

CUAI 스터디_MTs

2022.04.04

발표자 : 정 승 욱

스터디 모임 인증

Ch 04 모델 훈련 회귀 Regression

1. 선형 회귀
2. 경사 하강법
3. 다항 회귀
4. 학습 곡선
5. 규제 : 릿지, 라쏘, 엘라스틱 넷
6. 로지스틱 회귀

정규 방정식, 계산복잡도

장점: 한번에 최적 계수가 나옴! But 비용 높음!

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

#1. (a)

$$h_{\theta}(x^{(i)}) = \theta_0 + \theta_1 x_1^{(i)} + \dots + \theta_d x_d^{(i)}$$

$$= (X^{(i)})^T \theta$$

$(1 \times d+1) \quad (d+1) \times 1$

예측값

$$X\theta = \begin{pmatrix} (X^{(1)})^T \theta \\ (X^{(2)})^T \theta \\ \vdots \\ (X^{(n)})^T \theta \end{pmatrix} = \begin{pmatrix} h_{\theta}(x^{(1)}) \\ h_{\theta}(x^{(2)}) \\ \vdots \\ h_{\theta}(x^{(n)}) \end{pmatrix}$$

$(n \times d+1) \quad (d+1) \times 1$

$X\theta$: 예측값, Predicted value.

y : 실제값, Actual value (observed)

$$y = \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{pmatrix}$$

$n \times 1$

$$X\theta - y = \begin{pmatrix} h_{\theta}(x^{(1)}) - y^{(1)} \\ h_{\theta}(x^{(2)}) - y^{(2)} \\ \vdots \\ h_{\theta}(x^{(n)}) - y^{(n)} \end{pmatrix}$$

예측값 - 실제값

예측값 - 실제값의 Least Square Estimation
을 구해야 한다.

$$J(\theta) = \frac{1}{2} \sum_{i=1}^n [y^{(i)} - h_{\theta}(x^{(i)})]^2$$

NOTE That $Z^T Z = \sum_i Z_i^2$
Then Z is column vector.

$$J(\theta) = \frac{1}{2} (y - X\theta)^T (y - X\theta)$$

$$= \frac{1}{2} (y^T - \theta^T X^T) (y - X\theta)$$

Here,

$$(y^T - \theta^T X^T) (y - X\theta), \text{ --- ①}$$

$$= y^T y - y^T X\theta - \theta^T X^T y + \theta^T X^T X\theta$$

$$y^T X\theta = (\theta^T X^T y)^T = \theta^T X^T y$$

$\therefore 1 \times 1$ transposes.

$$\therefore y^T: 1 \times n \quad X: n \times (d+1) \quad \theta: (d+1) \times 1$$

Then $y^T X\theta: 1 \times 1$ matrix

$$\theta^T: 1 \times (d+1) \quad X^T: (d+1) \times n \quad y: n \times 1$$

Then $\theta^T X^T y: 1 \times 1$ matrix.

So that, ①

$$y^T y - 2y^T X\theta + \theta^T X^T X\theta$$

$\underbrace{\theta^T X^T X\theta}_{\text{is constant.}}$

$$J(\theta) = \frac{1}{2} [\theta^T X^T X\theta - 2y^T X\theta + y^T y]$$

Continue.

최적화 되는 θ , $\nabla_{\theta} J(\theta) = 0$ 을
만족하는 θ .

