

'overfitting'

selection

Inclass 13: Model Evaluation and Regularization

[SCS4049] Machine Learning and Data Science

Seongsik Park (s.park@dgu.ac.kr)

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

of cluster $\frac{2}{2}$ 많아서.



↓ 결정.

Clustering Model Evaluation

k-means : { inertia
silhouette score / coeff.

K-means clustering: inertia

A performance metric, called *inertia*

inertia = the mean squared distance
between each instance and its closest centroid $\leftarrow$
 $x_1, x_2, \dots, x_N$ $\mu_1, \mu_2, \dots, \mu_K$

= k-means clustering
minimizing cost.

Elbow rule: any lower value would be dramatic, while any higher value would not help much.

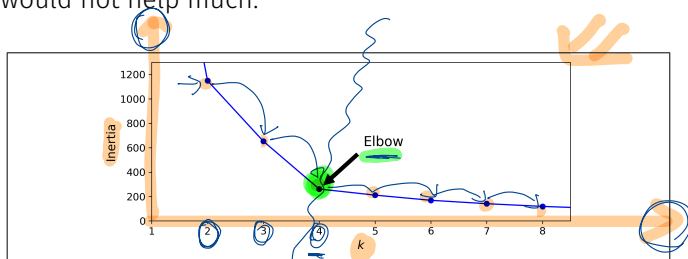
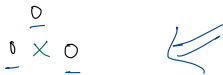
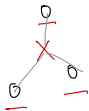


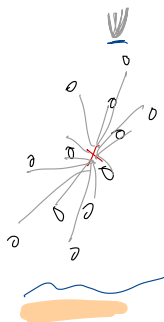
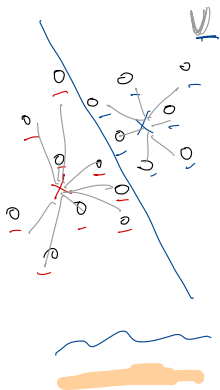
Figure 9-8. Selecting the number of clusters k using the “elbow rule”



$$D_{nk} = \|x_n - \mu_k\|_2$$

$$D = \begin{bmatrix} 1 & 2 & 4 \\ 1 & 3 & 5 \\ 4 & 1 & 6 \\ 4 & 3 & 1 \\ 5 & 3 & 1 \end{bmatrix}$$

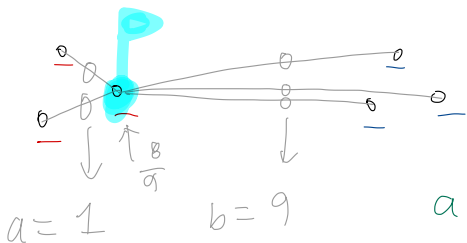
$$\text{inertia} = \frac{1}{5} [1^2 + 1^2 + 1^2 + 1^2 + 1^2]_{0/1}$$



K-means clustering: silhouette score

A more precise approach (but also more computationally expensive) is to use the *silhouette score*, which is the mean *silhouette coefficient* over all the instances. $\rightarrow \text{per sample}$

- An instance's silhouette coefficient is equal to $(b - a) / \max(a, b)$
- where a is the mean distance to the other instances in the same cluster (i.e., the mean intra-cluster distance)
- and b is the mean nearest-cluster distance (i.e., the mean distance to the instances of the next closest cluster, defined as the one that minimizes b , excluding the instance's own cluster).



a ① intra-cluster distance

b ② nearest-cluster distance.

