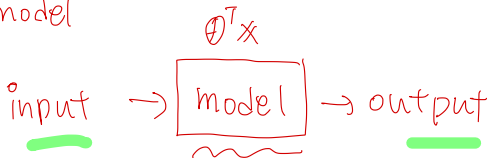


기본적인 model



Inclass 13: Model Evaluation and Regularization

[SCS4049] Machine Learning and Data Science

정규화

과적합.

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

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

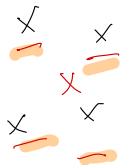
k-means : (#) of clusters $\sum_{j=1}^k D_{ij} < 1$



Clustering Model Evaluation

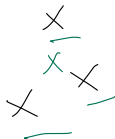
{ inertia

X12011
271



$$R_{nk} = \begin{cases} 1 \\ 0 \end{cases} \quad k = \arg\min D_{nk}$$


$$\mu_k = \frac{1}{N_k} \sum R_{nk} x_n$$

min $\frac{1}{N} \sum R_{nk} \|x_n - \mu_k\|^2$

inertia





$$D_{nlc} = \|x_n - u_{lc}\|^2$$

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

$$\begin{aligned} \text{inertia} &= 1+1+2+1+1 \\ &= 6/5 \end{aligned}$$

K-means clustering: inertia

A performance metric, called inertia

inertia = the mean squared distance
between each instance and its closest centroid

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

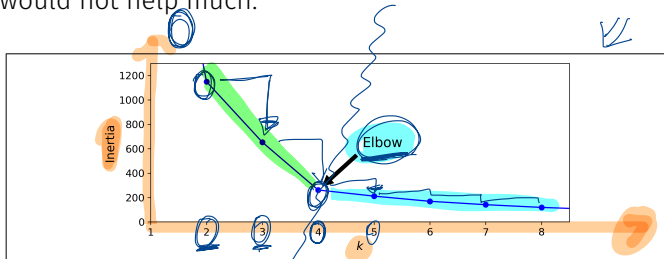
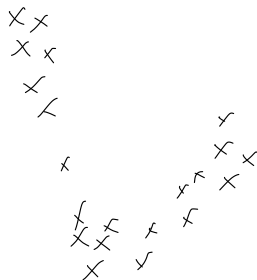
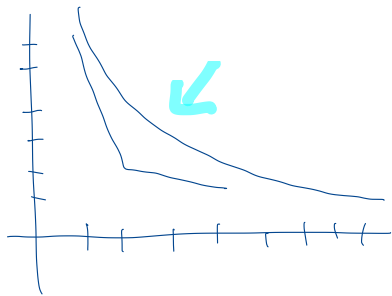
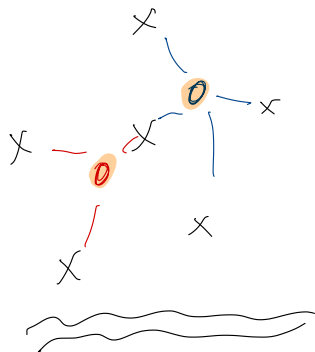
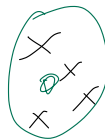
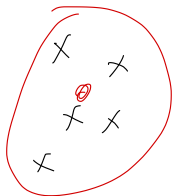


Figure 9-8. Selecting the number of clusters k using the “elbow rule”



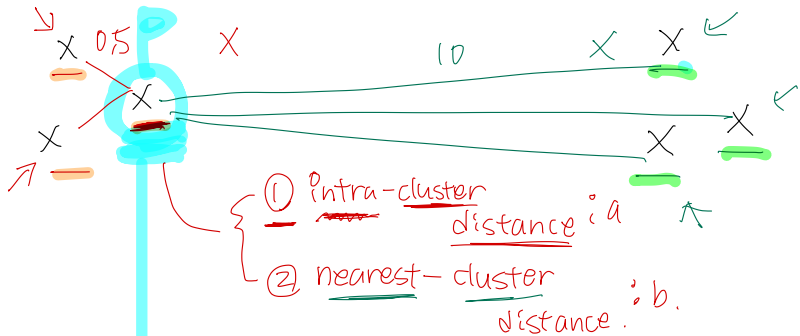




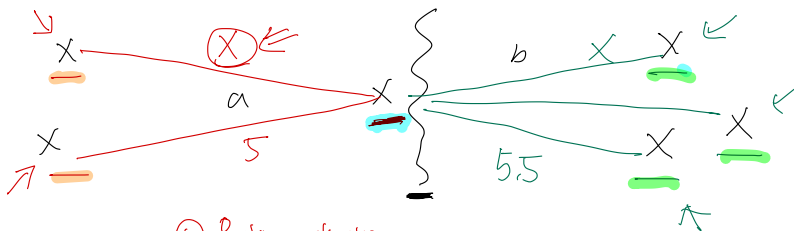
K-means clustering: silhouette score

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. 각 샘플마다.
[-1, +1] 실수.

