AI Project

2. 과적합 문제와 정규화

GNB 김도훈

Index

1. 1주차 교육 내용 복습
2. 모델 복잡도와 문제 복잡도
3. 과소적합, 과적합 문제
4. 정규화 (Regularization)
5. Lab 2 : Regularized Regression

복습: 연습 문제

예제 1) 입력 데이터가 다음 표와 같이 주어진다.

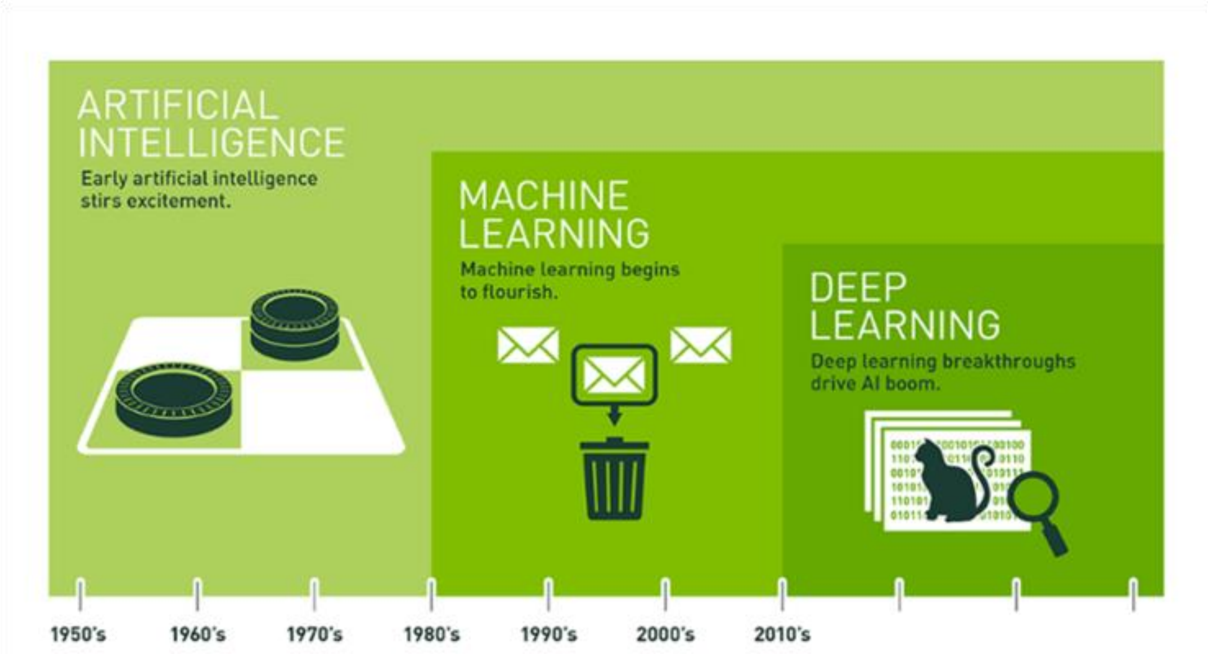
x1	x2	y
2	10	9

선형 회귀 모델이 다음과 같이 주어진다.

$$h_\theta(x) = \theta^T x \quad s.t. \;\; \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}$$

오차함수 E가 Sum of Squared Error일 때, $\frac{\partial E}{\partial \theta_1}$을 구하여라.

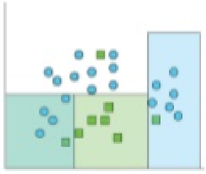
복습: 인공지능, 기계학습, 딥러닝



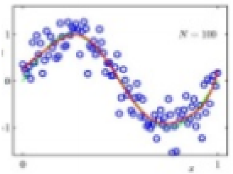
- 인공지능 : 인간의 사고방식을 모방하는 기계를 구현한다.
- 기계학습 : 경험으로부터 학습할 수 있는 소프트웨어를 구현한다.
- 딥러닝 : 깊게 쌓은 모델로 특정 분야에서 인간을 뛰어넘은 능력을 구현한다.

복습: 기계학습의 범주

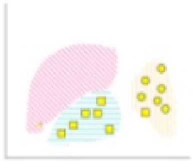
- 기계학습 (Machine Learnig)
  - 지도학습 (Supervised Learning)
    - 레이블된 데이터로 학습
    - 미래 데이터 예측
    - y = f(x)
    - 분류 (classification)

      - MNIST 필기체 인식
      - 스팸 메일 분류
    - 회귀 (regression)

      - 주가 예측
  - 비지도학습 (Unsupervised Learning)
    - 레이블 없이 학습
    - 데이터의 숨겨진 구조/특징 발견
    - x ~ p(x), x = f(x)
    - 클러스터링 (clustering)

      - 마케팅 고객 그룹화
  - 강화학습 (Reinforcement Learning)
    - 보상 시스템으로 학습
    - 의사결정을 위한 최적의 액션 선택

복습: 선형 회귀 (Linear Regression)

- 가설 함수 (Hypothesis function)

$$\hat{y} = h_\theta(x) = \theta_0 x_0 + \cdots + \theta_N x_N = [\theta_0 \quad \cdots \quad \theta_N] \begin{bmatrix} x_0 \\ \vdots \\ x_N \end{bmatrix} = \theta^T x \quad (x_0 = 1)$$

- 문제 정의 : 비용 극소화 문제(cost minimization problem)

  오차 함수 $E(\theta)$를 극소화시키는 θ구하기 (θ^* : 최적해 optimal solution)

