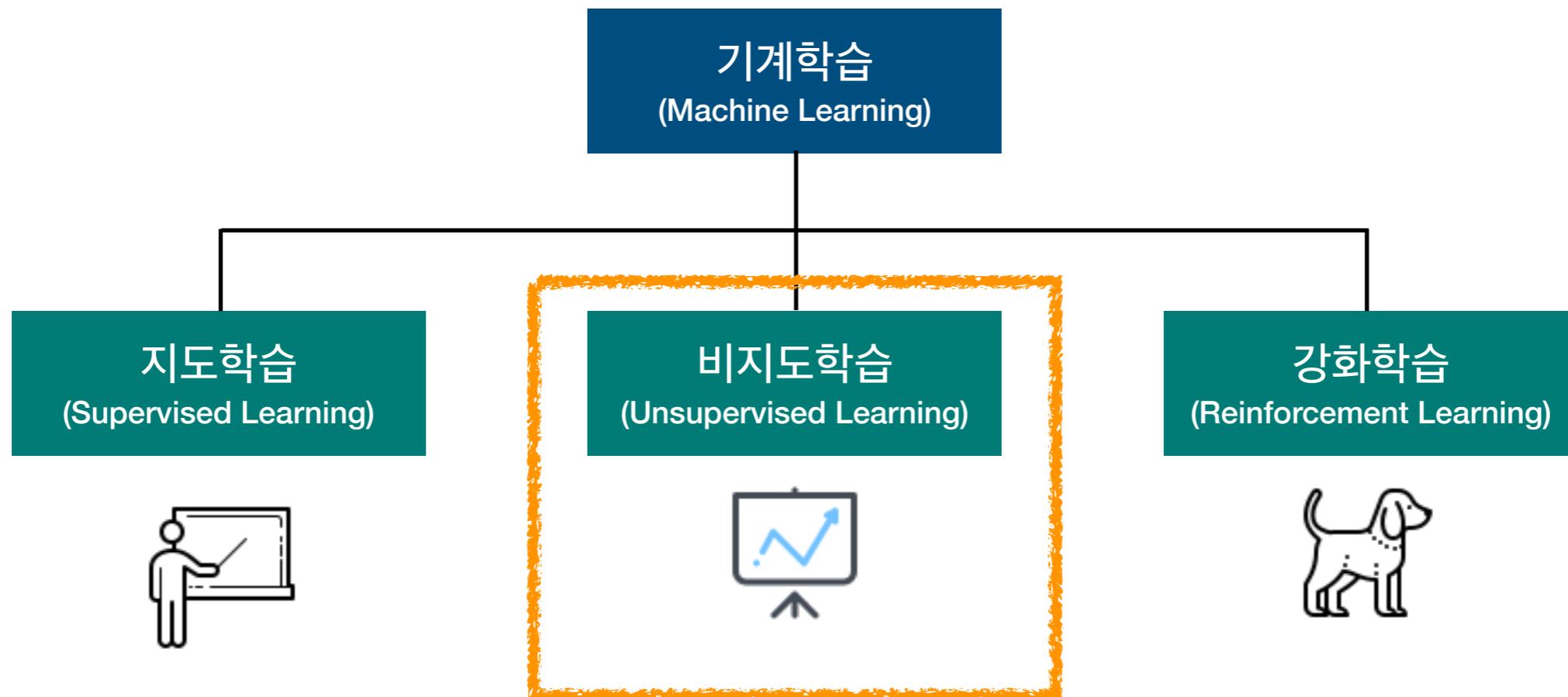


BIG DATA ANALYTICS

WEEK-09 | Unsupervised Learning

**Yonsei University
Jungwon Seo**

기계학습의 종류



비지도학습의 예

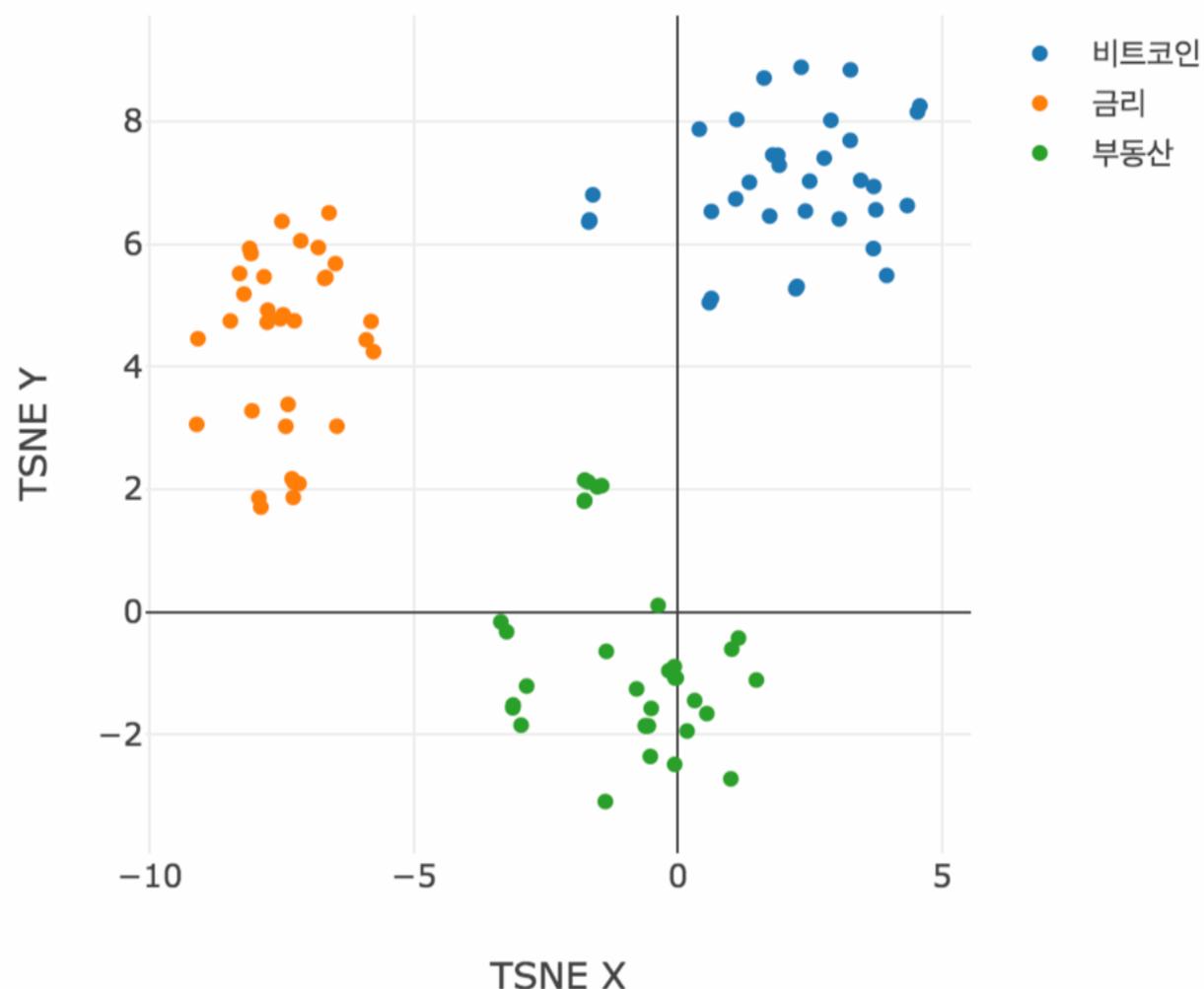
- 군집화 (Clustering)
 - Class가 명명되지 않은 상황에서 Feature들만을 이용해 데이터들을 Grouping
 - e.g., 학생들을 두 그룹으로 나누기, 뉴스 기사를 카테고리 별로 나누니
- 차원축소 (Dimension reduction)
 - 100,000개의 feature로 이루어진 데이터가 있을 때, 이 100,000개가 모두 필요한가?
 - 차원의 저주