- 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).

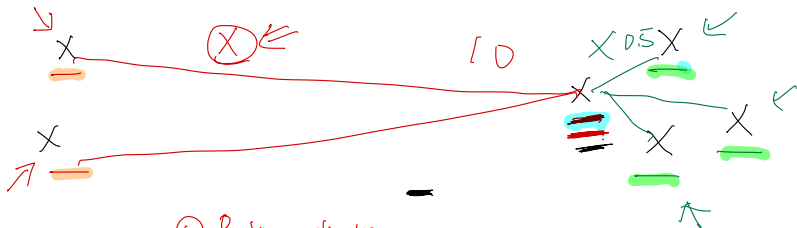


Silhouette coeff =
$$\frac{b - a}{\max(a, b)} = \frac{10 - 0.5}{10} = 0.95$$



- ① intra-cluster distance : a
- ② nearest-cluster distance : b .

Silhouette coeff = $\frac{b - a}{\max(a, b)} = \frac{0.5}{5.5} \approx 0.$



- ① intra-cluster distance : a
- ② nearest-cluster distance : b.

$$\text{Silhouette coeff} = \frac{b - a}{\max(a, b)} = \frac{-9.5}{10} = -0.95$$

Silhouette
coeff

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

$$\downarrow$$
$$\underline{b} \gg \underline{a}$$

$$\frac{b}{b} \simeq \underline{+1}$$

$$a \simeq b$$

$$\frac{b-a}{\max(a,b)} \simeq \underline{0}$$

$$\underline{a} \gg \underline{b}$$

$$\frac{-a}{a} \simeq \underline{-1}$$

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

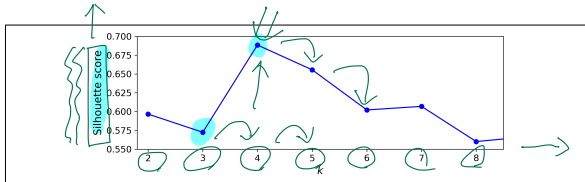


Figure 9-9. Selecting the number of clusters k using the silhouette score coefficient. This is called a *silhouette diagram* (see Figure 9-10):

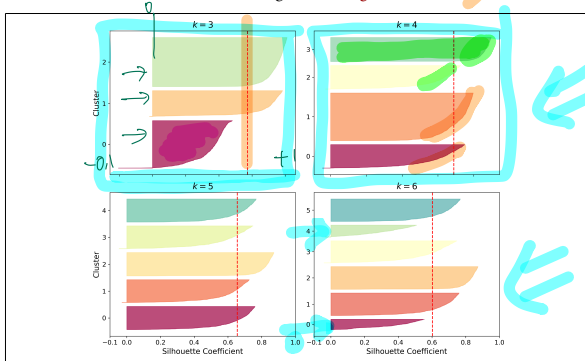


Figure 9-10. Silhouette analysis: comparing the silhouette diagrams for various values of k