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\, E(\theta) = \underset{\theta}{\operatorname{argmin}} \frac{1}{2} \sum_{i=0}^{I} \left(\hat{y}^{(i)} - y^{(i)}\right)^2$$

복습: 최소 제곱법 (Ordinary Least Square)

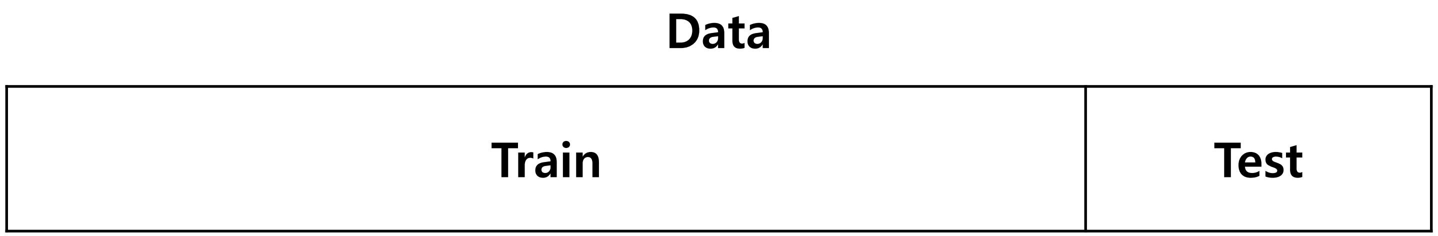
$$\theta^* = \left(x^T x\right)^{-1} x^T y$$

- 선형 회귀 문제의 닫힌 형태의 최적해를 유도한 것
- 최소 제곱법의 결과는…
  1. 선택한 모델($h_\theta(x)$: 선형 모델)과
  2. 훈련 데이터셋(x, y)에 대한

  최적의 파라미터 값이다.

⇒ 훈련 데이터셋에 없는 새로운 데이터에 대해서도 θ^*가 유효한 값인가?

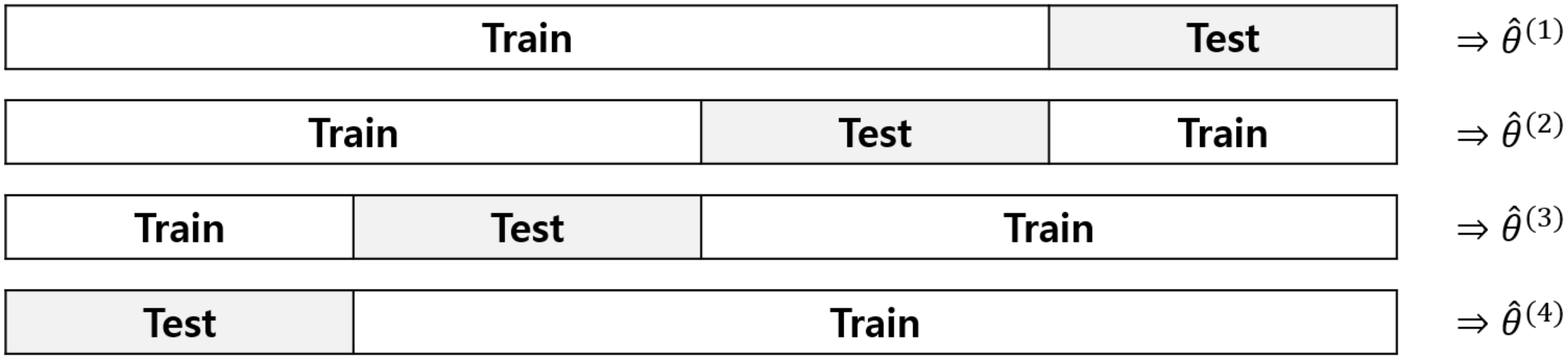
Train / Test data



- Train set : 최적해 θ^*를 구하기 위한 데이터
- Test set : 새로운 데이터에 대한 θ^*의 적합도를 평가하기 위한 데이터

- Train error와 Test error를 비교하여 학습이 잘 되었는지 평가하기 위함
- 학습에 사용하지 않은 데이터로 모델의 성능을 평가

교차 검증 (Cross validation)