데이터 유사도 (Data Similarity)

벡터공간 모델 (Vector Space Model)

- ▶ 각 데이터를 N 차원 상의 벡터공간에 표현하는 방법
- ▶ N은 데이터를 표현하는 특성 (feature)의 수에 따라서 결정됨

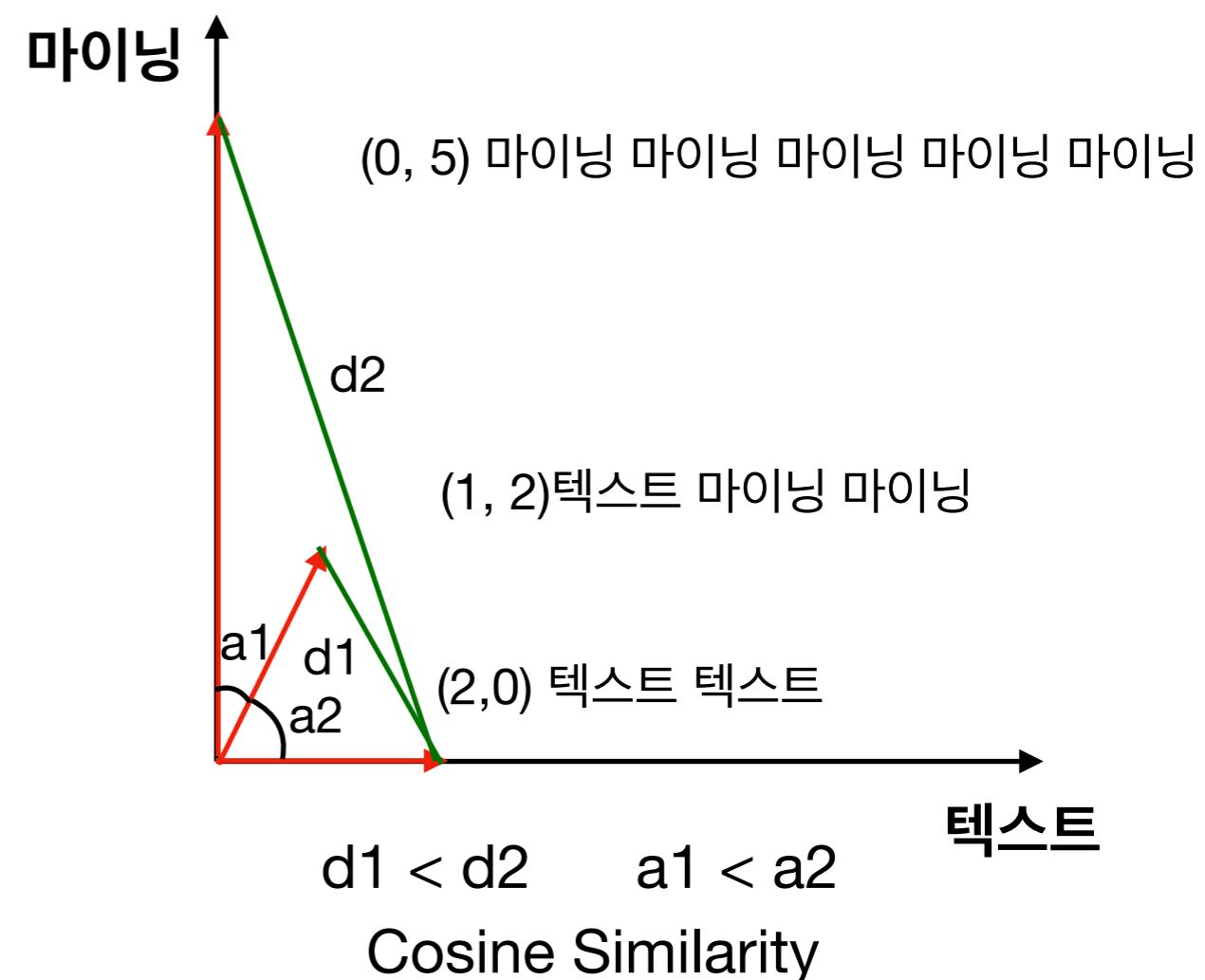
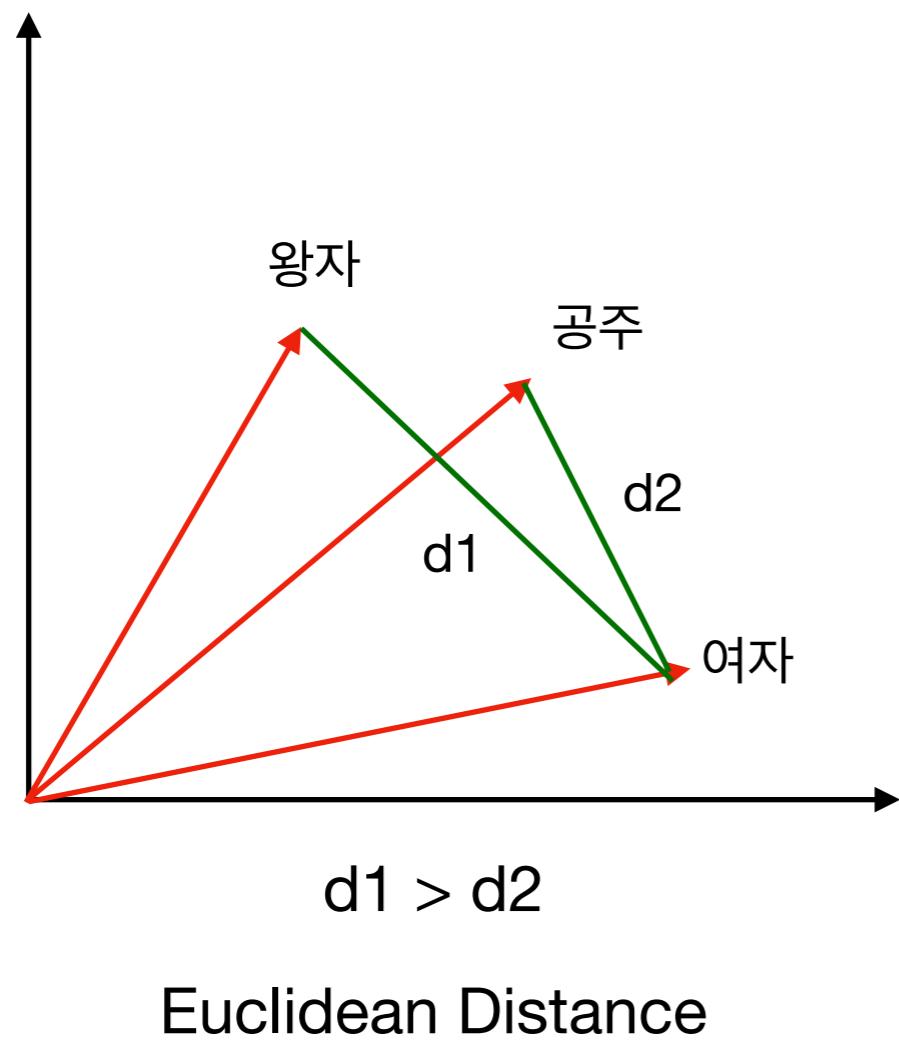
K-Means Clustering Graph - label



