Inclass 13: Model Evaluation and Selection

[SCS4049] Machine Learning and Data Science

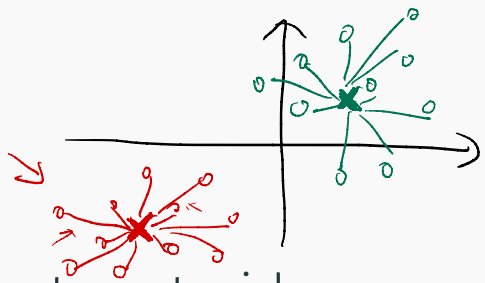
Seongsik Park (s.park@dgu.edu)

AI Department, Dongguk University

K-means clustering: inertia

A performance metric, called *inertia*

intertia = 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.

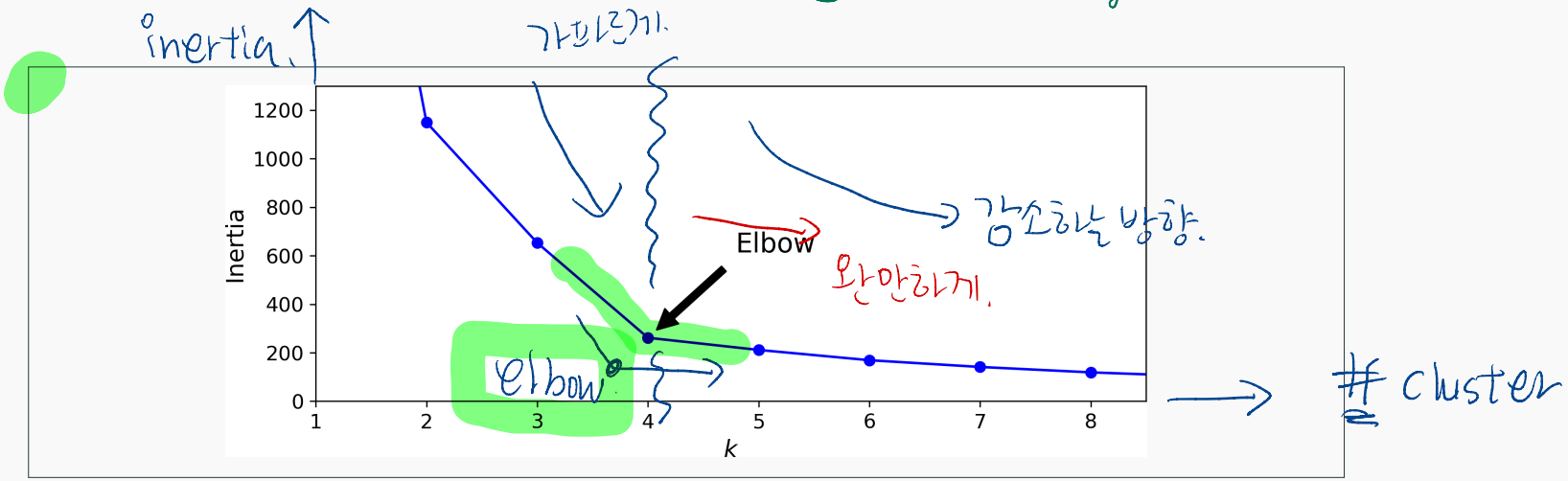


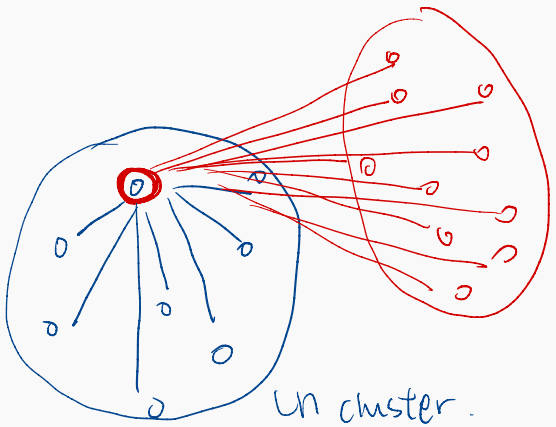
Figure 9-8. Selecting the number of clusters k using the "elbow rule"

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.

