k-means clustering.

- cluster의 개수 → 우리가 미리 설정.
- cluster의 개수를 어떻게 선택할것인지?
- 우리가 더 좋은 결과를 보는지.
  → measure, 척도.

Inclass 16: Model Evaluation and Selection

[SCS4049] Machine Learning and Data Science

Seongsik Park (s.park@dgu.edu)

Department of Artificial Intelligence, Dongguk University

k-means clustering.

① inertia (cost).

② silhouette score.

#cluster 개수.

- 중간고사 점수 → 27일(수), 가형, eclass.
- 이번주 과제 → 구글 설문, 강의 의견, 전반적 의견 제출. eclass.
  개별으로 (제출여부 확인) 이름·학번 입력 한번은.
- 자습시간 → 음악·MV·live 틀어볼까? 추천곡·영상 제보.

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.

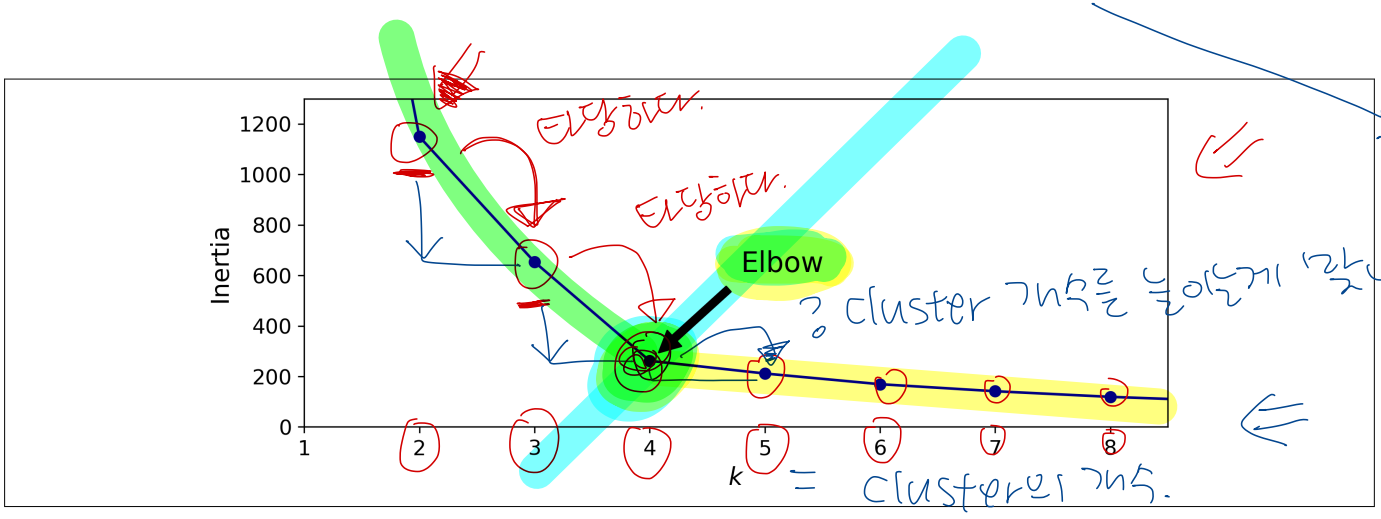
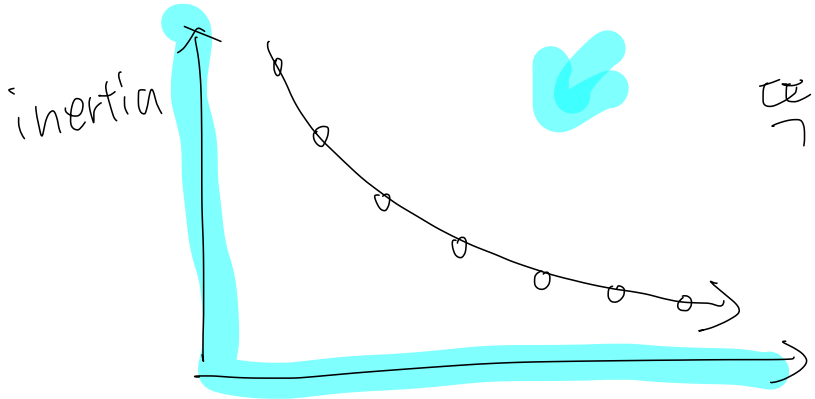
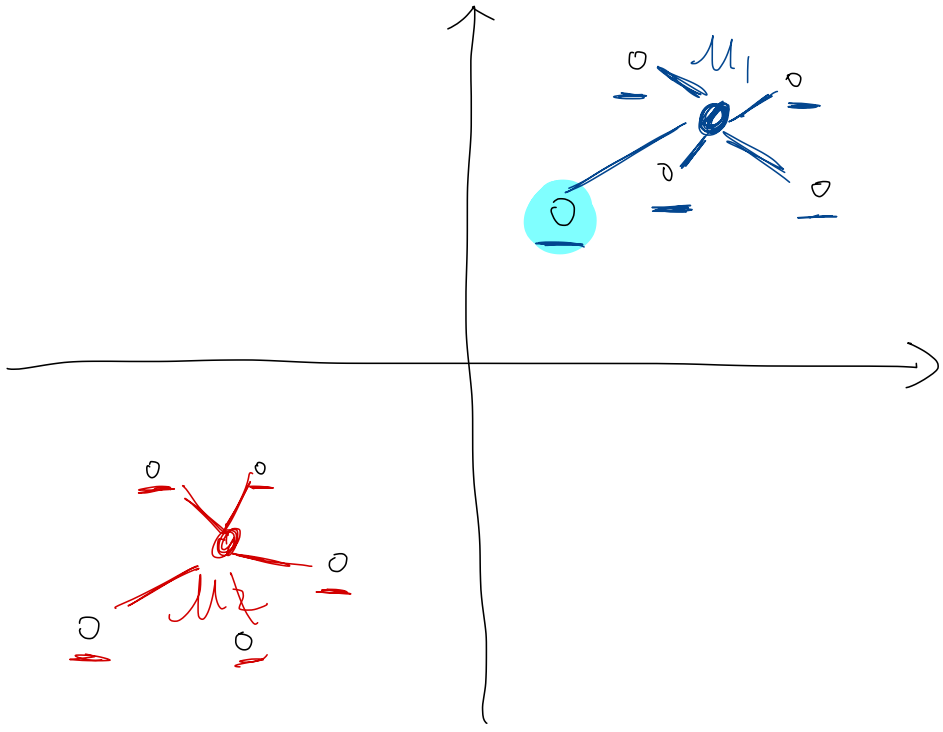