데이터 유사도 (Data Similarity)

벡터공간에서의 유사도

- ▶ 유클리디안 거리 (Euclidean Distance)
- ▶ 코사인 유사도 (Cosine Similarity)

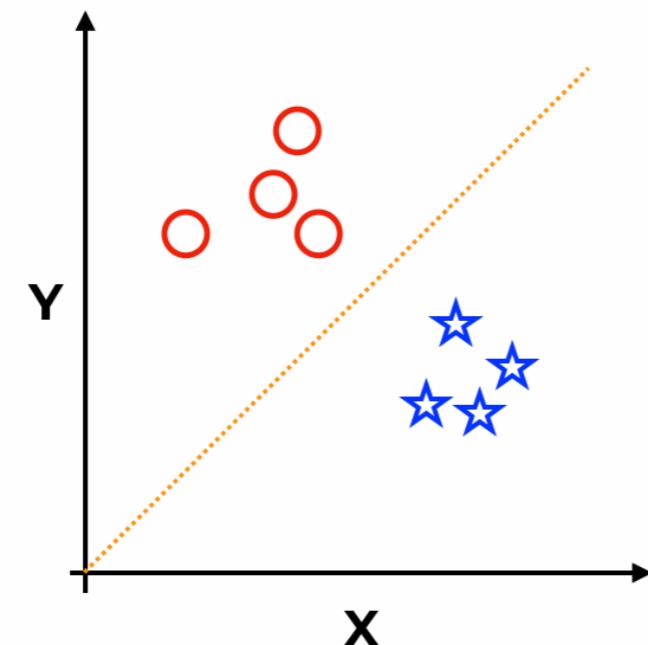


군집화 (Clustering)

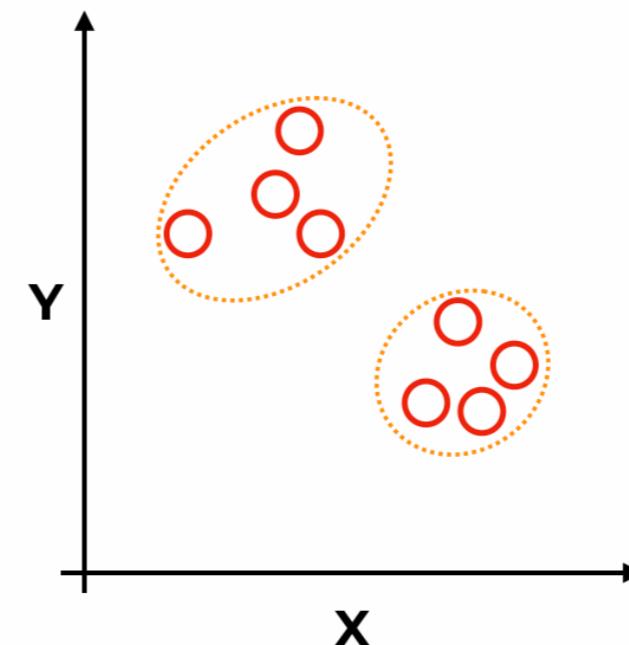
군집화란?

- ▶ 비지도 학습 (unsupervised learning)으로 데이터에서 자연스럽게 분류되는 그룹을 찾아내는 방법
- ▶ 지도 학습 (supervised learning) vs 비지도 학습
 - 지도학습 :
 y 값(타깃 값)이 사전에 정해진 경우 지도학습이며 ($y=0$ or $y=X$), 타깃값을 알고 있는 데이터를 이용해 새로운 객체의 타깃값을 예측할 수 있는 패턴을 찾아내는 것
 - 비지도학습 :
 y 값이 정해지지 않고 자율적으로 학습하는 경우 비지도 학습이라 하며, 타깃 변수에 신경쓰지 않고 데이터 집합에 있는 어떤 규칙성을 찾으려는 것

지도학습 (*Supervised Learning*)