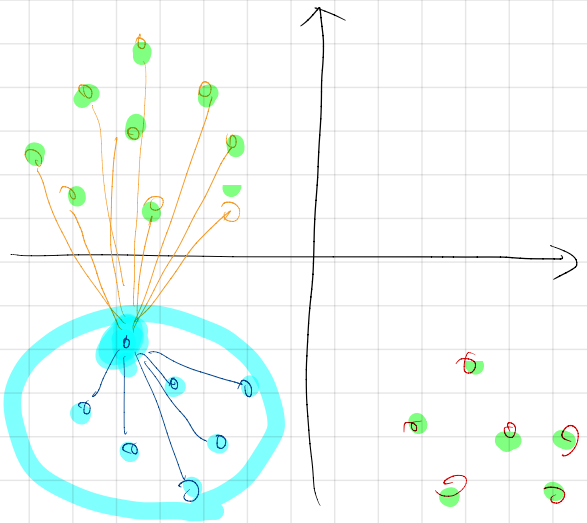
실루엣 계수 (coeff) = 각 개별의 sample (instance) 의 신뢰도 수치.

- 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 = 같은 cluster의 다른 sample까지의 거리.

b = 그 다음 가까운 cluster의 sample까지의 거리.



a = 파란색 거리의 평균

b = 초록색 거리의 평균

이 때의 실루엣 coeff

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

내가 내 클러스터에 가깝고, 잘 뭉쳐져 있음.

좋은 경우: $a \simeq 0$ $b >> a$, $b >> 0$. $\Rightarrow \frac{b-a}{\max(a,b)} \simeq \frac{b}{b} = +1$

안 좋은 경우: $a >>> b$ $b \simeq 0$. $\Rightarrow \frac{b-a}{\max(a,b)} \simeq -\frac{a}{a} = -1$

내가 내 클러스터보다 다른 클러스터에 더 가깝다.

K-means clustering: silhouette score

The silhouette coefficient can vary between -1 and +1. [-1, +1] 실수.

- 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, $a \simeq b$.
- and finally a coefficient close to -1 means that the instance may have been assigned to the wrong cluster.

Silhouette score $= \frac{1}{N} \sum_{n=1}^{N}$ Silhouette coeff.

K-means clustering: silhouette score

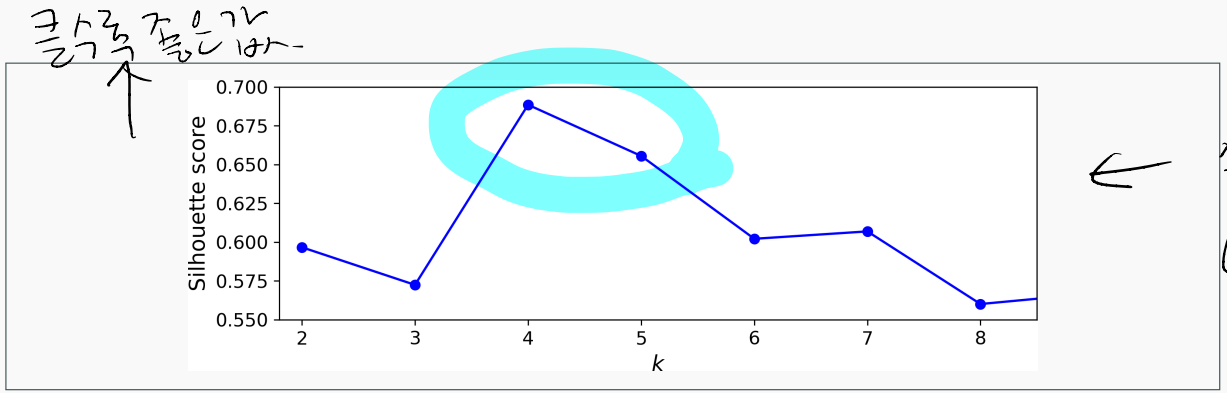


Figure 9-9. Selecting the number of clusters k using the silhouette score coefficient. This is called a *silhouette diagram* (see Figure 9-10):