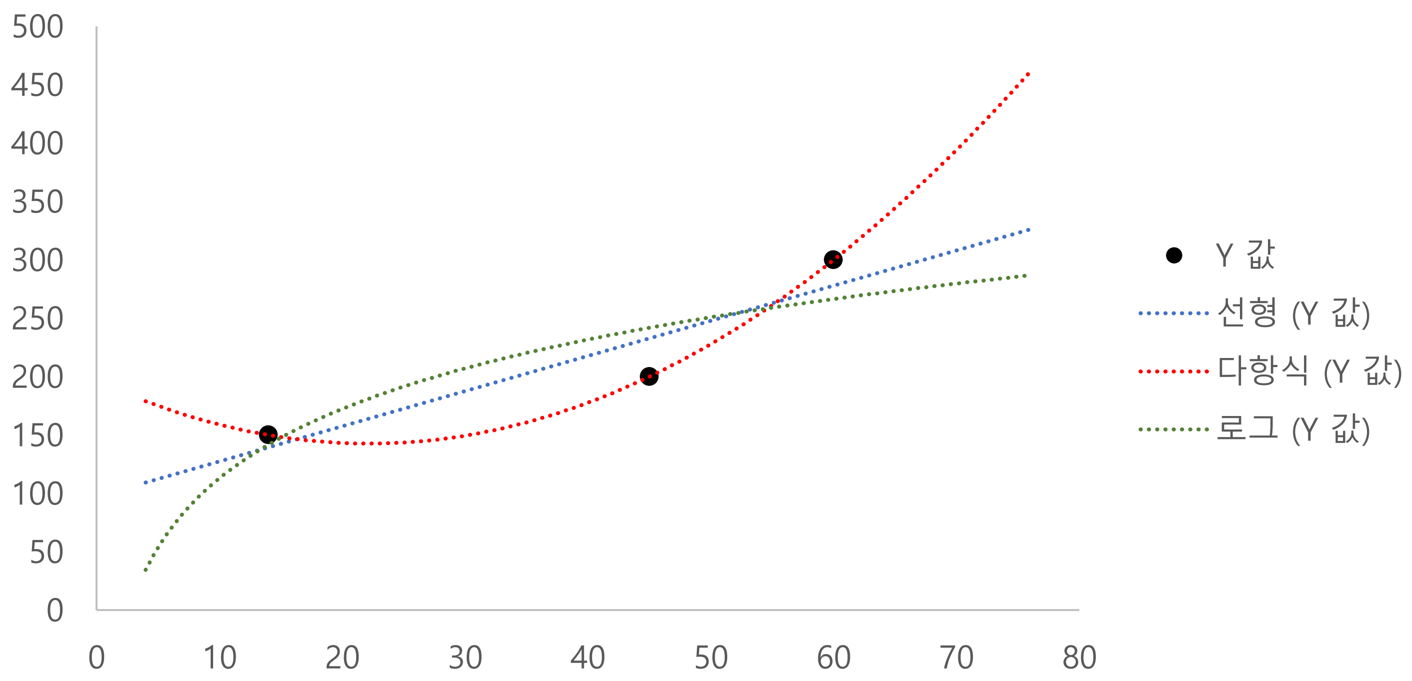


대표적인 교차 검증법인 K-Fold Cross Validation

- 새로운 데이터에도 정확도가 가장 좋은 $\hat{\theta}$를 선택하여 사용
- 관측되지 않은 데이터에 대한 일반성(모델의 신뢰도)를 높이기 위함

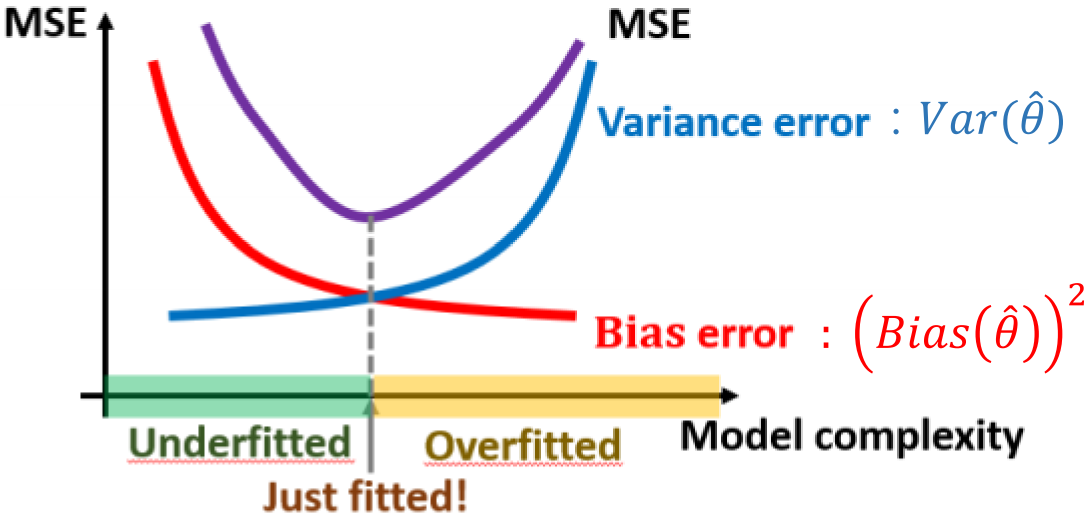
모델 복잡도 (Model complexity)

#	X	Y
1	14	150
2	45	200
3	60	300



- 이때까지 다룬 선형 모델 외에도 많은 복잡한 모델들이 존재
- 주어진 데이터셋에 대해 최고의 성능을 내는 모델을 선택해야함

모델 복잡도 (Model complexity)



- Variance error (분산)
  test data로 구한 error

- Bias error (편향)
  train data로 구한 error

- 모델 복잡도 = 모델의 표현력
- 모델이 복잡하다고 무조건 성능이 좋은 것이 아니다!
- 부족하지도, 과하지도 않은 적절한 복잡도의 모델을 선택해야함

모델 복잡도 (Model complexity)

$$\mathrm{MSE}\left(\hat{\theta}\right) = \mathrm{Var}\left(\hat{\theta}\right) + \left(Bias\left(\hat{\theta}\right)\right)^2$$

Proof)

$\theta \triangleq true\ parameters\ (deterministic)$

$\hat{\theta} \triangleq estimated\ parameters\ (random)$

$$Var(X) = E[X^2] - (E[X])^2, \qquad X = \hat{\theta} - \theta$$

$$Var\left(\hat{\theta} - \theta\right) = E\left[\left(\hat{\theta} - \theta\right)^2\right] - \left(E\left[\hat{\theta} - \theta\right]\right)^2$$

