model complexity ~ overfitting

↑

regularization.

feature,

Inclass 19: Polynomial Regression and Regularization 정규화

[SCS4049] Machine Learning and Data Science

Seongsik Park (s.park@dgu.edu)

School of AI Convergence & Department of Artificial Intelligence, Dongguk University

$$\underset{x}{\text{input}} \rightarrow \boxed{\text{feature 특징, 특성.}} \xrightarrow{\overset{z}{\text{feature}}} \boxed{\text{model}} \rightarrow \overset{y}{\text{output.}}$$

regression

classification

clustering

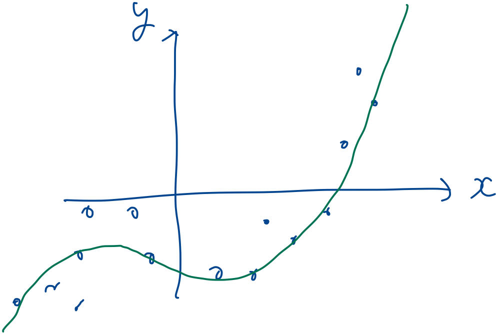
⋮

1D input, 1D output,
+bias

다항, polynomial

$$y = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$$

$$(x^{(1)}, y^{(1)}), \; (x^{(2)}, y^{(2)}) \; \dots \; (x^{(100)}, y^{(100)})$$



min SSE(θ)

→ normal

→ GD

$$\mathbb{X} = \begin{bmatrix} 1 & x^{(1)} \\ 1 & x^{(2)} \\ \vdots & \vdots \\ 1 & x^{(100)} \end{bmatrix} \qquad \mathbb{Y} = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(100)} \end{bmatrix}$$

normal. $\hat{\Theta} = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbb{Y}$

BGD. $\Theta \leftarrow \Theta - \eta \, 2\mathbb{X}^T(\mathbb{X}\Theta - \mathbb{Y})$

$$X = \begin{bmatrix} 1 & x^{(1)} & x^{2(1)} & x^{3(1)} \\ 1 & x^{(2)} & x^{2(2)} & x^{3(2)} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x^{(100)} & x^{2(100)} & x^{3(100)} \end{bmatrix} \in \mathbb{R}^{100\times 4}$$

$$\hat{y} = X\theta = \begin{bmatrix} | & | & | & | \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix}$$

$$= \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}\theta_0 + \begin{bmatrix} x^{(1)} \\ \vdots \\ x^{(100)} \end{bmatrix}\theta_1 + \begin{bmatrix} x^{2(1)} \\ x^{2(2)} \\ \vdots \\ x^{2(100)} \end{bmatrix}\theta_2 + \begin{bmatrix} x^{3(1)} \\ \vdots \\ x^{3(100)} \end{bmatrix}\theta_3$$

$$\underset{4\times100}{} \; \underset{100\times4}{} \; \underset{4\times100}{} \; \underset{100\times1}{}$$

$$\hat{\theta} = (X^T X)^{-1} X^T y \rightarrow 4\times1$$

$$\theta \leftarrow \theta - 2\eta X^T(X\theta - y) \qquad 4\times1$$

$$y = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$$

$$= \theta_0 + \theta_1 z_1 + \theta_2 z_2 + \theta_3 z_3$$

$$x \rightarrow \boxed{\text{feature}} \rightarrow z_1, z_2, z_3$$

x	y
1	2
2	4
3	5
4	7
5	9

1D $\rightarrow$ 3D poly + bias $\rightarrow$ 1D

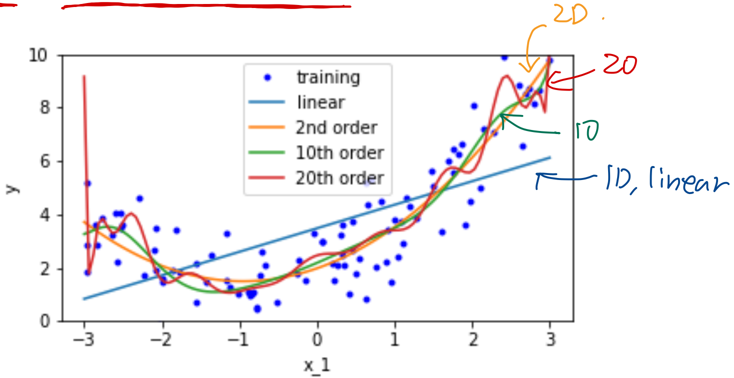
$$y = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$$