$$\text{coeff} = \frac{b-a}{\max(a,b)}$$

$$\frac{9-1}{9}$$

$$= \frac{8}{9}$$

0
0
0

0
0
0

$a \downarrow _ b \uparrow$

$b \gg a$

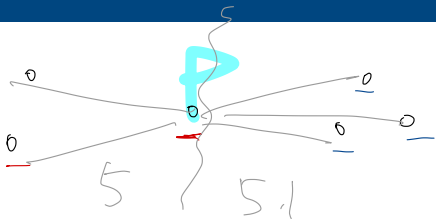
a ① intra-cluster distance

b ② nearest-cluster distance

$$\text{coeff} = \frac{b - a^k}{\max(a, b)}$$

0
0
0

$$\approx \frac{b}{b} = \underline{\underline{+1}}$$



$$a \approx b$$

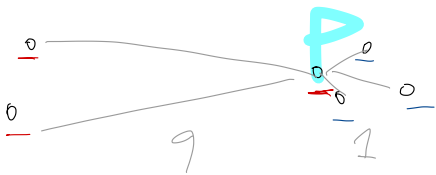
a ① intra-cluster distance

b ② nearest-cluster distance

$$\text{coeff} = \frac{b - a^k}{\max(a, b)}$$

$$\frac{0}{0} = 0$$

$$= \frac{0}{a \text{ or } b} = 0$$



$a \gg b$

a ① intra-cluster distance

b ② nearest-cluster distance

$$\text{coeff} = \frac{b - a^k}{\max(a, b)}$$

$$\frac{0}{0}$$

$$= \frac{-a}{a} = -1$$

K-means clustering: silhouette score

The silhouette coefficient can vary between -1 and +1.

- A coefficient close to +1 means that the instance is well inside its own cluster and far from other clusters,
- while a coefficient close to 0 means that it is close to a cluster boundary,
- and finally a coefficient close to -1 means that the instance may have been assigned to the wrong cluster.

K-means clustering: silhouette score

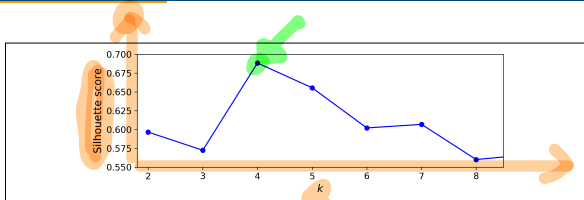


Figure 9-9. Selecting the number of clusters k using the silhouette score coefficient. This is called a *silhouette diagram* (see Figure 9-10):

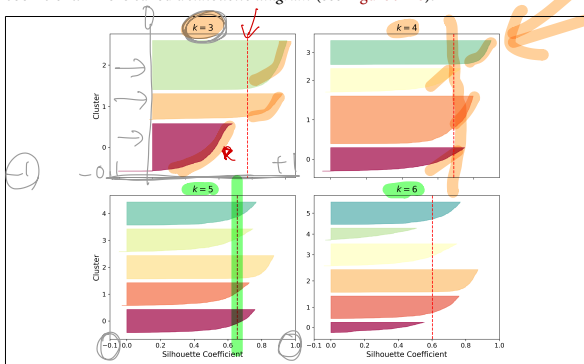
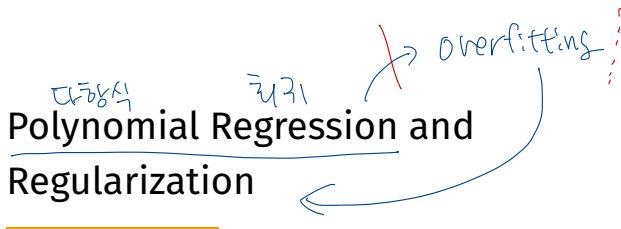
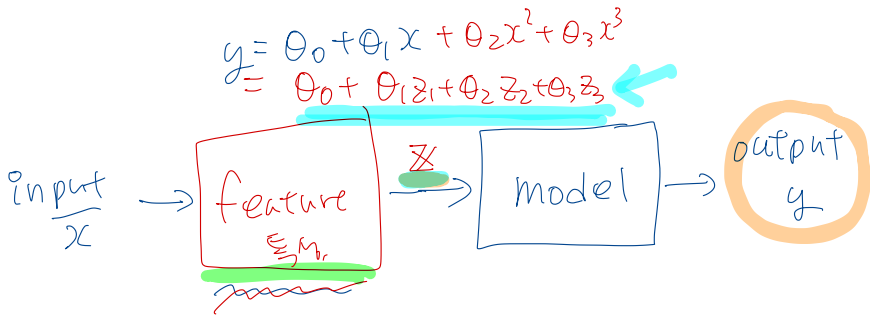


Figure 9-10. Silhouette analysis: comparing the silhouette diagrams for various values of k

다항식 차수 ~~→ overfitting~~ !