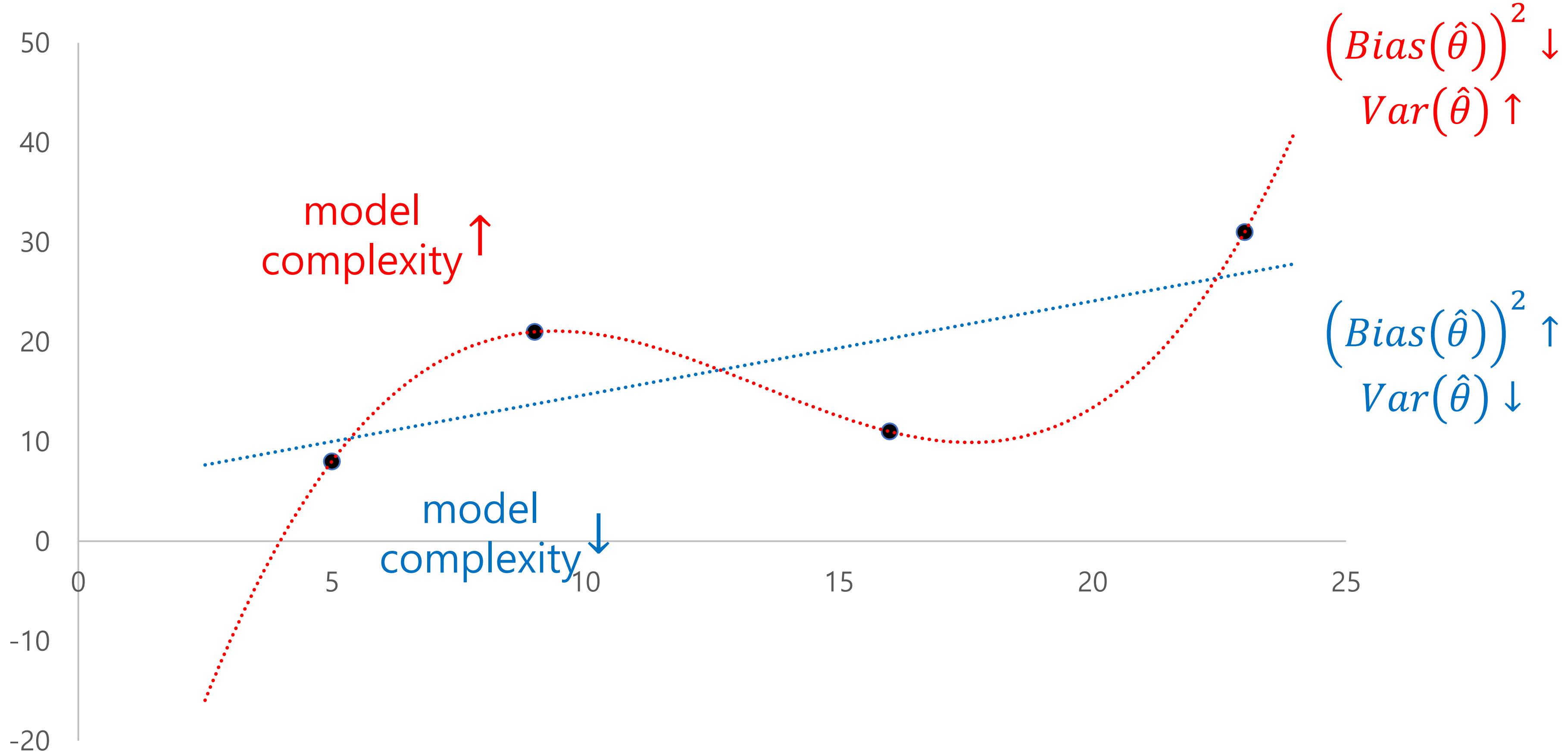
$$Var\left(\hat{\theta} - \theta\right) = Var\left(\hat{\theta}\right) \quad (\because \theta : detetministic)$$

$$E\left[\left(\hat{\theta} - \theta\right)^2\right] = MSE(\hat{\theta})$$

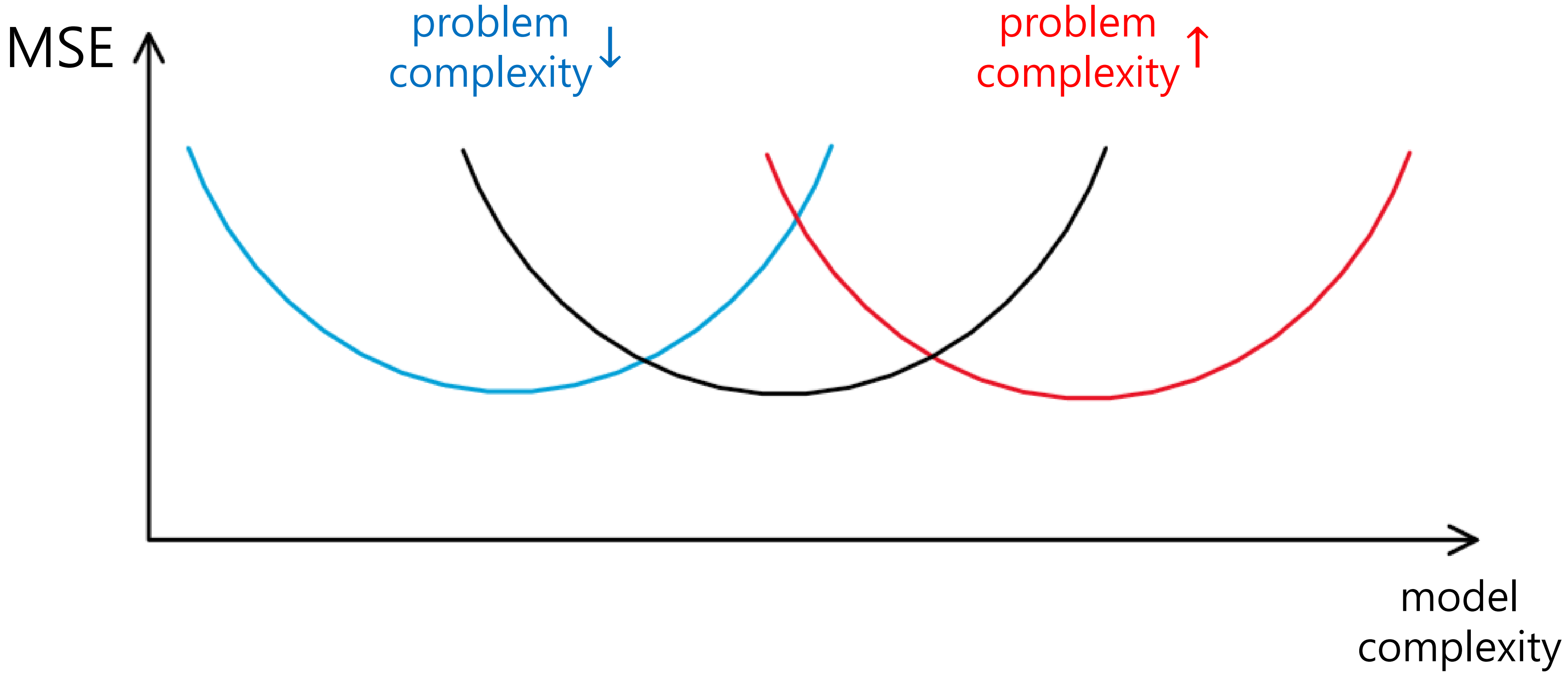
$$\left(E\left[\hat{\theta} - \theta\right]\right)^2 = \left(E\left[\hat{\theta}\right] - E[\theta]\right)^2 = \left(E\left[\hat{\theta}\right] - \theta\right)^2 = \left(Bias\left(\hat{\theta}\right)\right)^2$$

