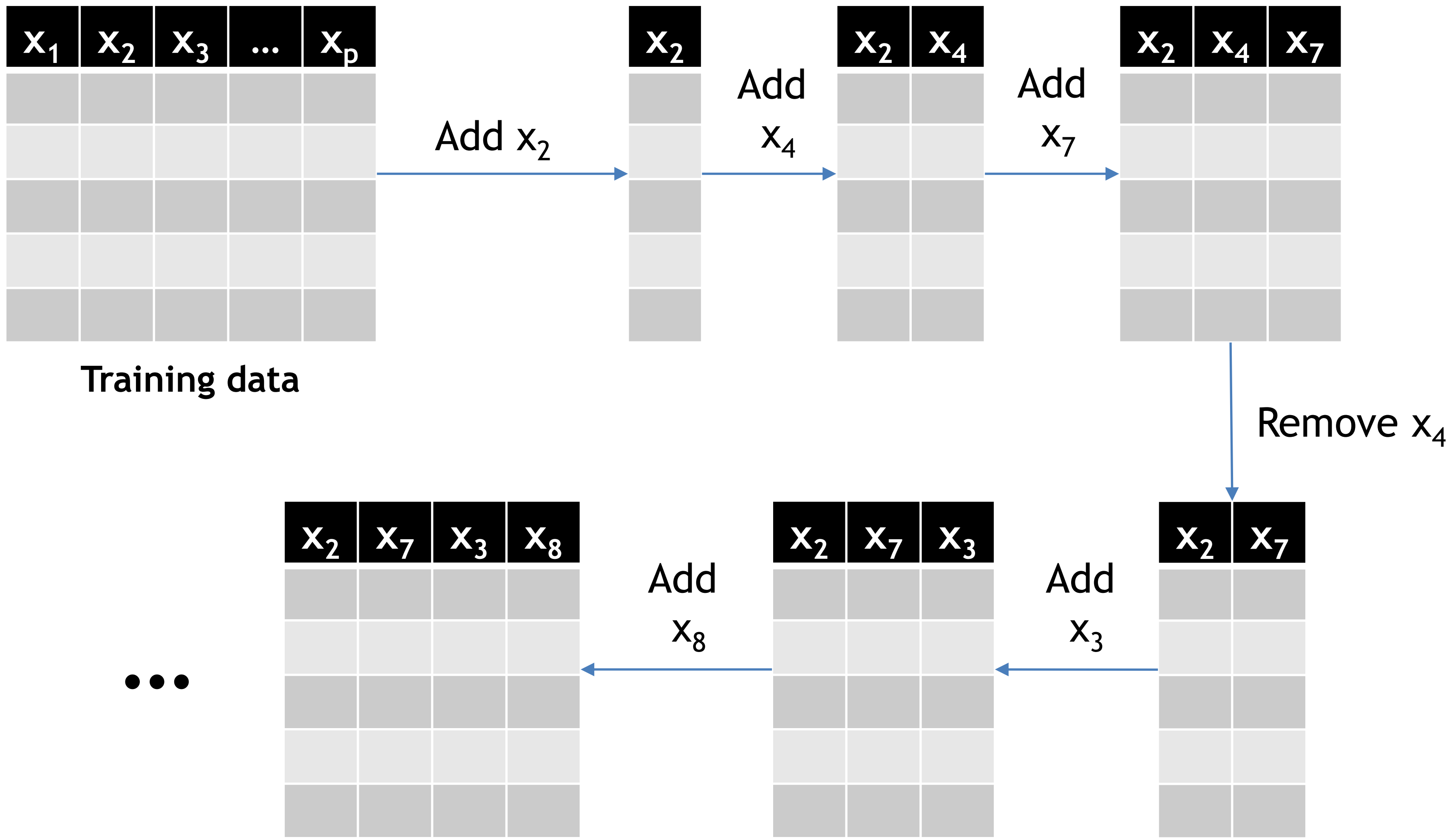
- Exhaustive search (전역 탐색)
  - The simplest method for finding an optimal feature subset
    : 모든 변수 집합을 탐색
  - But, we need too much time.
    : 변수의 수가 n개이면, 가능한 모든 부분집합의 수가 $2^n - 1$