Here, $\nabla_x b^T x = b$. $\nabla_x (x^T A x) = 2Ax$
 A : symmetric.

$$\nabla_{\theta} J(\theta) \text{ (0이 되는 } \theta \text{)}$$

$$= \frac{1}{2} \nabla_{\theta} [\theta^T X^T X\theta - 2y^T X\theta]$$

$X^T X: (d+1) \times 1 \times (d+1) = (d+1) \times (d+1)$ symmetric.
여기서 θ 도 $(d+1) \times 1$ 이므로 곱할 수 있음.

$$\nabla_{\theta} (\theta^T X^T X\theta) = 2(X^T X)\theta$$

$$= 2X^T X\theta$$

$$\nabla_{\theta} (-2y^T X\theta) = \nabla_{\theta} [-2(X^T y)^T \theta]$$

$$= -2X^T y$$

$$\Rightarrow \frac{1}{2} (2X^T X\theta - 2X^T y)$$

$$= X^T X\theta - X^T y$$

$$\nabla_{\theta} J(\theta) = 0$$

$$X^T X\theta - X^T y = 0$$

$$X^T X\theta = X^T y$$

$$\underbrace{(X^T X)^{-1}}_I (X^T X)\theta = (X^T X)^{-1} X^T y$$

$$\therefore \theta = (X^T X)^{-1} X^T y$$

$\nabla_x b^T x$
 $(b_1 \dots b_n) \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$

$\frac{\partial}{\partial x} (b_1 x_1 + \dots + b_n x_n)$

$\Rightarrow \frac{\partial}{\partial x_1} \quad \frac{\partial}{\partial x_2} \quad \dots \quad \frac{\partial}{\partial x_n}$

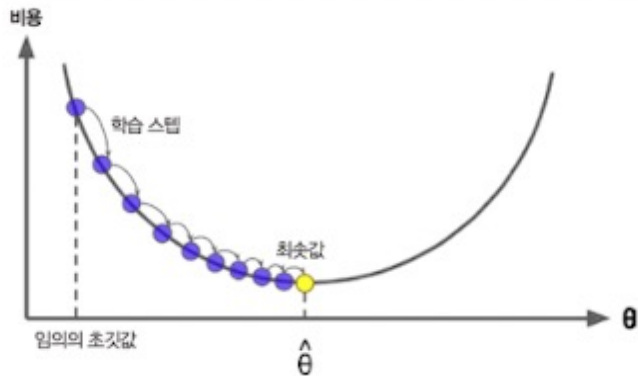
$\Rightarrow (b_1 \quad b_2 \quad \dots \quad b_n)$

$= \underline{\underline{b}}$

$\nabla_x (x^T A x)$
 $= (x_1 \dots x_n) \begin{pmatrix} A_{11} & A_{12} & \dots & A_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \dots & A_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$

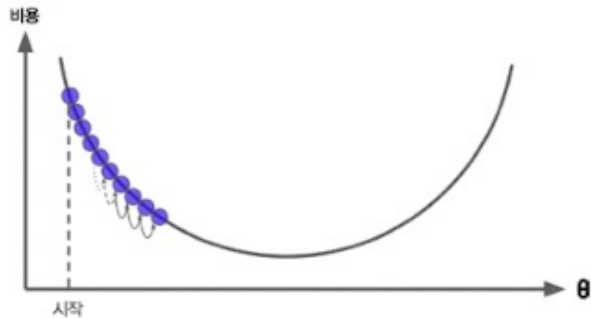
$= A_{11}x_1^2 + A_{12}x_1x_2 + \dots + A_{1n}x_1x_n$
 $+ 2A_{21}x_1x_2 + 2A_{22}x_2^2 + \dots + 2A_{2n}x_2x_n$
 \vdots

경사 하강법



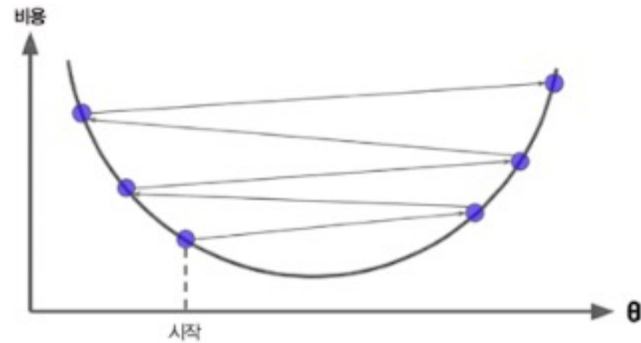
$$\theta^{(nextstep)} = \theta - \eta \nabla_{\theta} MSE(\theta)$$