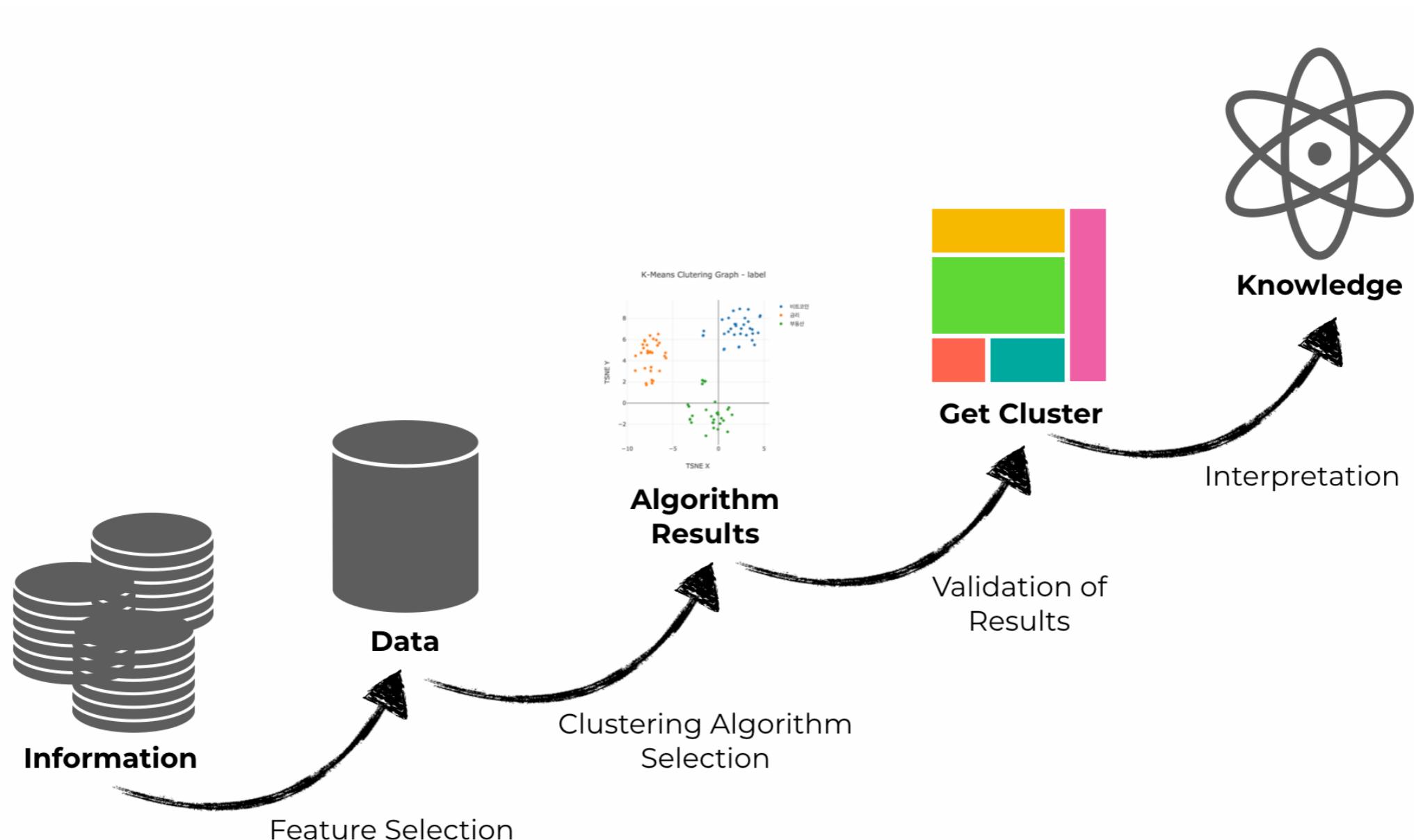


비지도학습 (*Unsupervised Learning*)



군집화 (Clustering)

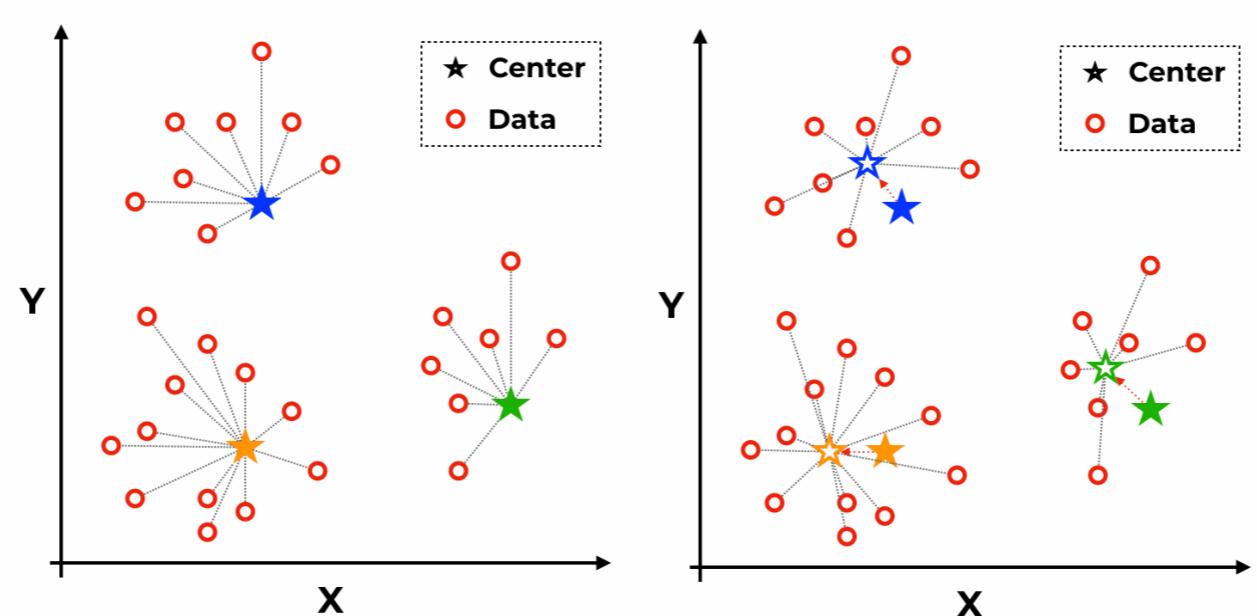
문서를 이해하기 위한 과정



군집화 (Clustering)

k-평균 군집 (K-means Clustering)

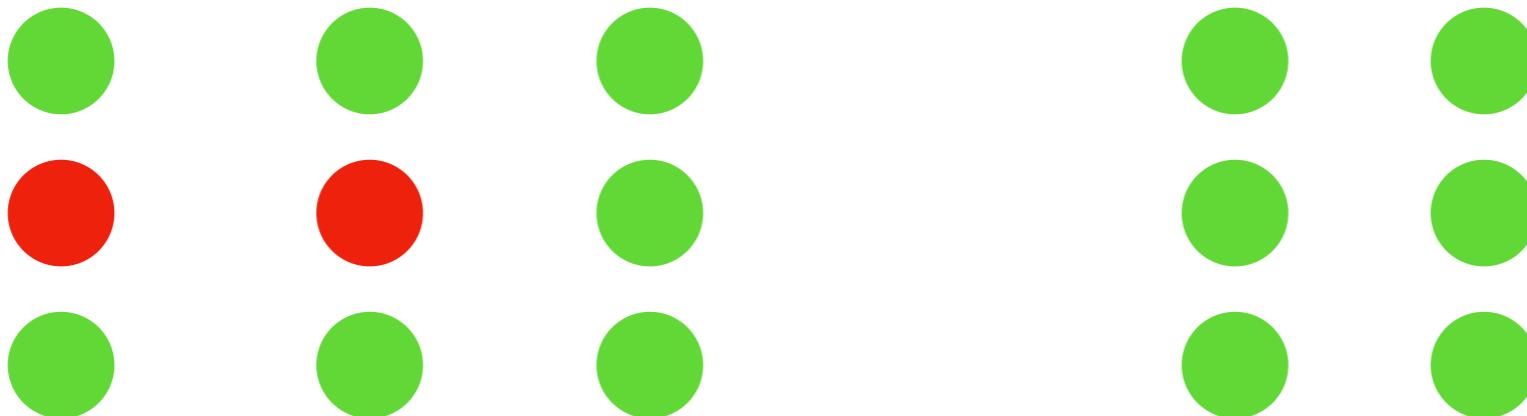
- ▶ 중점 주변 군집화 중 가장 널리 사용되며, 찾아내려는 군집 k개를 파악 ($k=$ 군집수)
- ▶ 평균은 각 군집의 중점이며, 군집에 들어있는 객체들의 각 차원별 산술평균값으로 표현
- ▶ k-평균 알고리즘의 3단계
 - Step 1 : 임의로 k 개의 데이터 포인트를 시드로 선택
 - Step 2 : 각 레코드를 가장 가까운 시드에 배정
 - Step 3 : 군집의 중심점 찾기 (다시 계산)
- ▶ 군집 중점이 이동하면 각 점마다 어느 군집에 속하는지 다시 계산해야 하고, 각 점이 속한 군집을 다시 계산한 후에는 군집의 중점을 다시 계산
- ▶ 더 이상 군집에 변화가 생기지 않을 때까지 Step 2와 Step 3 과정을 반복