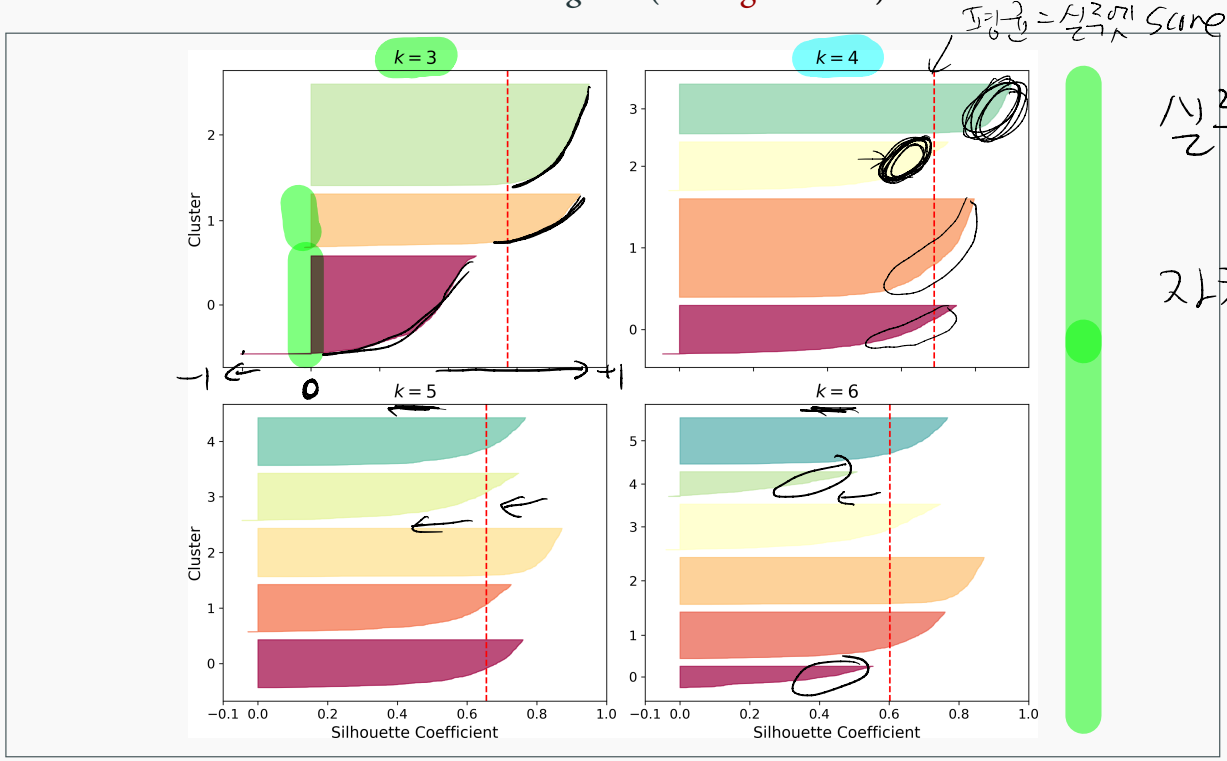


Figure 9-10. Silouhette analysis: comparing the silhouette diagrams for various values of k

MoG clustering: information criterion

- Akaike information criterion (AIC)

model complexity.

작은게 좋은거 정의.

$$\mathrm{AIC} = -\frac{2}{N} \cdot \log \Pr(\mathbf{x}_1, \ldots, \mathbf{x}_N \mid \boldsymbol{\theta}) + 2 \cdot \frac{d}{N} \quad (1)$$

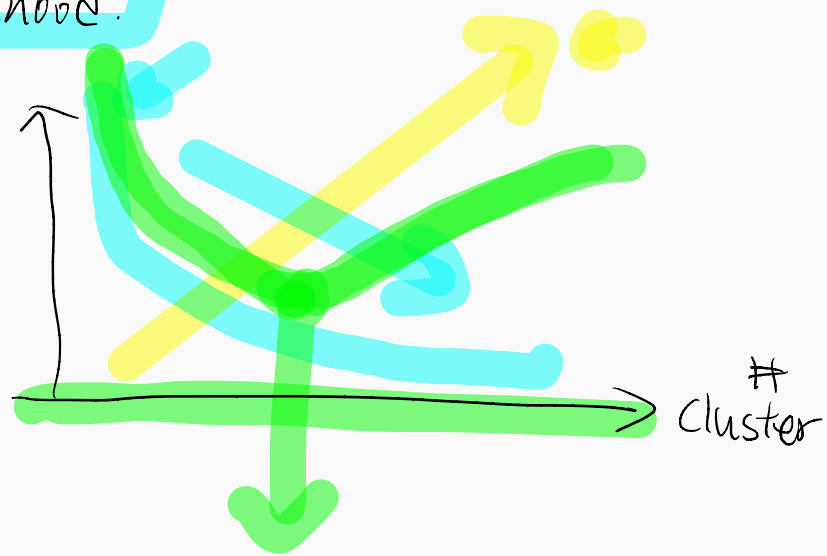
- Bayesian information criterion (BIC)