$$X = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 \\ 1 & 3 & 9 & 27 \\ 1 & 4 & 16 & 64 \\ 1 & 5 & 25 & 125 \end{bmatrix} \qquad y = \begin{bmatrix} 2 \\ 4 \\ 5 \\ 7 \\ 9 \end{bmatrix}$$

$$x \rightarrow \boxed{\text{feature}} \rightarrow x, x^2, x^3 \rightarrow \boxed{\text{model}} \rightarrow y$$

?

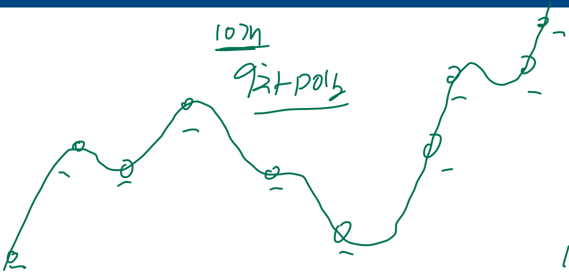
모델의 차수는 어떻게 결정해야 하는가?



poly ↑ , SSE ↓

10개

9차 poly



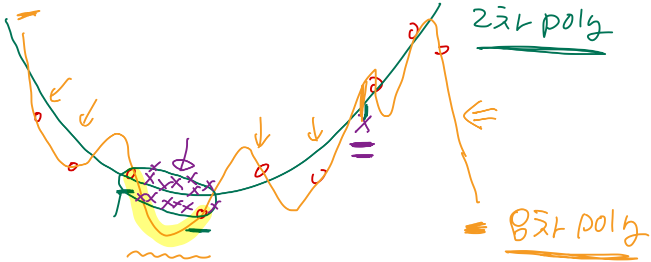
10개
data

과적합 overfitting

- 학습 데이터에 대해서는 좋은 성능을 보이지만 모델 복잡도↑,
- 처음 보는 데이터에 대해서는 일반화하지 못함
- 학습 데이터의 양이나 노이즈에 비해 모델이 너무 복잡할 때
- 모델이 학습 데이터를 일반화하는 범위를 넘어서 과도하게 의존함
- 모델이 학습 데이터를 기억하는듯한 양상을 보이기 시작

미적합 underfitting

- 모델이 너무 단순하여 학습 데이터에 대해서도 부정확

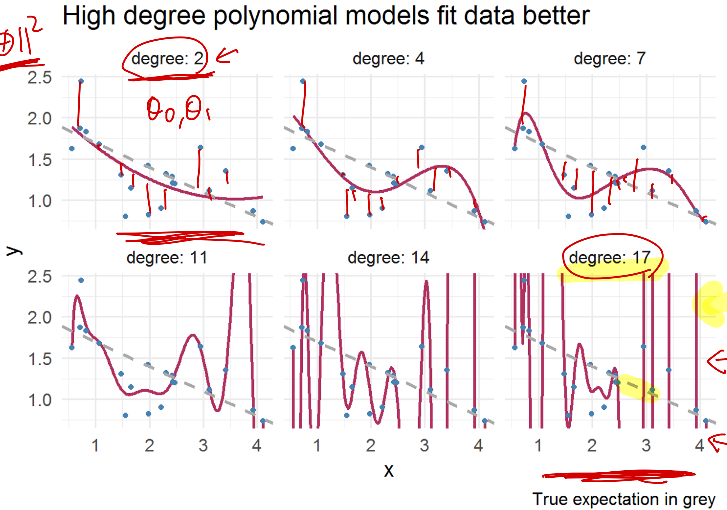


학습데이터. SSE > SSE

새데이터. error < error.