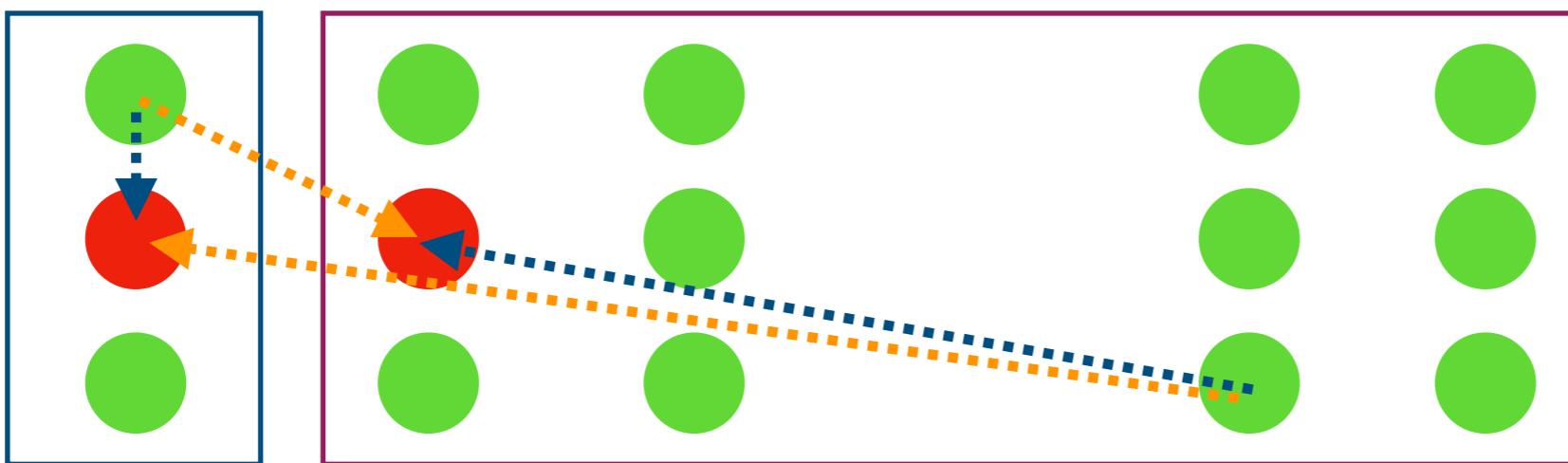
군집화 (Clustering)

k-평균 군집 (K-means Clustering)

- ▶ Step 1 : 임의로 k개의 데이터 포인트를 시드로 선택



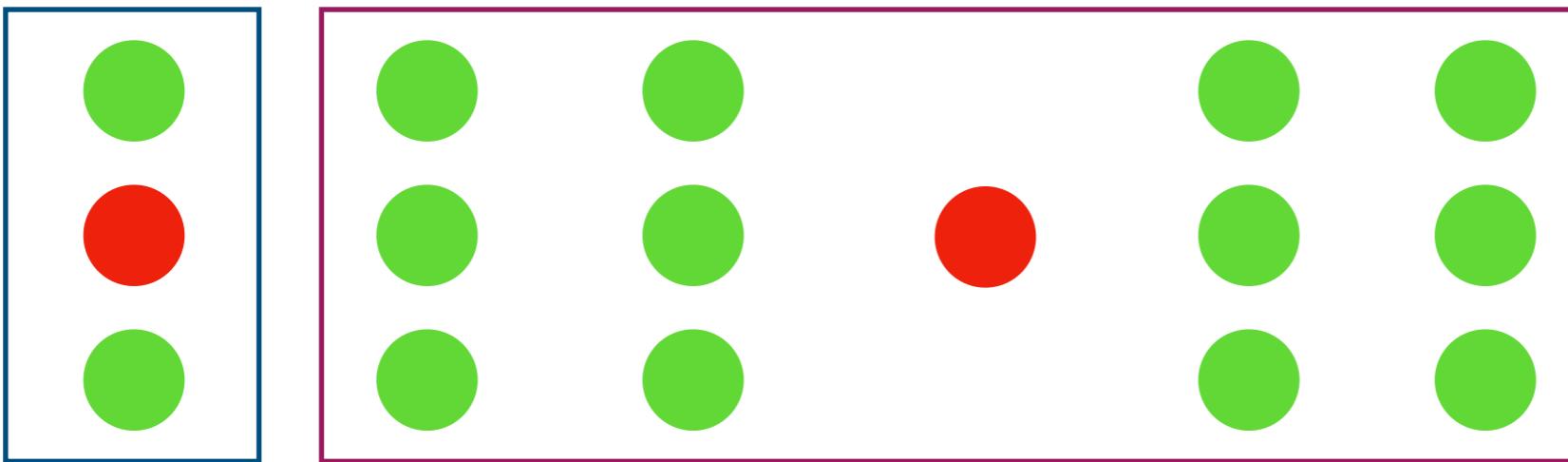
- ▶ Step 2 : 각 레코드를 가장 가까운 시드에 배정



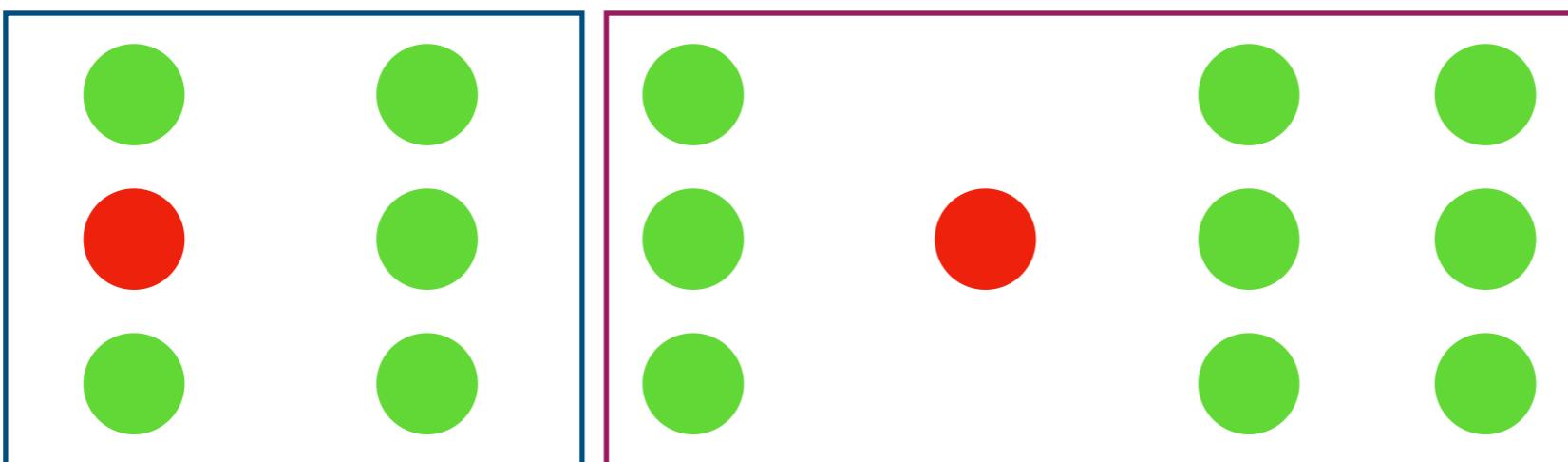
군집화 (Clustering)