$$\therefore\ \mathrm{Var}\left(\hat{\theta}\right) = \mathrm{MSE}\left(\hat{\theta}\right) - \left(Bias\left(\hat{\theta}\right)\right)^2$$

모델 복잡도 (Model complexity)

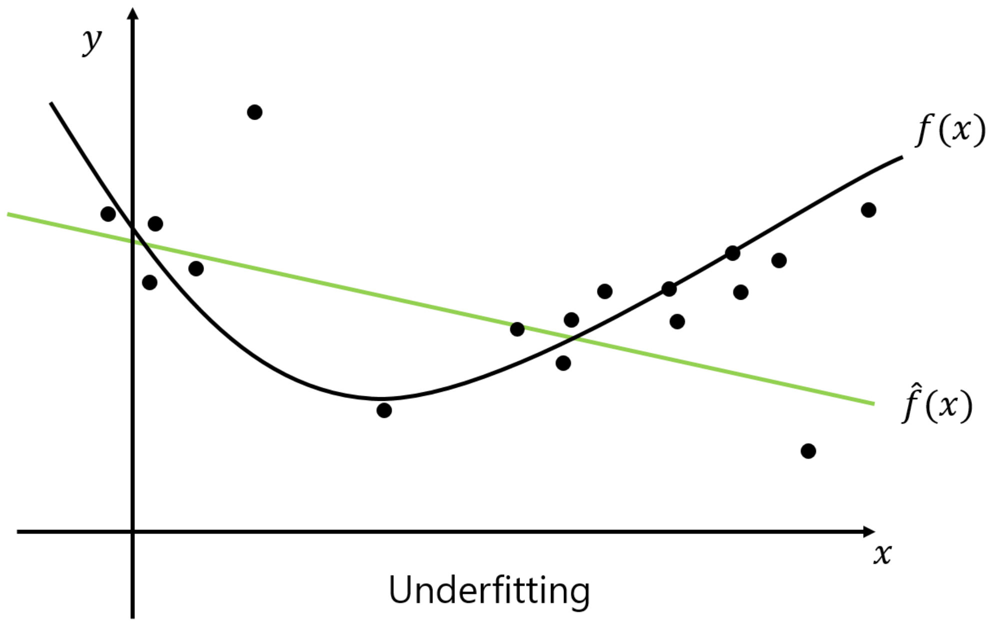


문제 복잡도 (Problem complextity)



문제의 복잡도에 맞는 복잡도를 가지는 모델을 선택해야한다!

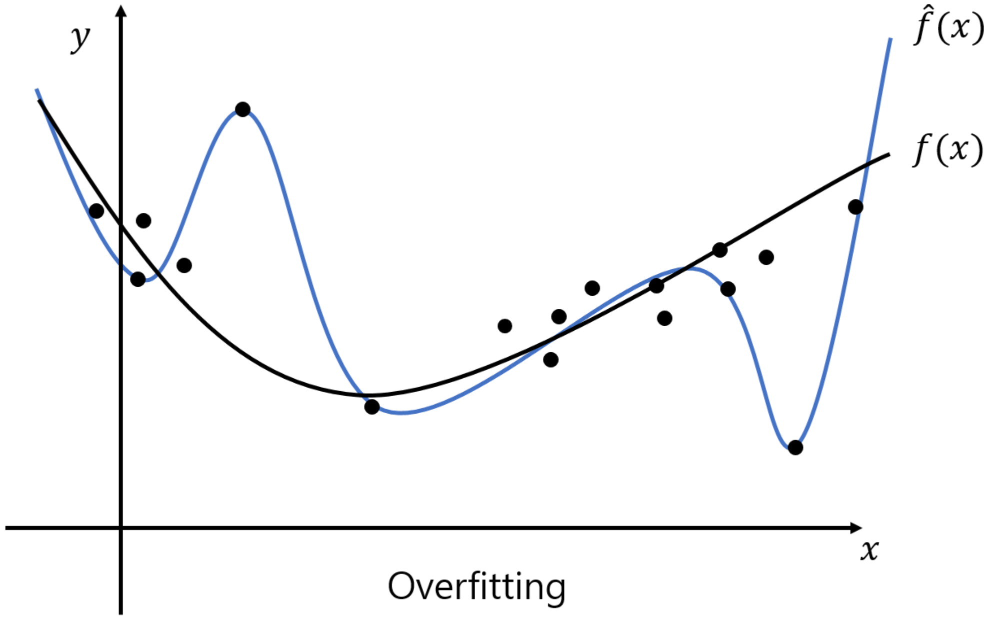
과소적합 (Underfitting or high bias)