Polynomial Regression and Regularization





$$z_1 = x$$

$$z_2 = x^2$$

$$z_3 = x^3$$

x	x^2	y
1	1	4
2	4	2
3	9	1
4	16	2
5	25	4
z_1	z_2	

① normal eqn. ✓

$$X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{bmatrix} \quad y = \begin{bmatrix} 4 \\ 2 \\ 1 \\ 2 \\ 4 \end{bmatrix}$$

$$\hat{\theta} = \underbrace{(X^T X)^{-1}}_{3 \times 3} \underbrace{X^T y}_{3 \times 1}$$

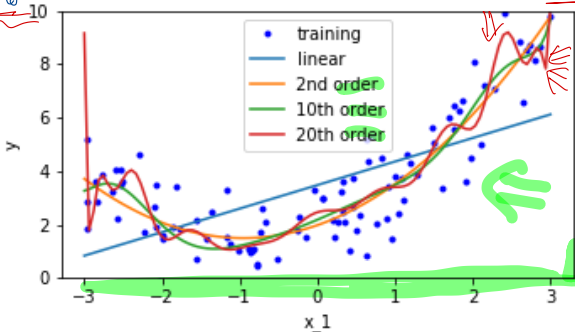
$$y = \theta_0 + \theta_1 x + \theta_2 x^2$$

② BGD. $\theta = \theta - 2\eta \sum_{n=1}^5 (\theta_0 + \theta_1 x + \theta_2 x^2 - y)$

Polynomial regression

모델의 차수는 어떻게 결정해야 하는가?

(모델의 복잡도)



정규성

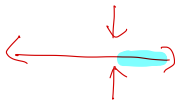


모델 복잡도

모델 복잡도

단순함

모델 복잡도 ↑



복잡함

이반정규성, 정규성, 노이즈 포함

Overfitting and underfitting

① 과적합 overfitting → regularization 정규화.

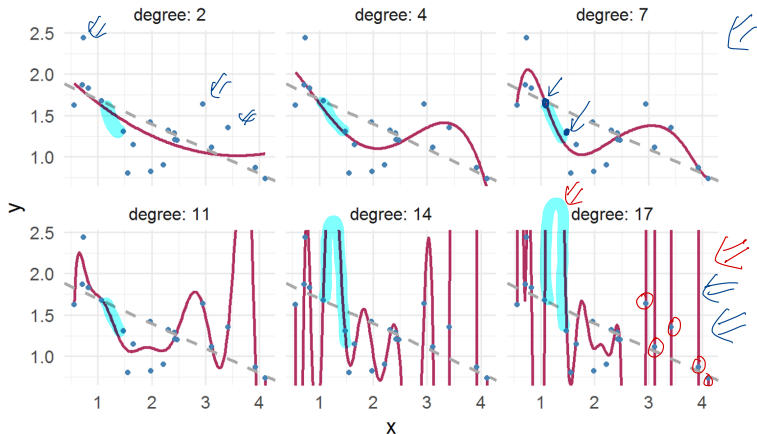
- 학습 데이터에 대해서는 좋은 성능을 보이지만 복잡성 ↑, 일반성 ↑.
- 처음 보는 데이터에 대해서는 일반화하지 못함
- 학습 데이터의 양이나 노이즈에 비해 모델이 너무 복잡할 때
- 모델이 학습 데이터를 일반화하는 범위를 넘어서 과도하게 의존함
- 모델이 학습 데이터를 기억하는 듯한 양상을 보이기 시작