k-평균 군집 (K-means Clustering)

- ▶ Step 3 : 군집의 중심점 찾기 (다시 계산)



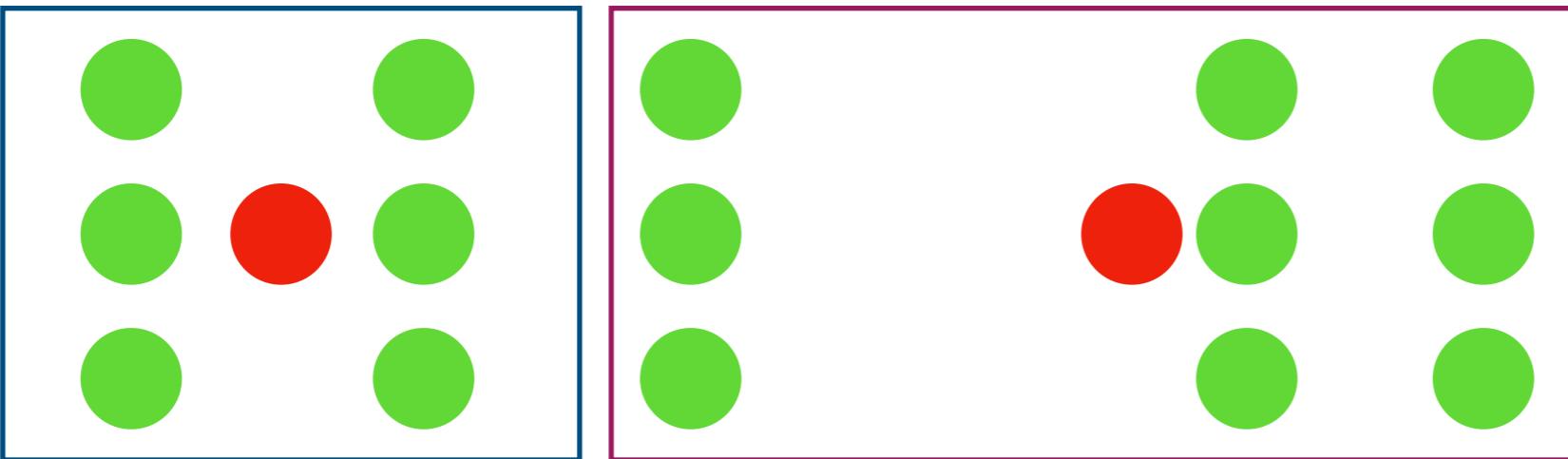
- ▶ Step 2 : 각 레코드를 가장 가까운 시드에 배정



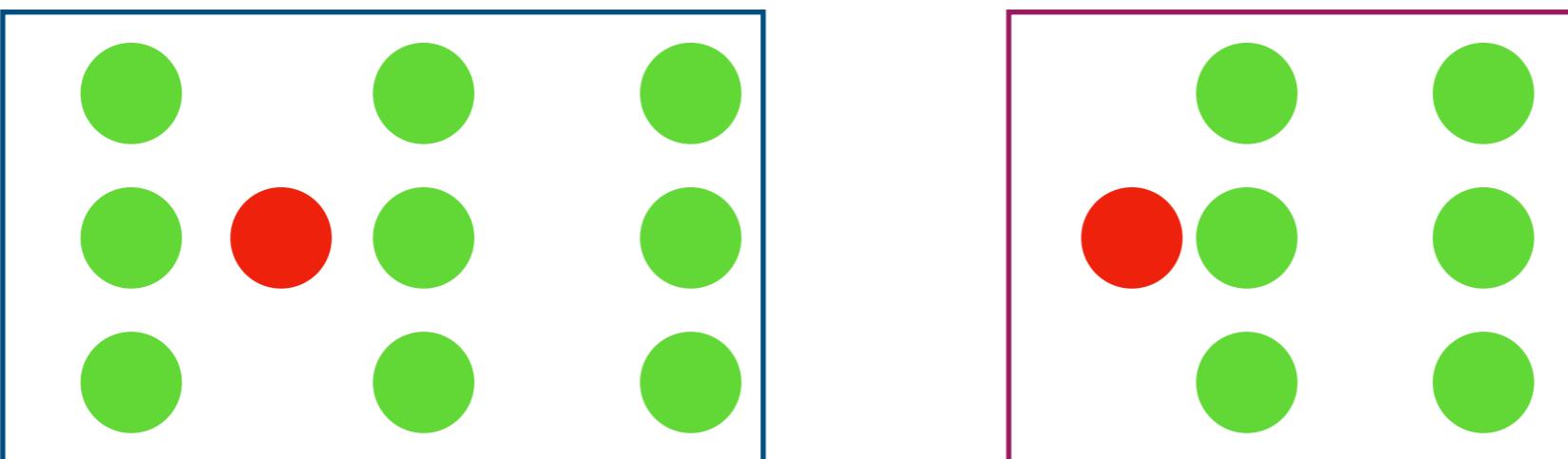
군집화 (Clustering)

k-평균 군집 (K-means Clustering)

- ▶ Step 3 : 군집의 중심점 찾기 (다시 계산)