Underfitting

- 선택한 모델이 강한 선입견(또는 편견)을 가지고 있음
- 데이터로부터 얻는 정보보다 선입견(편견)에 지배됨
- 이로 인해 학습의 결과가 나빠지는 상황
- 문제에 비해 모델의 복잡도가 부족할 때 나타남

과적합 (Overfitting or high variance)



Overfitting

- 문제에 비해 모델의 복잡도가 클 때
- 데이터의 양이 충분하지 않을 때
- train error에 비해 test error가 비정상적으로 높은 상태
- 기출문제는 100점 받는데 수능은 60점인 상황

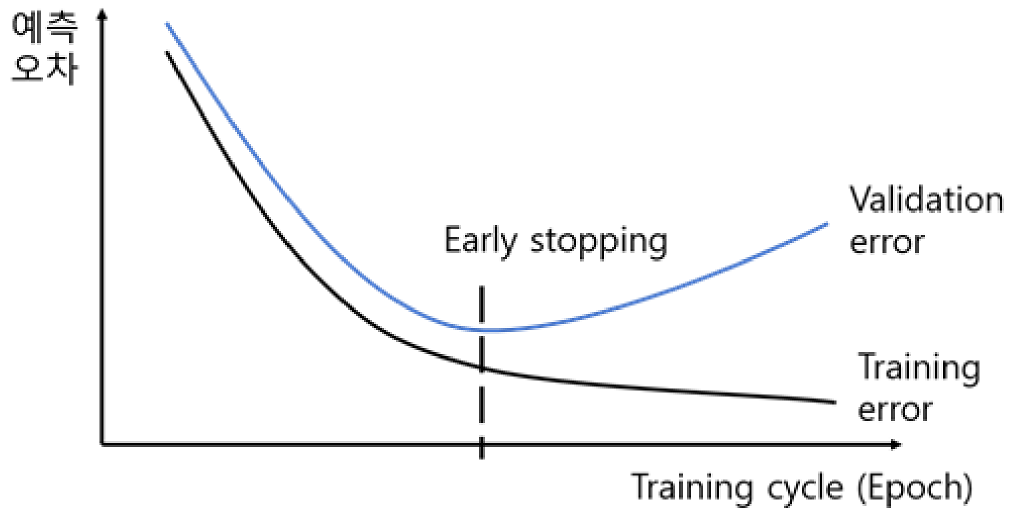
과적합 (Overfitting or high variance)

- 과적합 문제를 해결하는 방법

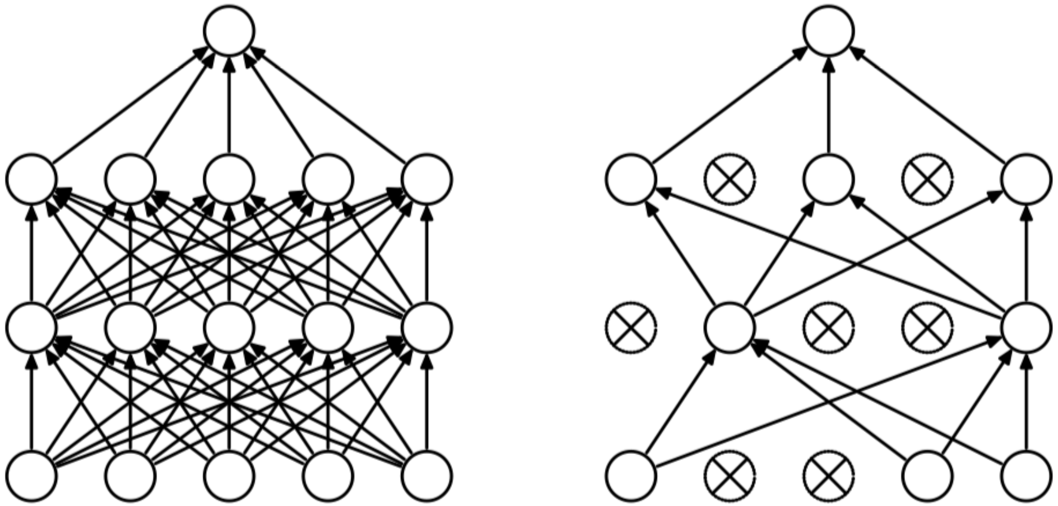
1. 더 많은 데이터를 사용한다.
   과적합 문제가 발생하는 근본적인 원인은 데이터의 부족이다.
2. 학습을 조기에 종료한다.
   학습을 조기에 종료함으로써 모델이 train 데이터에 완전히 맞춰지지 않게 한다.
3. 모델의 복잡도를 줄인다.
   모델의 표현력을 줄여 모델이 train 데이터에 완전히 맞춰지지 않게 한다.
4. 정규화 기법을 사용한다.
   정규화 기법은 train 데이터를 완전히 믿지 않음으로써 과적합문제를 완화한다.