MoG 학습 $\theta = \{(\pi_n, \mu_n, \sigma_n)\}$

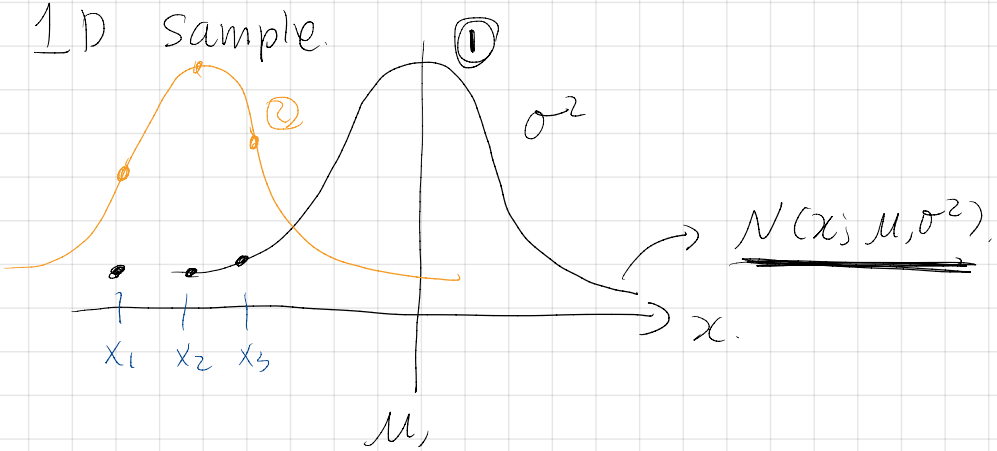
$$\mathrm{BIC} = -2 \cdot \log \Pr(\mathbf{x}_1, \ldots, \mathbf{x}_N \mid \boldsymbol{\theta}) + d \cdot \log N \quad (2)$$

log-likelihood.

- lg-like = - 클수록 좋은값.
= 작을수록 좋은값



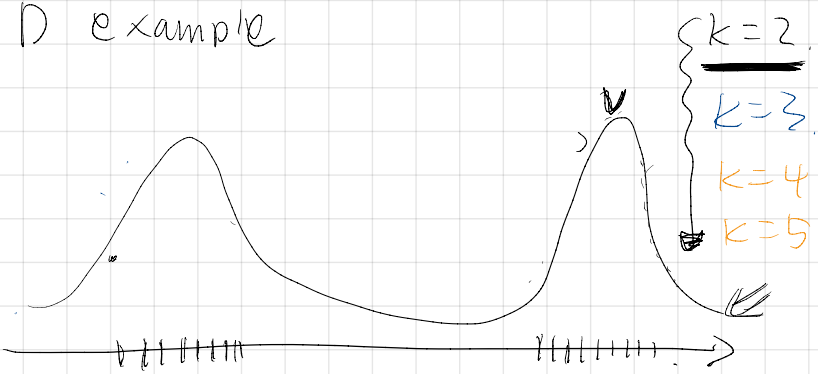
1D sample.



① $\mathcal{L}(\mu, \sigma^2; x_1, x_2, x_3) = \prod N(x_n; \mu, \sigma^2)$

② $\mathcal{L}(\mu, \sigma^2; x_1, x_2, x_3) = \prod N(x_n; \mu, \sigma^2)$ ← 큰값.

1D example



model eval, selection

① $k=2$, $f(x) = f(x; \pi_1, \pi_2, \mu_1, \mu_2, \sigma_1, \sigma_2)$

$$\mathcal{L}(\pi_1, \pi_2, \mu_1, \mu_2, \sigma_1, \sigma_2 ; x_1, x_2, \cdots, x_N) = \prod_{n=1}^{N} f(x_n)$$

② $k=3$ $f(x) = f(x; \pi_1, \pi_2, \pi_3, \mu_1, \mu_2, \mu_3, \sigma_1, \sigma_2, \sigma_3)$.

$$\mathcal{L}(\cdots\cdots ; x_1, \cdots, x_N) = \prod f(x_n)$$