Figure 9-8. Selecting the number of clusters k using the "elbow rule"



떨어지지 않을 수 있다.

2D



cluster의 개수 = 2개.

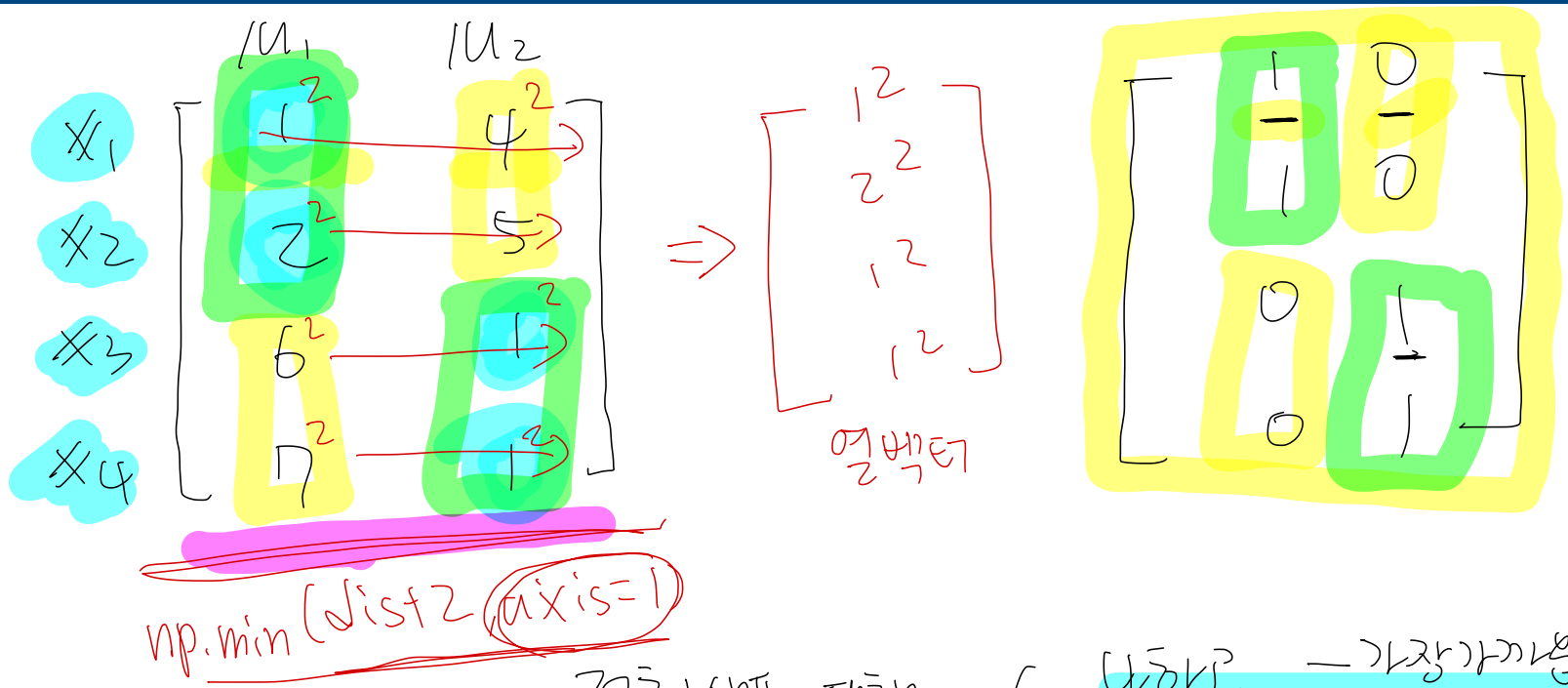
inertia

= 각각의 샘플에 대한 평균

$\left(\frac{\text{거리}^2}{\text{나 자신 = 샘플 S}} \right)$ 나랑 제일 가까운 평균.

