

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

Clustering Model Evaluation

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.

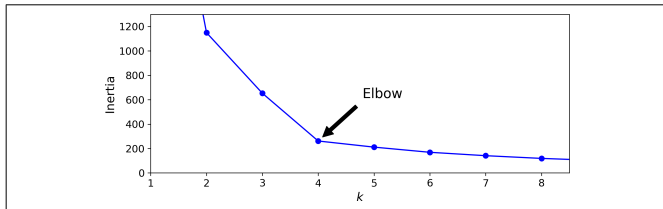


Figure 9-8. Selecting the number of clusters k using the “elbow rule”

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.

- 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).

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

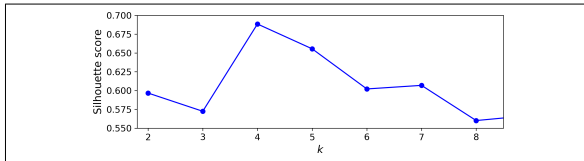


Figure 9-9. Selecting the number of clusters k using the silhouette score coefficient. This is called a *silhouette diagram* (see Figure 9-10):

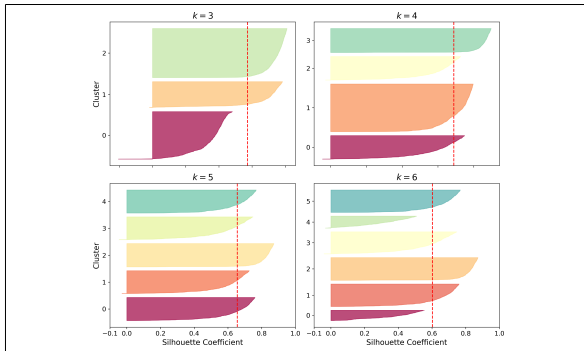
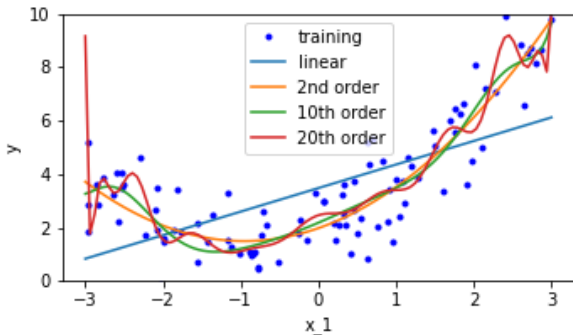


Figure 9-10. Silhouette analysis: comparing the silhouette diagrams for various values of k

Polynomial Regression and Regularization

Polynomial regression

모델의 차수는 어떻게 결정해야 하는가?



Overfitting and underfitting

과적합 overfitting

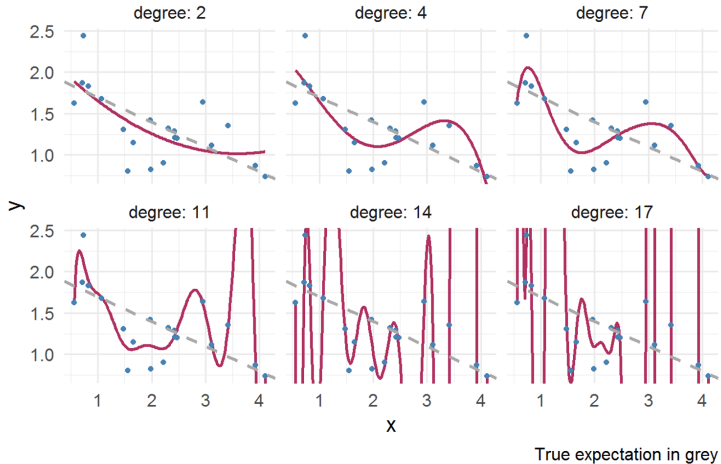
- 학습 데이터에 대해서는 좋은 성능을 보이지만
- 처음 보는 데이터에 대해서는 일반화하지 못함
- 학습 데이터의 양이나 노이즈에 비해 모델이 너무 복잡할 때
- 모델이 학습 데이터를 일반화하는 범위를 넘어서 과도하게 의존함
- 모델이 학습 데이터를 기억하는 듯한 양상을 보이기 시작

미적합 underfitting

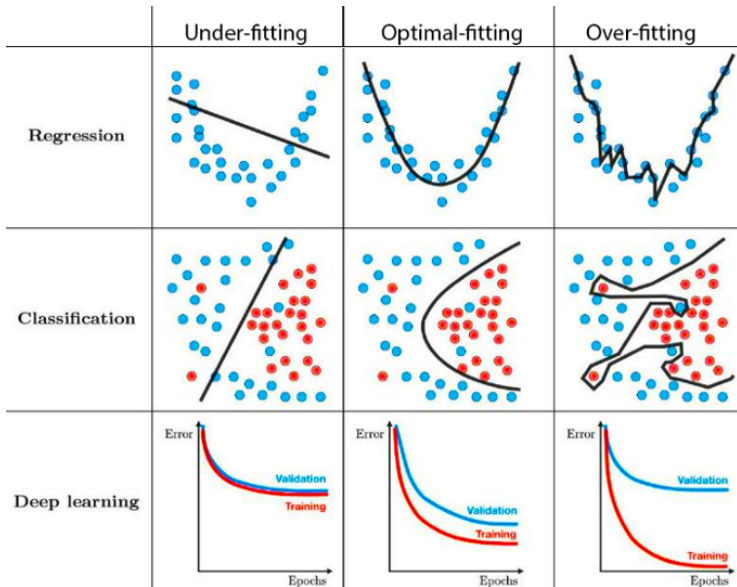
- 모델이 너무 단순하여 학습 데이터에 대해서도 부정확

Overfitting and underfitting

High degree polynomial models fit data better



Overfitting and underfitting



Regularization

정규화 regularization

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

Ridge regression

$$J(\boldsymbol{\theta}) = \text{SSE}(\boldsymbol{\theta}) + \frac{\alpha}{2} \|\boldsymbol{\theta}\|^2 \quad (1)$$

LASSO

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

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 $\hat{\boldsymbol{\theta}} = (\mathbf{X}^T \mathbf{X} + \alpha \mathbf{I}_n)^{-1} \mathbf{X}^T \mathbf{y}$

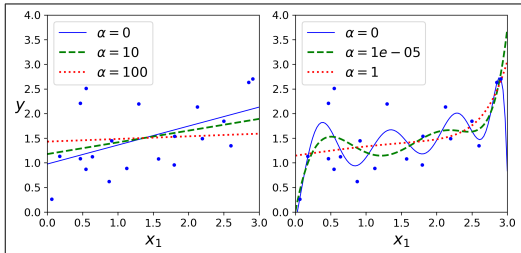


Figure 4-17. Ridge Regression

Lasso regression

LASSO

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

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

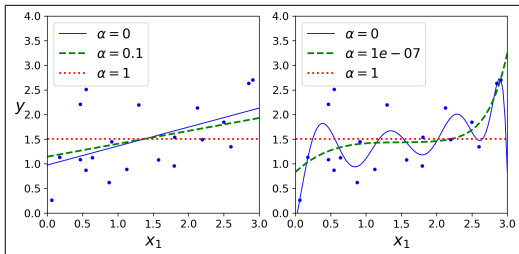


Figure 4-18. Lasso Regression