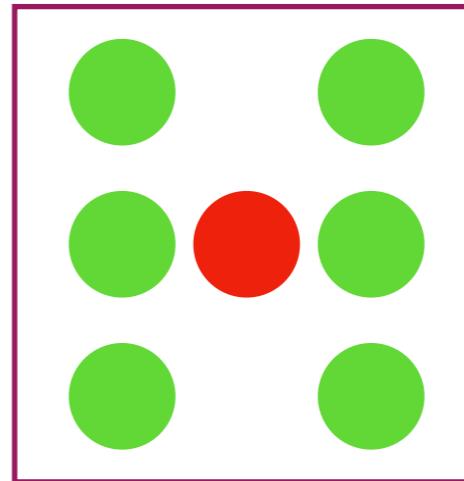
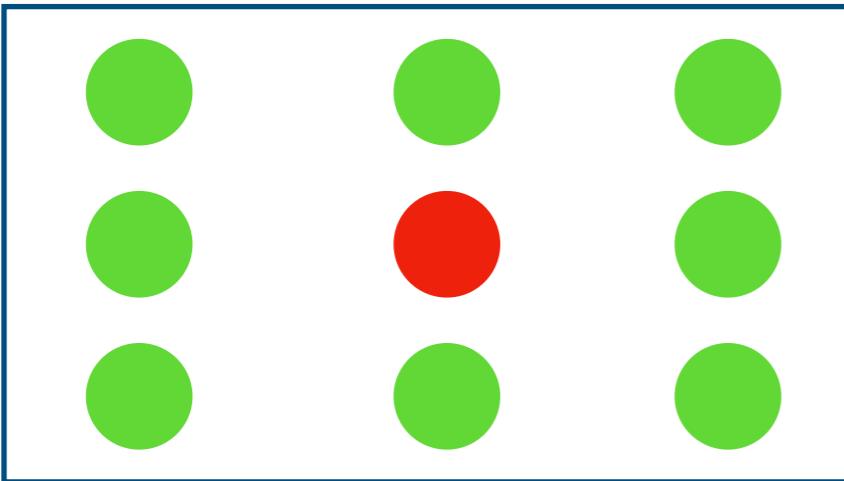
- ▶ Step 2 : 각 레코드를 가장 가까운 시드에 배정



군집화 (Clustering)

k-평균 군집 (K-means Clustering)

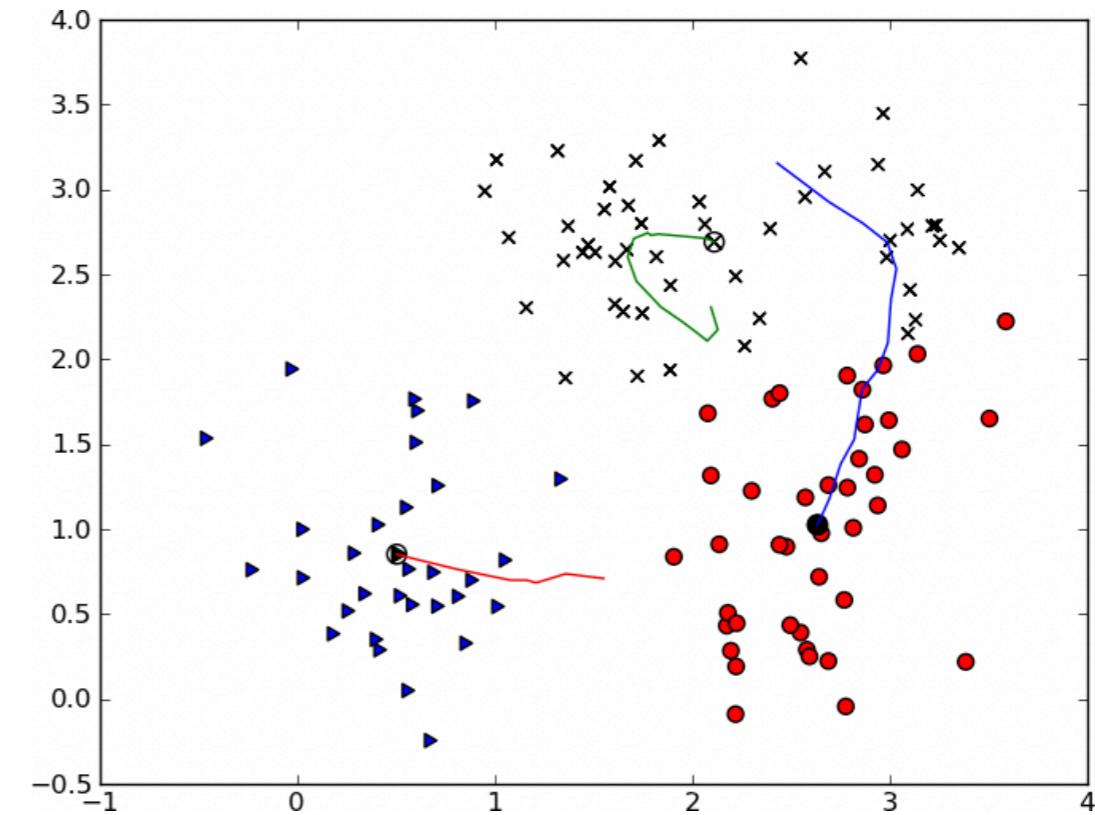
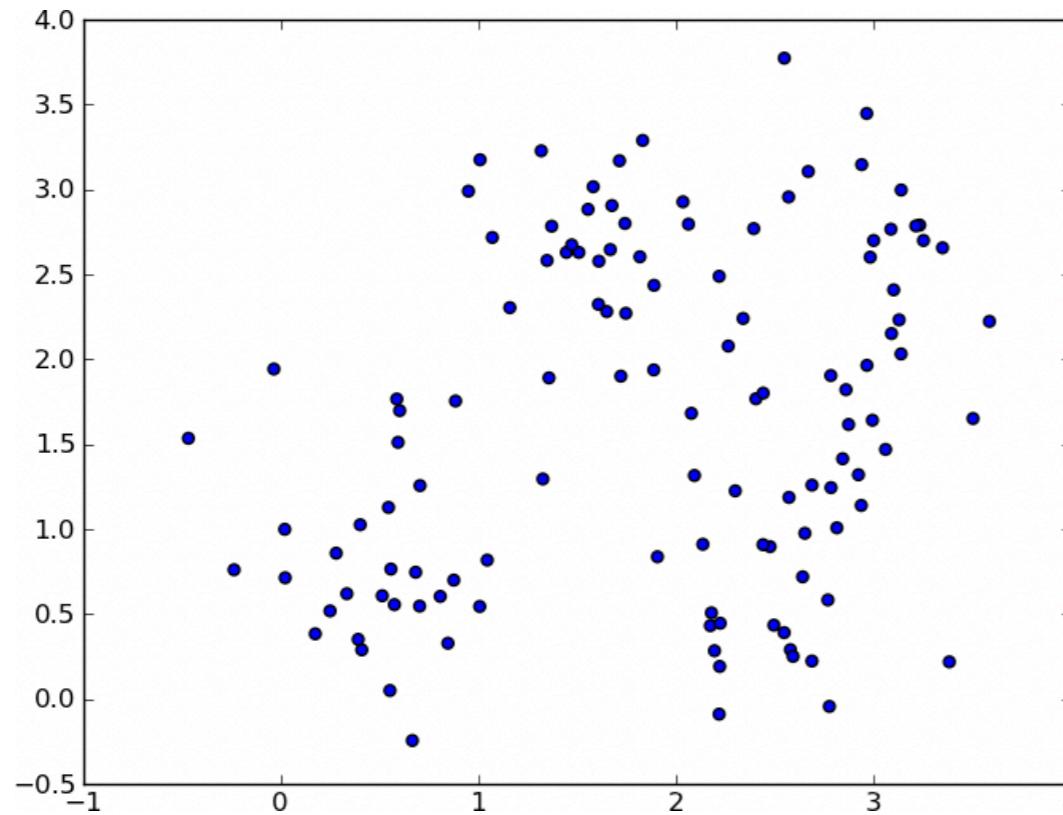
- ▶ Step 3 : 군집의 중심점 찾기 (다시 계산)