$$\text{inertia} = \frac{1}{N} \sum_{n=1}^{N} r_{nk} \|x_n - \mu_k\|^2$$

r_{nk}: $\begin{cases} 0 \\ 1 \end{cases}$ n번째 sample이 k번째 cluster에 속하는가.



inertia = 전체 샘플에 대하여 (내가 속한 군집 중심과의 거리의 제곱) 평균

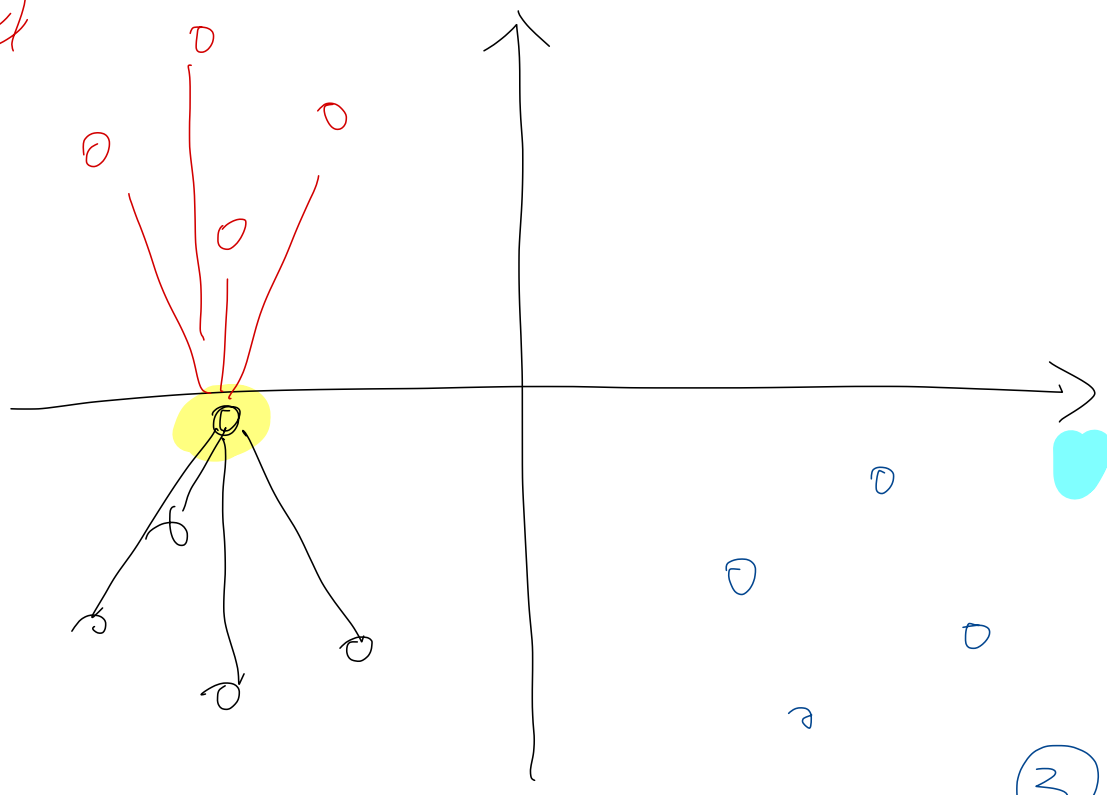
$$= \frac{1}{4}\left(1^2 + 2^2 + 1^2 + 1^2\right)$$