다항식

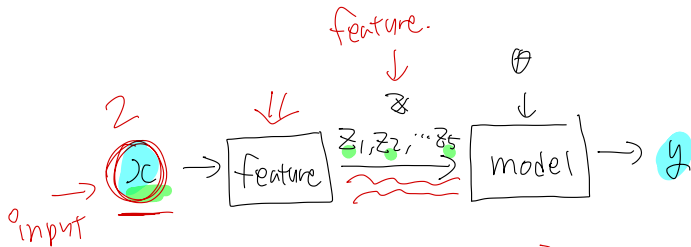
회귀

Polynomial Regression and Regularization

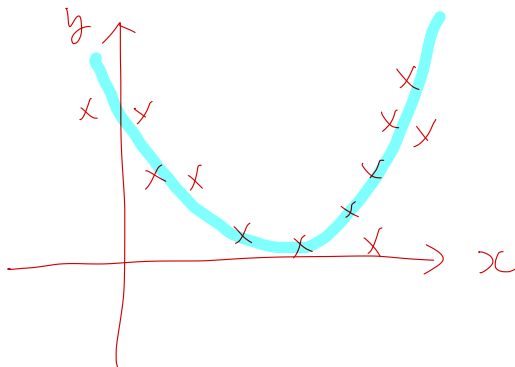
$$y(x) = \theta^T x$$

$$= \theta_1 z_1 + \theta_2 z_2 + \theta_3 z_3 + \theta_4 z_4 + \theta_5 z_5$$

$$= \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4 + \theta_5 x^5$$



$$z_1 = x, \quad z_2 = x^2, \quad z_3 = x^3, \quad z_4 = x^4, \quad z_5 = x^5$$



$$\begin{aligned}
 \underline{y} &= \theta_0 + \theta_1 x + \theta_2 x^2 \\
 &= \theta_0 z_0 + \theta_1 z_1 + \theta_2 z_2 \\
 &\quad \quad \quad 1 \quad \quad x \quad \quad x^2
 \end{aligned}$$

$$\begin{bmatrix} x & x^2 & y \\ 1 & 1 & 4 \\ 2 & 4 & 2 \\ 3 & 9 & 1 \\ 4 & 16 & 2 \\ 5 & 25 & 4 \end{bmatrix}$$

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

$$X = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \\ 1 & 5 & 25 \end{bmatrix} \quad y = \begin{bmatrix} 4 \\ 2 \\ 1 \\ 2 \\ 4 \end{bmatrix}$$

5×3

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

$3 \times 3 \quad 3 \times 5 \quad 5 \times 1$

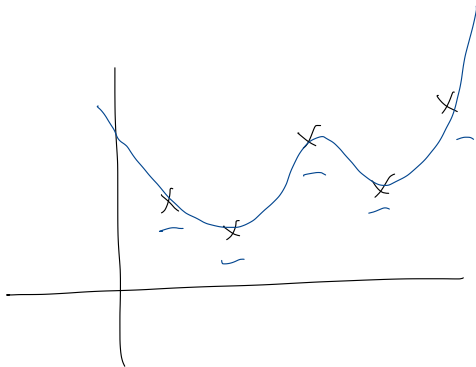
$$\Rightarrow \underline{3 \times 1}$$

$$\Theta = \Theta - 2\eta X^T(X\Theta - y)$$

$$= \Theta - 2\eta \sum (\Theta^T x_n - y_n) x_n$$

$$= \Theta - 2\eta \sum_{n=1}^N \left\{ (\theta_0 + \theta_1 x_n + \theta_2 x_n^2) - y_n \right\} \begin{bmatrix} 1 \\ x_n \\ x_n^2 \end{bmatrix}$$

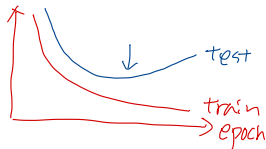
GR



Overfitting and underfitting

과적합 overfitting

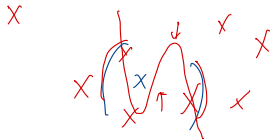
- 학습 데이터에 대해서는 좋은 성능을 보이지만
- 처음 보는 데이터에 대해서는 일반화하지 못함
- 학습 데이터의 양이나 노이즈에 비해 모델이 너무 복잡할 때 ↑
- 모델이 학습 데이터를 일반화하는 범위를 넘어서 과도하게 의존함
- 모델이 학습 데이터를 기억하는 듯한 양상을 보이기 시작