과적합 (Overfitting or high variance)

- 학습을 조기에 종료한다.



- 모델의 복잡도를 줄인다.

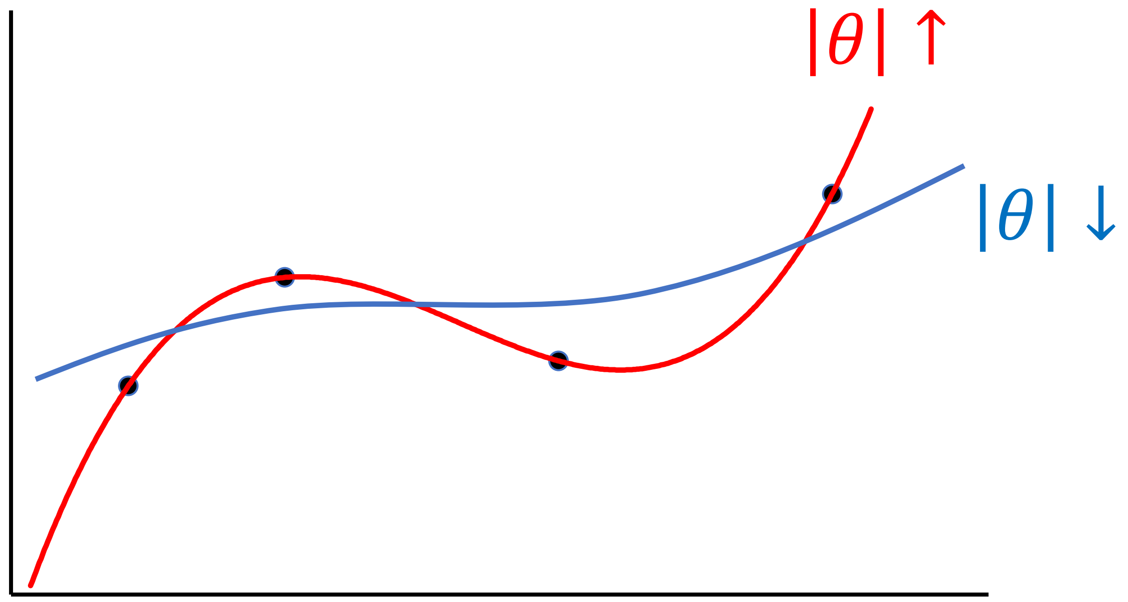


(a) Standard Neural Net

(b) After applying dropout.

정규화 (Regularization)

- 직관 : 파라미터 값이 작으면…
  1. 가설 함수 $h_\theta(x)$가 단순해진다.
  2. 과적합에 덜 취약해진다.



정규화 (Regularization)

- 문제 정의 : 제한적 최적화 문제 (constrained optimization problem)

  제한된 θ의 범위 내에서 오차 함수 $E(\theta)$를 극소화시키는 θ구하기

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \boxed{E(\theta)}$$: 목적 함수 (Objective function)

$$s.t. \boxed{\|\bar{\theta}\|_p^p \leq c^p}$$: 제한 조건 (Constraint)

$$where\ \theta = \begin{bmatrix} \theta_0 \\ \bar{\theta} \end{bmatrix}, c : constant$$

$$cf1)\ \|\bar{\theta}\|_p^p = \left(\sum\left(|\bar{\theta}_n|^p\right)^{\frac{1}{p}}\right)^p = \sum|\bar{\theta}_n|^p \qquad cf2)\ \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_N \end{bmatrix} \in R^{(N+1)\times 1}, \qquad \bar{\theta} = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_N \end{bmatrix} \in R^{N\times 1}$$

정규화 (Regularization)

- 문제 재정의

$$\boxed{\begin{aligned} \theta^* &= \underset{\theta}{\operatorname{argmin}}\, E(\theta) \\ s.t.\ & \|\bar{\theta}\|_p^p \le c^p \end{aligned}}$$

제한적 최적화 문제

$\downarrow$ Lagrangian method

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \left(E(\theta) + \lambda \left(\|\bar{\theta}\|_p^p - c^p \right) \right)$$

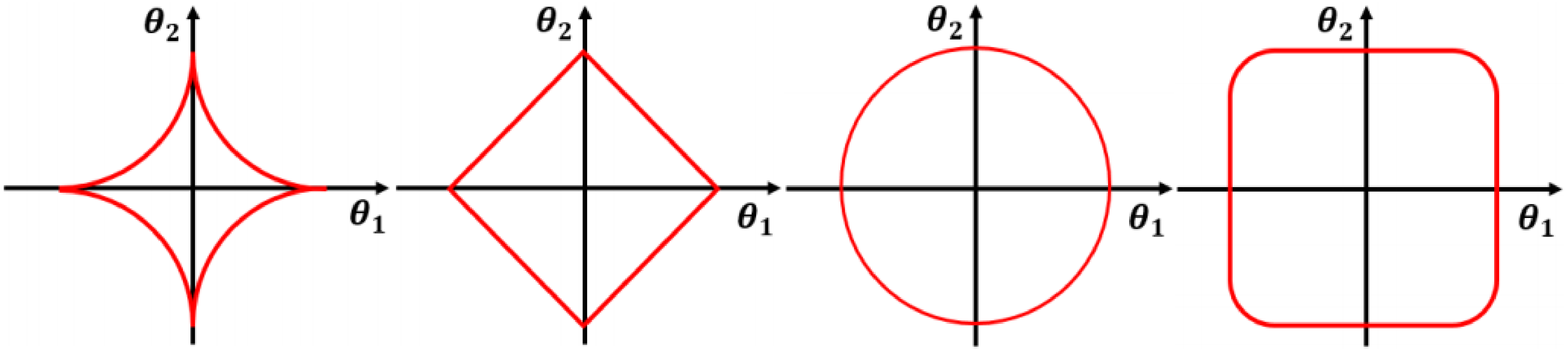
제한조건이 없는 최적화 문제

$$\boxed{\theta^* = \underset{\theta}{\operatorname{argmin}} \left(E(\theta) + \lambda \|\bar{\theta}\|_p^p \right)}$$