Stepwise linear regression

❖ 단계적 선택법

- 입력변수 집합에 변수를 하나씩 추가하거나(전진선택법: Forward selection) 하나씩 제거하는(후진소거법: Backward elimination) 과정을 반복함
- 입력변수 집합이 생성될 때마다 선형회귀모델을 학습하고 이를 평가하여 최적의 입력변수 집합을 탐색
- 한 번 선택되거나 제거된 변수가 다시 선택/제거될 수 있음

Stepwise linear regression



Stepwise linear regression: Algorithm

Initialize:

- Start with model with no input variables.
- *Selected* = null

Loop

- For each variable which is not in *Selected*:
  - *Selected* = *Selected* + candidate variable
  - Build submatrix of X using *Selected*
  - Train a linear regression model and evaluate it.
- Find the best model and responding *Selected*.

Forward selection phase

- For each variable which is in *Selected*:
  - *Selected* = *Selected* - candidate variable
  - Build submatrix of X using *Selected*
  - Train a linear regression model and evaluate it.
- Find the best model and responding *Selected*.

Backward elimination phase

Stepwise linear regression

❖ How to evaluate candidate linear regression models (1)

- Akaike Information Criteria (AIC)
- Bayesian Information Criteria (BIC)
- Adjusted-R^2: 기존의 R^2에 변수의 수를 고려
- Mallow's C_k

$$\mathrm{AIC} = n \cdot \ln\left(\frac{\mathrm{SSE}_k}{n}\right) + 2k$$

$$\mathrm{BIC} = n \cdot \ln\left(\frac{\mathrm{SSE}_k}{n}\right) + k \cdot \ln(n)$$

$$\text{Adjusted-}R^2 = 1 - \left(\frac{n-1}{n-k-1}\right)(1 - R^2)$$

$$C_k = \frac{SSE_k}{s^2} - (n - 2k)$$

n : number of samples

k : number of selected variables

SSE_k : sum of squared error of regression model with k variables

s : sum of squared error of full regression model

Stepwise linear regression

❖ How to evaluate candidate linear regression models (2)

- Using train error
  - 앞에서의 AIC, BIC, Mallow's C_k, Adjusted-R^2 와 마찬가지로 Regression model이 학습데이터에 잘 적합했는가를 살펴보는 지표
- Using validation / test error
  - Regression model이 앞으로 새롭게 발생하는 데이터의 Y를 얼마나 잘 예측할 것인가를 살펴보는 지표

Regularization

❖ Regularization (제약)

- 기계학습으로 학습한 모델의 복잡도에 대한 제약/페널티를 부여
- 하는 이유?
  - 학습 데이터에 너무 과적합(overfitting)하여, 새롭게 등장하는 데이터에 대한 예측 성능이 떨어지는 것을 방지 ➔ "Generalization"
  - 더 자세한 내용은 추후 [편향-분산 트레이드오프 (Bias-variance tradeoff)] 에서 더욱 자세히 다룰 예정

Regularization

❖ Regularization (제약)

- 학습을 위한 최적화 문제의 목적식에 penalty term 추가
- For regression models,

$$\min \sum_{i=1}^{n} \{ y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_p x_{ip})^2$$

$$\min \sum_{i=1}^{n} \{ y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_p x_{ip})^2 + \boxed{\sum_{j=1}^{p} \left| \beta_j \right|}$$

$$\min \sum_{i=1}^{n} \{ y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_p x_{ip})^2 + \boxed{\sum_{j=1}^{p} \beta_j{}^2}$$

Ridge regression (능형회귀분석)

❖ Ridge regression의 회귀계수

$$\hat{\boldsymbol{\beta}}_{ridge} = \arg\min_{\boldsymbol{\beta}} \left\{ \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|^2 \right\}$$

$$= ((\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda \mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$$

$$\min_{\boldsymbol{\beta}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 \quad \text{subject to } \|\boldsymbol{\beta}\|^2 = \sum_{j=1}^{p} \beta_j^{\ 2} \le s$$