미적합 underfitting

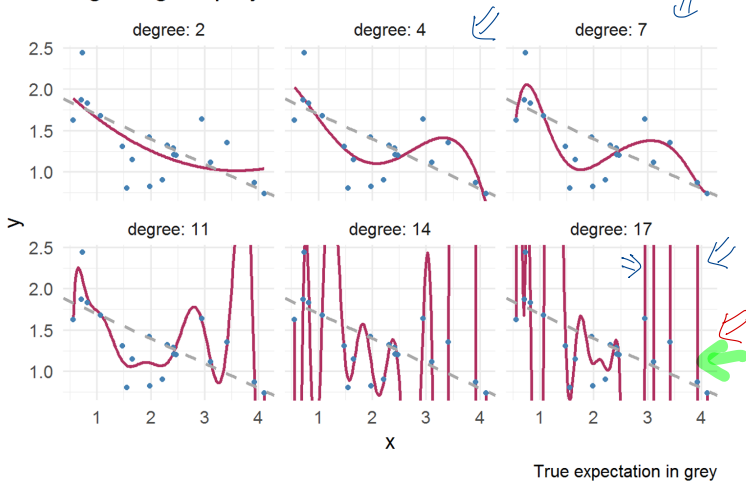
- 모델이 너무 단순하여 학습 데이터에 대해서도 부정확

training
정확.
X

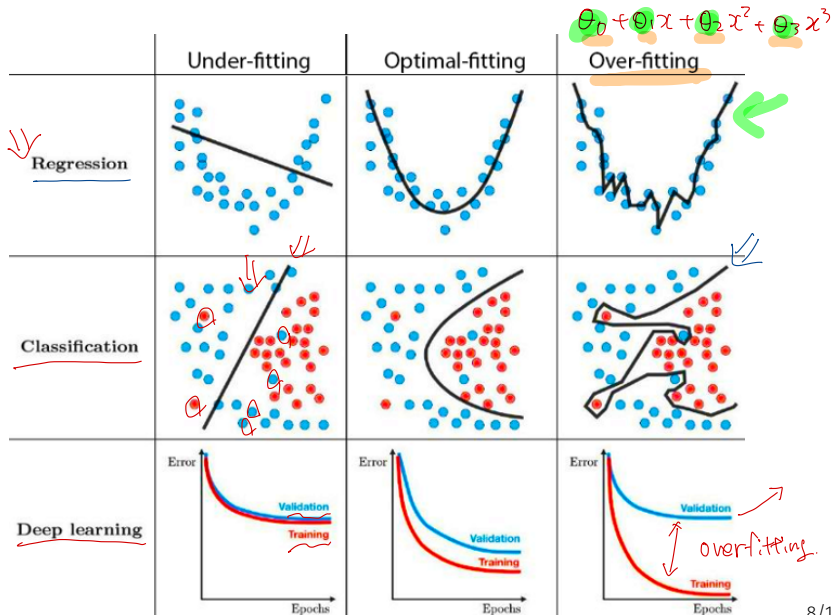


Overfitting and underfitting

High degree polynomial models fit data better



Overfitting and underfitting



Regularization

정규화 regularization

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

모델의 복잡도

$$\Rightarrow \|\theta\|_1, \|\theta\|_2$$

Ridge regression

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

trade-off

$\alpha \uparrow$, 복잡도 낮추는데
 $\alpha \downarrow$,
(1) 정중.

LASSO

$$J(\theta) = \text{SSE}(\theta) + \alpha \sum_i |\theta_i| \quad (2)$$

Ridge regression

Ridge regression

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

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

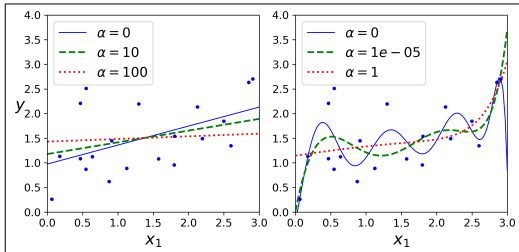


Figure 4-17. Ridge Regression

Lasso regression

LASSO

~~closed~~ 근사.

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

feature selection ←

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

매개변수

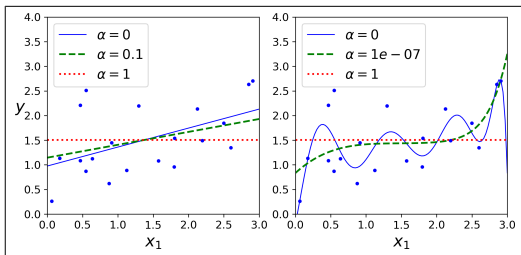


Figure 4-18. Lasso Regression

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

LASSO ← data

$$\theta = \begin{bmatrix} 1 & 2 & 0 & 4 \end{bmatrix}^T$$

$\theta_0 \theta_1 \theta_2 \theta_3$