$$= \frac{1}{N}\sum r_{nk}\,\|x_n - \mu_k\|^2$$

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).



a = 나 vs. 내가 속한 곳의 샘플들.

=

b = 나 vs. 두번째로 가까운 cluster에 속한 샘플들.

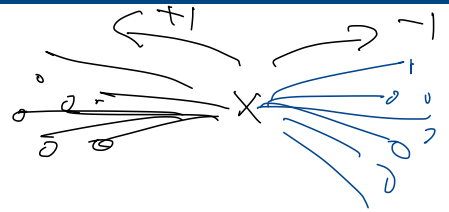
의 silhouette coeff = $\frac{b-a}{\max(a,b)}$

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.

Silhouette coeff $= \dfrac{b-a}{\max(a,b)}$

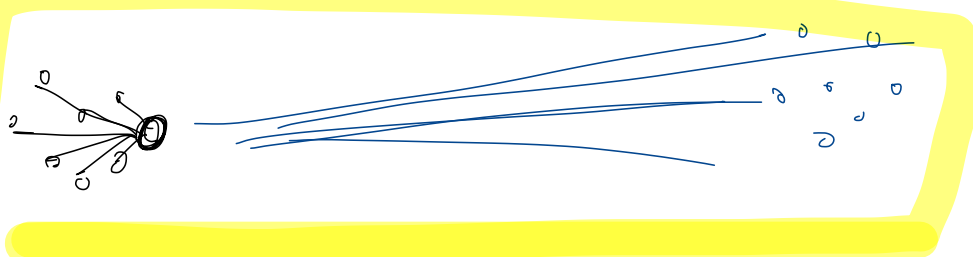


$\simeq 0$.

$a, b \geq 0$.

① $a \downarrow, b \uparrow$ = 나는 내가속한애들이랑 뭉쳐있고(a↓) 그다음가까운 cluster까지는 출발히멀리있음(b↑)

$\dfrac{b-a}{\max(a,b)} \simeq \dfrac{b}{b} = 1$



② $a \gg b$, $a \uparrow, b \downarrow$ = 나는 내가속한데랑 멀고(a↑) 가장이웃한곳이랑 가깝다(b↓)

$\dfrac{b-a}{\max(a,b)} \simeq \dfrac{-a}{a} = -1$.