$\eta = \text{Learning Rate}$



학습률 너무 작을 때

:



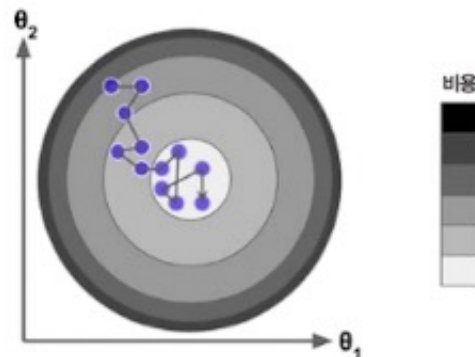
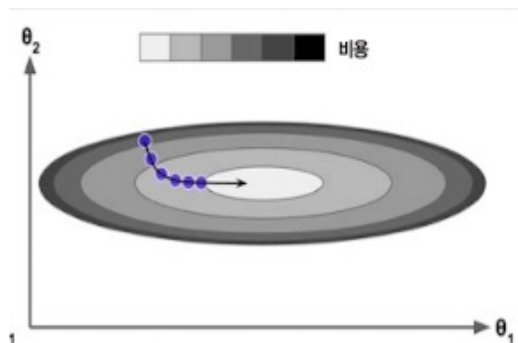
학습률 너무 클 때

배치 경사 하강법 & 확률적 경사 하강법 & 미니 배치 경사 하강법

배치 : 매 스텝 **모든** 샘플

확률 : 매 스텝 **1개** 샘플

미니 배치 : 매 스텝 **작은 랜덤** 샘플



<https://www.youtube.com/watch?v=sDv4f4s2SB8>
<https://www.youtube.com/watch?v=vMh0zPT0tLI>

Epoch & Batch & Iteration



1 Epoch : 모든 데이터 셋을 한 번 학습

1 iteration : 1회 학습

minibatch : 데이터 셋을 batch size 크기로 쪼개서 학습

ex) 총 데이터가 100개, batch size가 10이면,

1 iteration = 10개 데이터에 대해서 학습

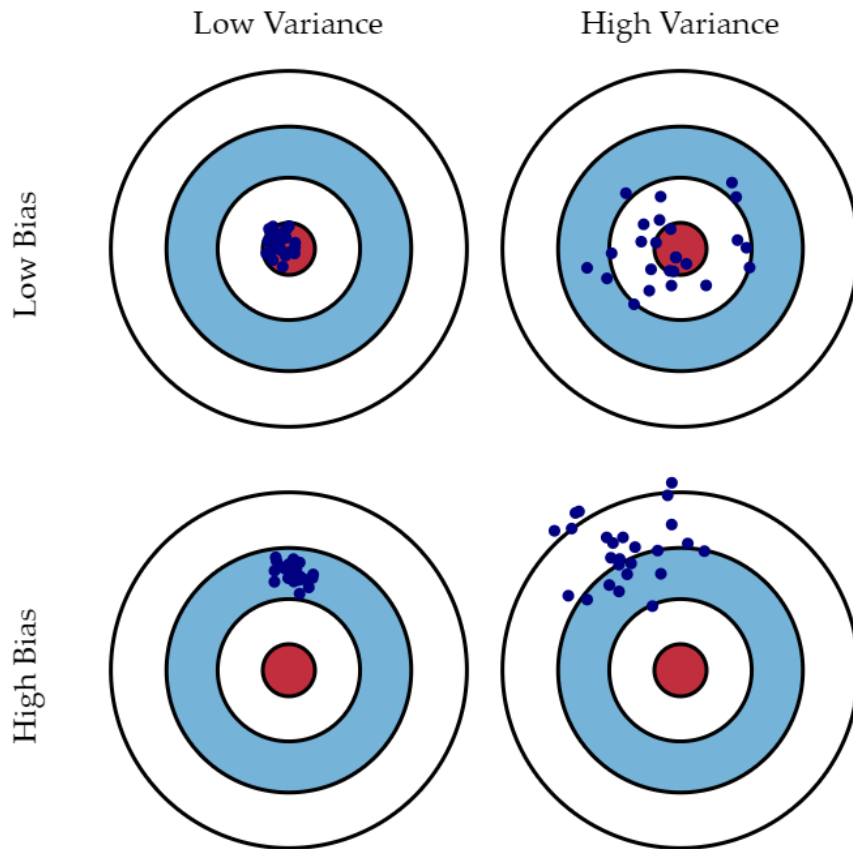
1 Epoch = $100 / \text{batch size} = 10$ iteration