Overfitting and underfitting

$\lambda \|\theta\|^2$

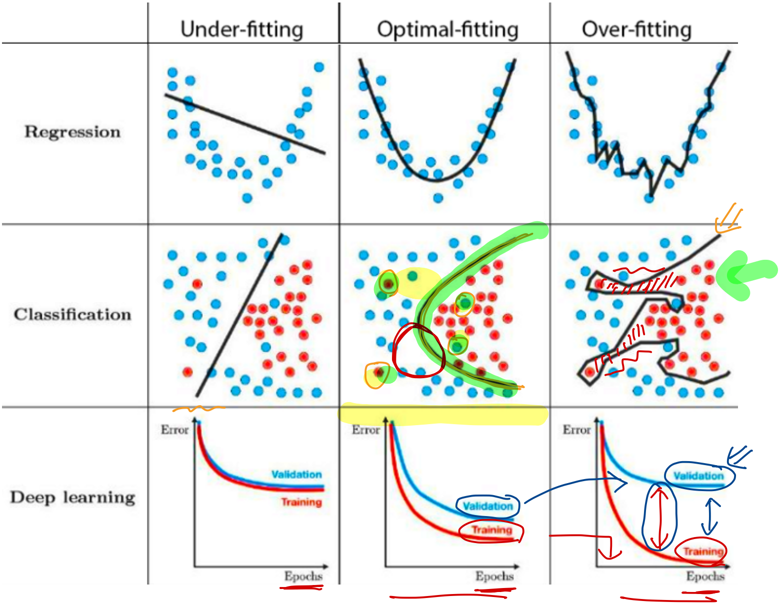


$\|\theta\|^2$

θ_0
$\vdots$
θ_{17}

$SSE(\theta)\ \lambda$

Overfitting and underfitting



정규화 regularization

min · 비용 함수에 복잡도에 대한 penalty를 추가

· 과적합에 대한 비용을 optimizing cost에 반영

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_D \end{bmatrix}$$

Ridge regression

$\|\theta\|$

$$\min \quad J(\boldsymbol{\theta}) = \mathrm{SSE}(\boldsymbol{\theta}) + \frac{\alpha}{2}\|\boldsymbol{\theta}\|^2 \quad \leftarrow \quad \text{2norm. } (L2) \tag{1}$$

$$\mathrm{SSE}(\theta) + 100\|\theta\|^2$$

LASSO

$$J(\boldsymbol{\theta}) = \mathrm{SSE}(\boldsymbol{\theta}) + \alpha \sum_i |\theta_i| \quad \leftarrow \quad \text{1 norm } (L1) \tag{2}$$

$\theta \leftarrow \theta - \eta(2X(X\theta - y) + \alpha\theta)$

Ridge regression

$$J(\boldsymbol{\theta}) = \mathrm{SSE}(\boldsymbol{\theta}) + \frac{\alpha}{2}\|\boldsymbol{\theta}\|^2 \quad (3)$$

- Hyperparameter α로 두 항목 간의 상대적인 비중을 조절
- Closed form solution $\hat{\boldsymbol{\theta}} = (\mathbf{X}^T\mathbf{X} + \alpha\mathbf{I}_n)^{-1}\mathbf{X}^T\mathbf{y}$

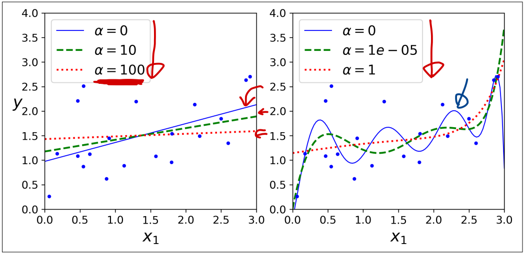


Figure 4-17. Ridge Regression

LASSO

$$J(\boldsymbol{\theta}) = \mathrm{SSE}(\boldsymbol{\theta}) + \alpha \sum_{i=1}^{n} |\theta_i| \tag{4}$$

- Least Absolute Shrinkage and Selection Operator Regression
- 중요하지 않은 feature들의 weight를 0으로 만드는 경향

(feature selection)

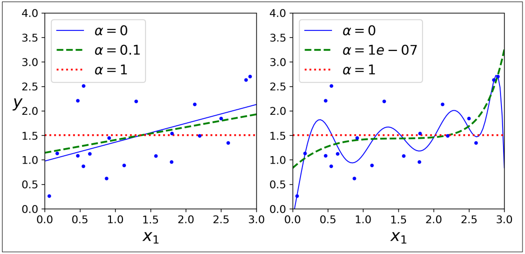


Figure 4-18. Lasso Regression

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3$$

0

$$\hat{\theta} = [\ 1 \quad 0.5 \quad 0 \quad 0.3\ ]$$

x_2는 중요하지 않다.