$\lambda : Lagrangian\ multiplier$

정규화 (Regularization)

- p값에 따른 정규화 함수$\left(\|\bar{\theta}\|_p^p = c^p\right)$의 형태



(a) $p = 0.5$ (b) $p = 1$ (c) $p = 2$ (d) $p = 4$

Lasso Ridge

정규화 (Regularization)

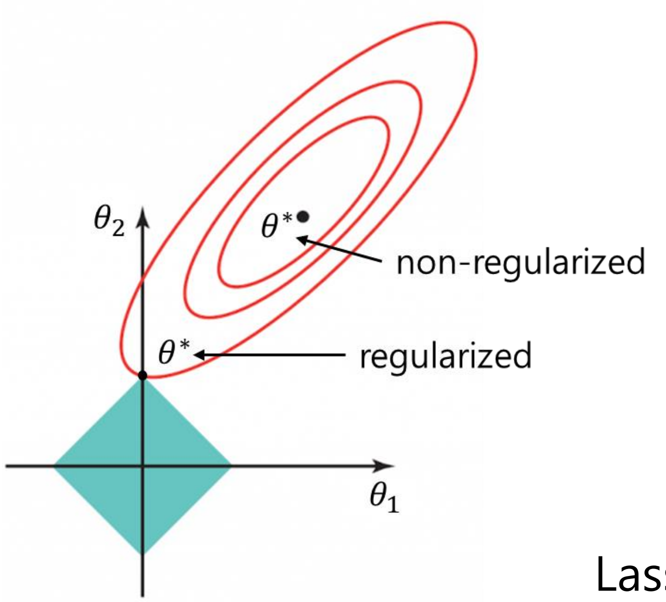
- Lasso 정규화 : $p = 1$일 때의 정규화 문제

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\left(E(\theta) + \lambda\|\bar{\theta}\|_1^1\right)$$

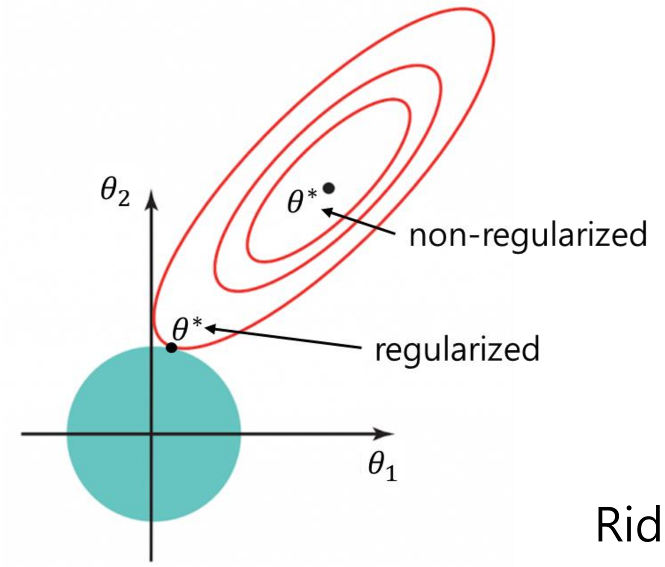
- Ridge 정규화 : $p = 2$일 때의 정규화 문제

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\left(E(\theta) + \lambda\|\bar{\theta}\|_2^2\right)$$

- 문제를 풀기위한 방법
  1. 경사 하강법 : Gradient descent
  2. Closed-form solution



Lasso : p=1



Ridge : p=2

Ridge 정규화

1. 경사 하강법

데이터로부터 학습한 정보를 덜 신뢰한다!

$$\theta_n \leftarrow \theta_n - \alpha \left(\sum_{i=1}^{I} \left(\theta^T x^{(i)} - y^{(i)} \right) x_n^{(i)} - \lambda \theta_n \right)$$

2. Closed-form solution

$$\theta^* = \left(x^T x + \lambda \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \right)^{-1} x^T y$$

$shape : \ [N+1, N+1]$

- 행렬의 크기
  - $\theta : [N+1, 1]$
  - $x : [I, N+1]$
  - $y : [I, 1]$

Lab 2 : Regularized Regression

- Tasks
  1. data_lab2.txt의 데이터를 읽고, train 데이터 70%, test 데이터 30%로 분할하여라.
  2. 다음의 가설 함수들에 대해 파라미터의 최적값을 구하여라.

     (비정규화 회귀는 최소 제곱법, 정규화 회귀는 closed-form solution 사용)

     a. 비정규화 회귀 – 선형 모델
     b. 비정규화 회귀 – 2차 다항식 모델
     c. 비정규화 회귀 – 5차 다항식 모델
     d. 정규화 회귀 – 5차 다항식 모델 (Ridge)
  3. train 데이터 위에 학습된 회귀자들을 plot하고, train error를 구하여라.
  4. test 데이터 위에 학습된 회귀자들을 plot하고, test error를 구하여라.