미적합 underfitting

- 모델이 너무 단순하여 학습 데이터에 대해서도 부정확

Overfitting and underfitting

High degree polynomial models fit data better

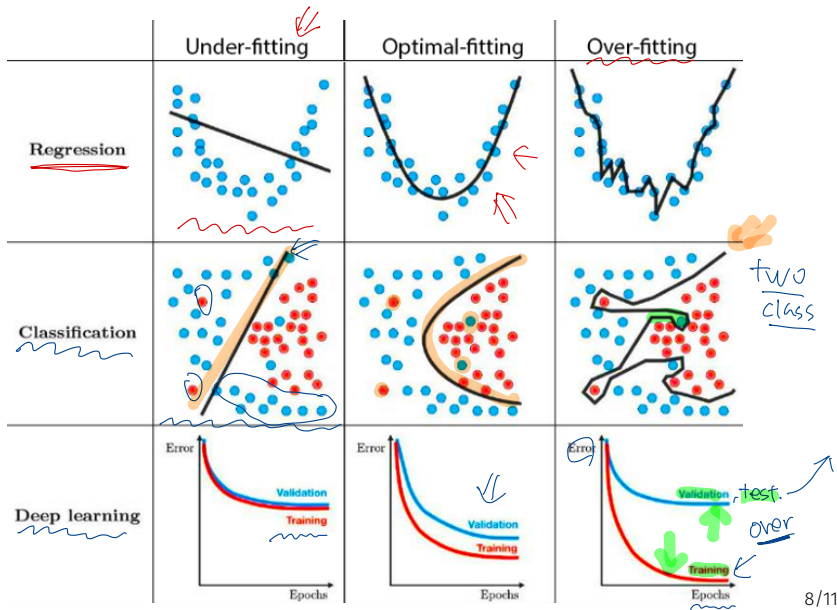


True expectation in grey

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

7/11

Overfitting and underfitting



Regularization

정규화 regularization

- 비용 함수에 복잡도에 대한 penalty를 추가
- 과적합에 대한 비용을 optimizing cost에 반영

$$\|\theta\|_1, \|\theta\|_2 \dots$$

복잡도.

Ridge regression

$$\min J(\theta) = \text{SSE}(\theta) + \frac{\alpha}{2} \|\theta\|^2$$

trade-off

$\alpha \uparrow \|\theta\|^2$ 증가, 과적합.
 $\alpha \downarrow \|\theta\|^2$ 감소, 정규화.
 $\alpha = 0$ SSE regularization.

LASSO

$$J(\theta) = \text{SSE}(\theta) + \alpha \sum_i |\theta_i|$$

$$\|\theta\|_1. \quad (2)$$

Ridge regression

Ridge regression

$$J(\theta) = \text{SSE}(\theta) + \frac{\alpha}{2} \|\theta\|^2 \quad \text{closed (normal) GD (approx).} \quad (3)$$

(Handwritten notes: The equation is boxed in red. An orange arrow points to the α term, and a blue arrow points to the $\frac{1}{2}$ term. A red arrow points to the $\|\theta\|^2$ term.)

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

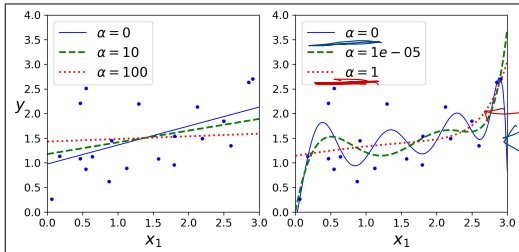


Figure 4-17. Ridge Regression

Lasso regression

LASSO

$$\hat{y} = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$$
$$\hat{\theta} = [1, 2, 0, 4]^T \quad (4)$$

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

← feature selection.

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

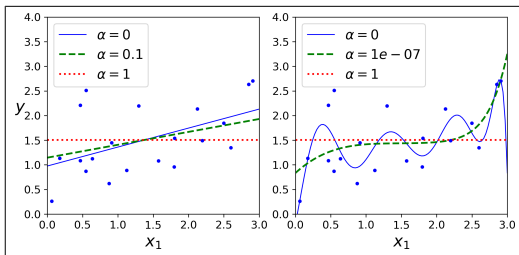


Figure 4-18. Lasso Regression

~~closed~~
approx.