- 계수의 크기에 대한 L2-norm penalty를 부여하여 모델의 overfitting을 방지

Lasso regression

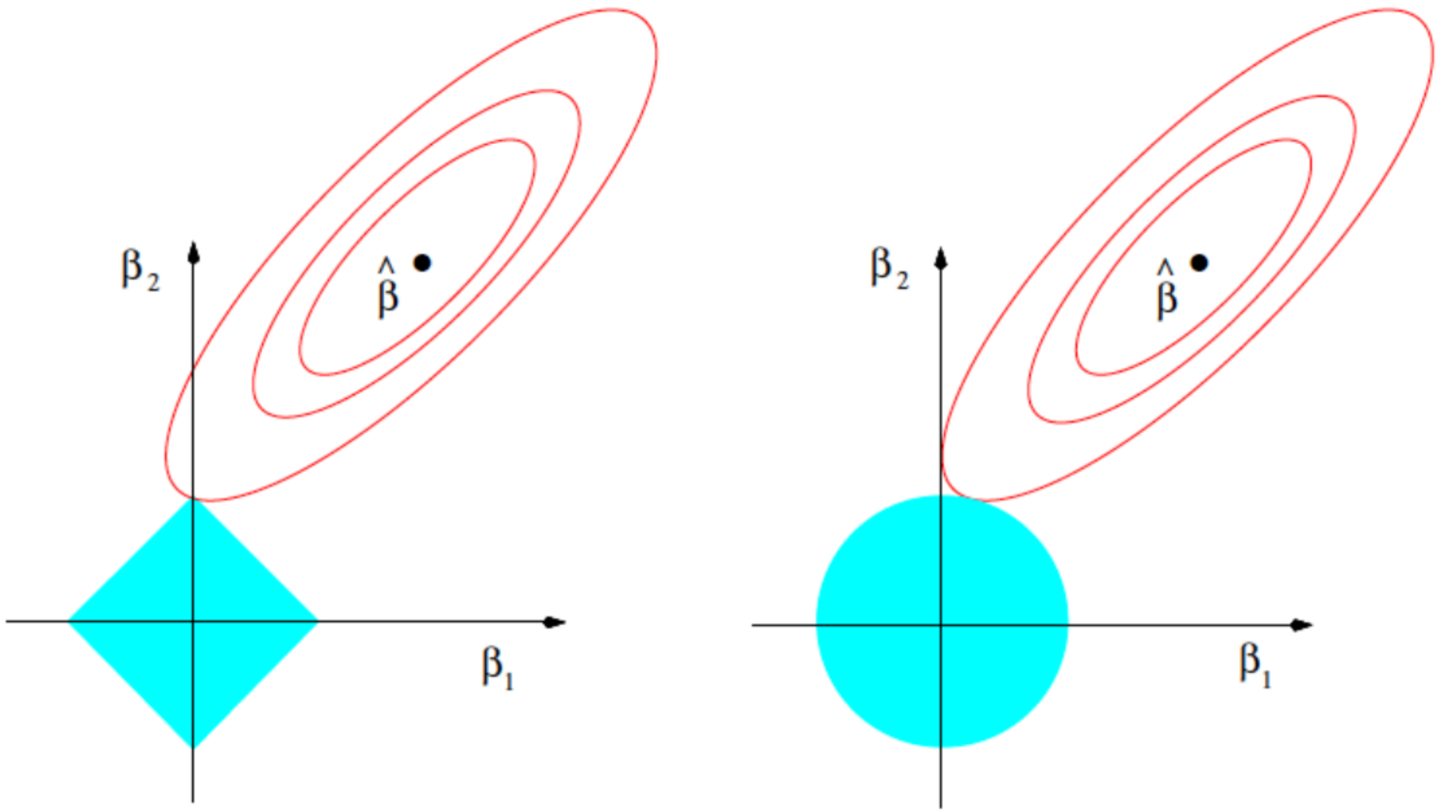
- **Least absolute shrinkage and selection operator (LASSO) regression**

$$\hat{\boldsymbol{\beta}}_{lasso} = \arg\min_{\boldsymbol{\beta}} \left\{ \left\| \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \right\|^2 + \lambda \left\| \boldsymbol{\beta} \right\| \right\}$$

$$\min_{\boldsymbol{\beta}} \left\| \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \right\|^2 \quad \text{subject to} \ \left\| \boldsymbol{\beta} \right\| = \sum_{j=1}^{p} \left| \beta_j \right| \le t$$

- 계수의 크기에 대한 L1-norm penalty를 부여하여 모델의 overfitting을 방지
- Lasso regression은 전체 입력변수 계수 중 일부를 0으로 만들어 입력변수를 선택하는 효과가 있음 ➔ Sparse modeling

Lasso regression vs. Ridge regression



T. Hastie, R. Tibshirani, J. Friedman. *The elements of statistical learning : data mining, inference , and prediction*. Springer, 2011