K-means clustering: silhouette score

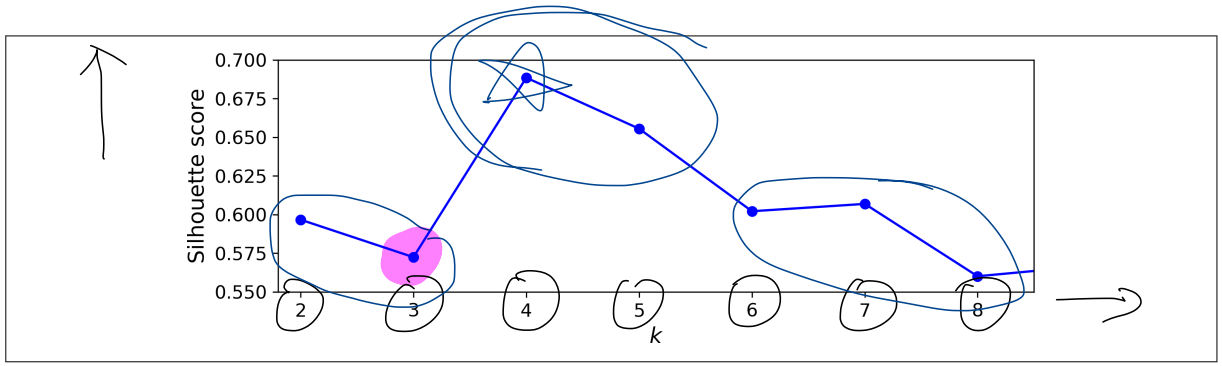


Figure 9-9. Selecting the number of clusters k using the silhouette score

coefficient. This is called a *silhouette diagram* (see Figure 9-10):

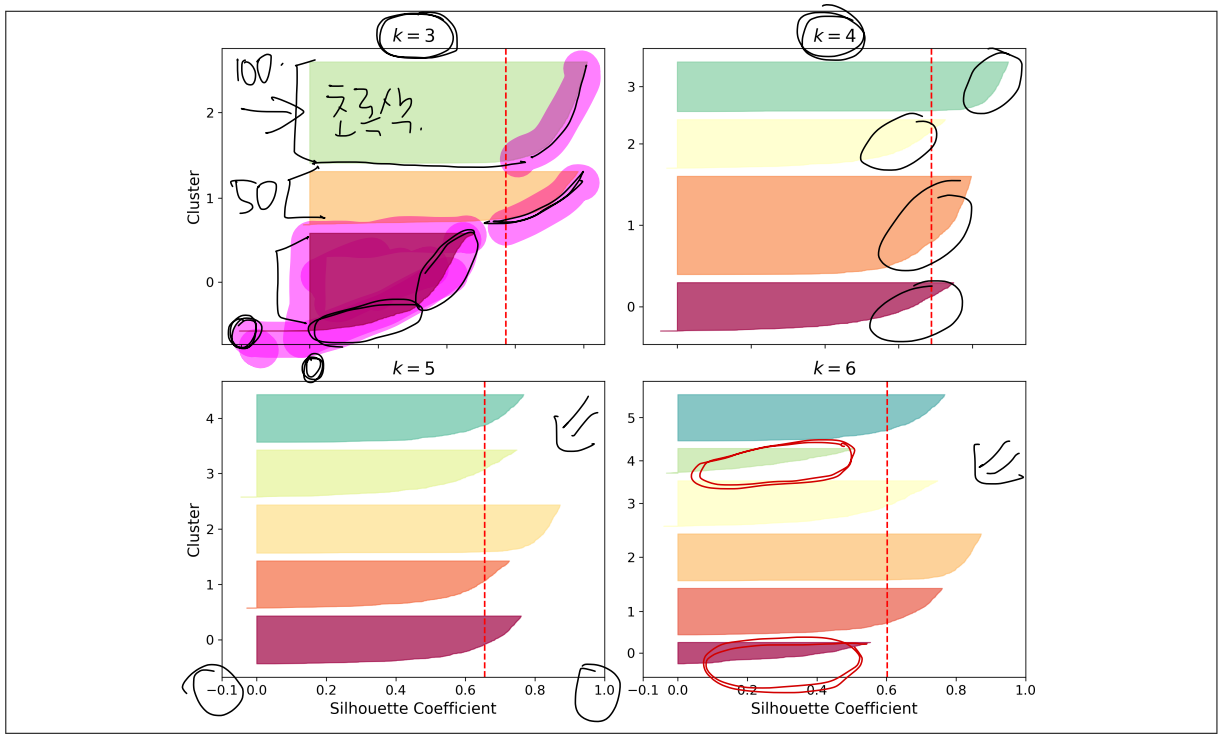


Figure 9-10. Silouhette analysis: comparing the silhouette diagrams for various values of k