분산과 편향

편향: 잘못된 가정
분산: 작은 변동에 모델이 민감

데이터 안의 진동: 엡실론으로 표기!

편향 종류!! bias vs variance & ϵ 인지!



규제 : 릿지, 라쏘, 엘라스틱넷

-> 가중치에 제한, 과대적합 방지

$$J(\boldsymbol{\theta}) = \text{MSE}(\boldsymbol{\theta}) + \alpha \frac{1}{2} \sum_{i=1}^n \theta_i^2$$

릿지: 덜 중요한 특성의 가중치 낮춤 | 2 규제!

$$J(\boldsymbol{\theta}) = \text{MSE}(\boldsymbol{\theta}) + \alpha \sum_{i=1}^n |\theta_i|$$

라쏘: 덜 중요한 특성의 가중치 삭제 | 1 규제!

$$J(\boldsymbol{\theta}) = \text{MSE}(\boldsymbol{\theta}) + r\alpha \sum_{i=1}^n |\theta_i| + \frac{1-r}{2}\alpha \sum_{i=1}^n \theta_i^2$$

릿지와 라쏘의 절충: 엘라스틱 넷

혼합 비율 $r, 1-r$

- 특성 수가 훈련 샘플 수 보다 많거나
- 특성 몇 개가 강하게 연관 되어있을 때 라쏘가 문제를 일으킴

로지스틱 회귀: 범주형 자료 회귀, 분류 모델

이진 분류 하는 문제! 1 또는 0

시그모이드 함수를 이용

$$\log(\text{Odds}(p)) = wx + b$$

오즈 Odds

실패 비율 대비 성공 비율

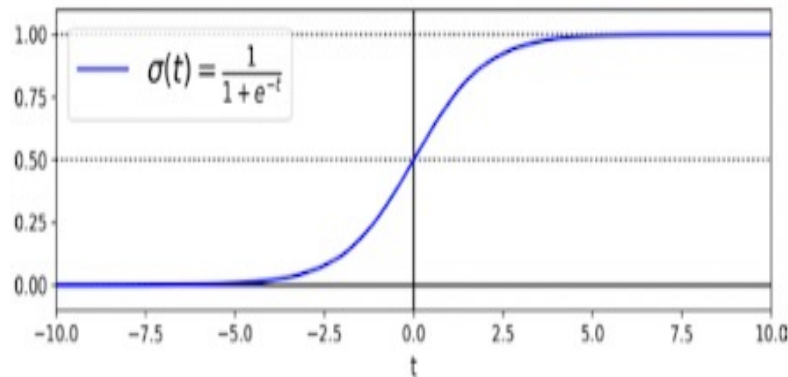
$$\frac{p}{1-p}$$

로짓 Logit Log+Odds

오즈에 자연로그!

$$L = \ln \frac{p}{1-p}$$

$$p = \frac{e^{-L}}{e^{-L} + 1}$$



<https://www.youtube.com/watch?v=yIYKR4sgzI8>

감사합니다!