군집화 (Clustering)

k-평균 군집 (K-means Clustering)

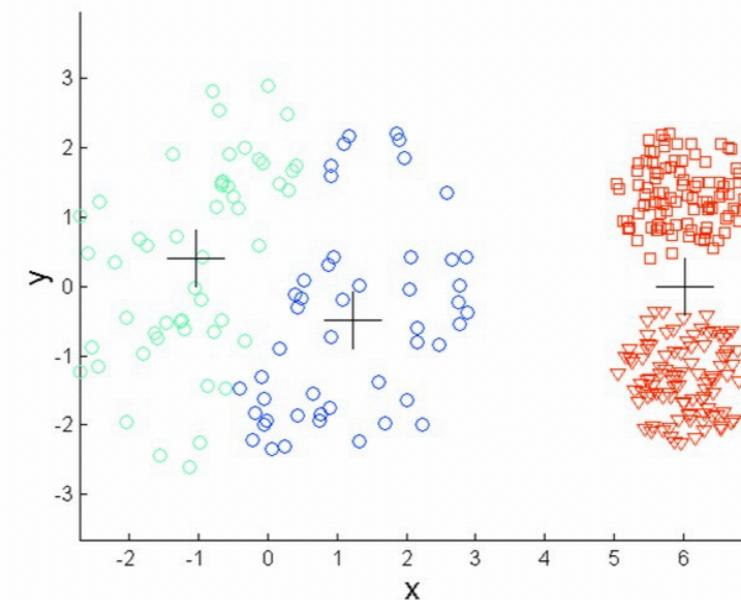
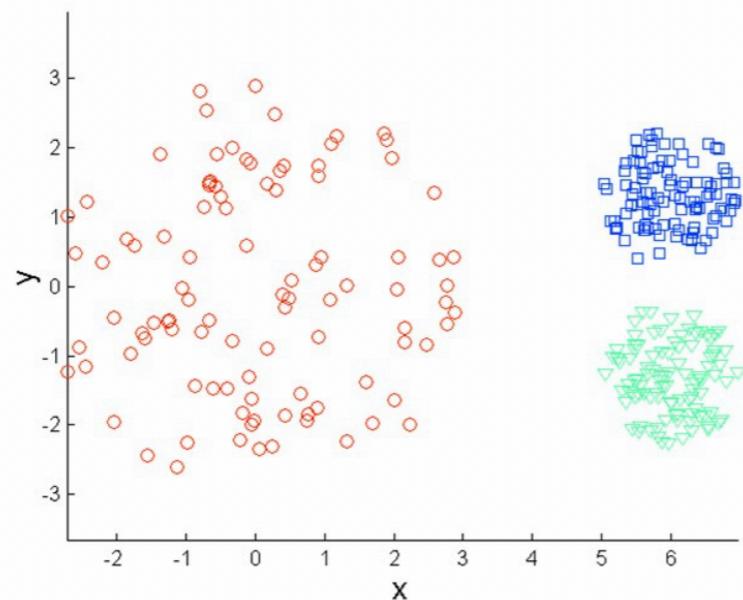
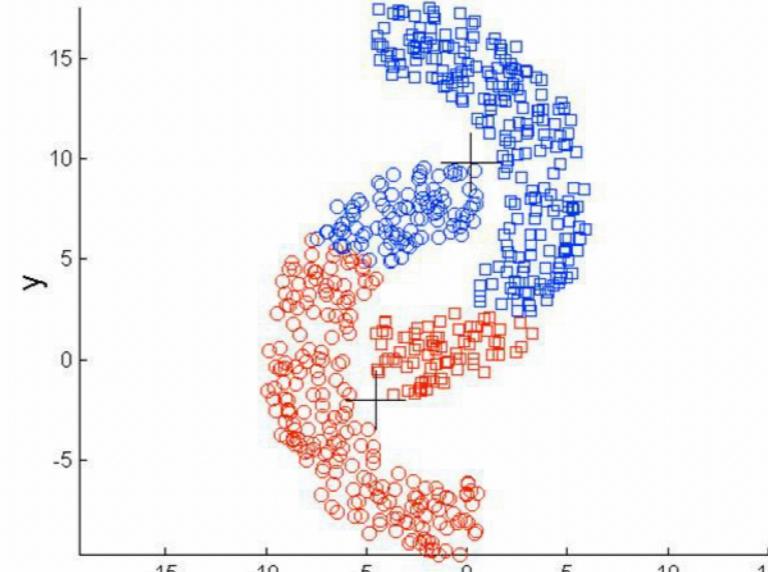
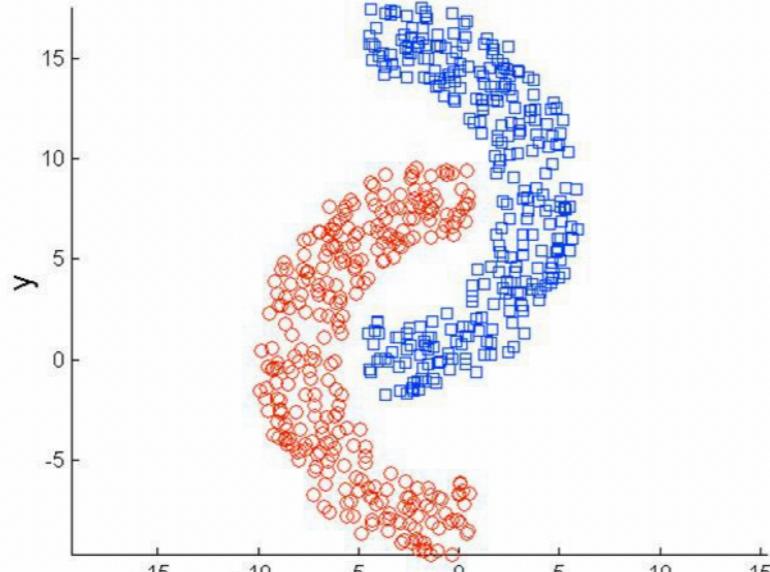
- ▶ K-평균 알고리즘은 각 단계마다 중점을 변경해가면 여러 번 반복
- ▶ 군집의 왜곡 (distortion)은 각 데이터에서 해당 중점까지의 거리의 제곱을 모두 합한 값으로, 이를 이용할 경우 왜곡값이 가장 작은 군집이 제일 좋음
- ▶ 실행 시간 측면에서 보면 k-평균 알고리즘은 효율이 좋음 (비교적 실행속도가 빠름)



군집화 (Clustering)

k-평균 군집의 한계점

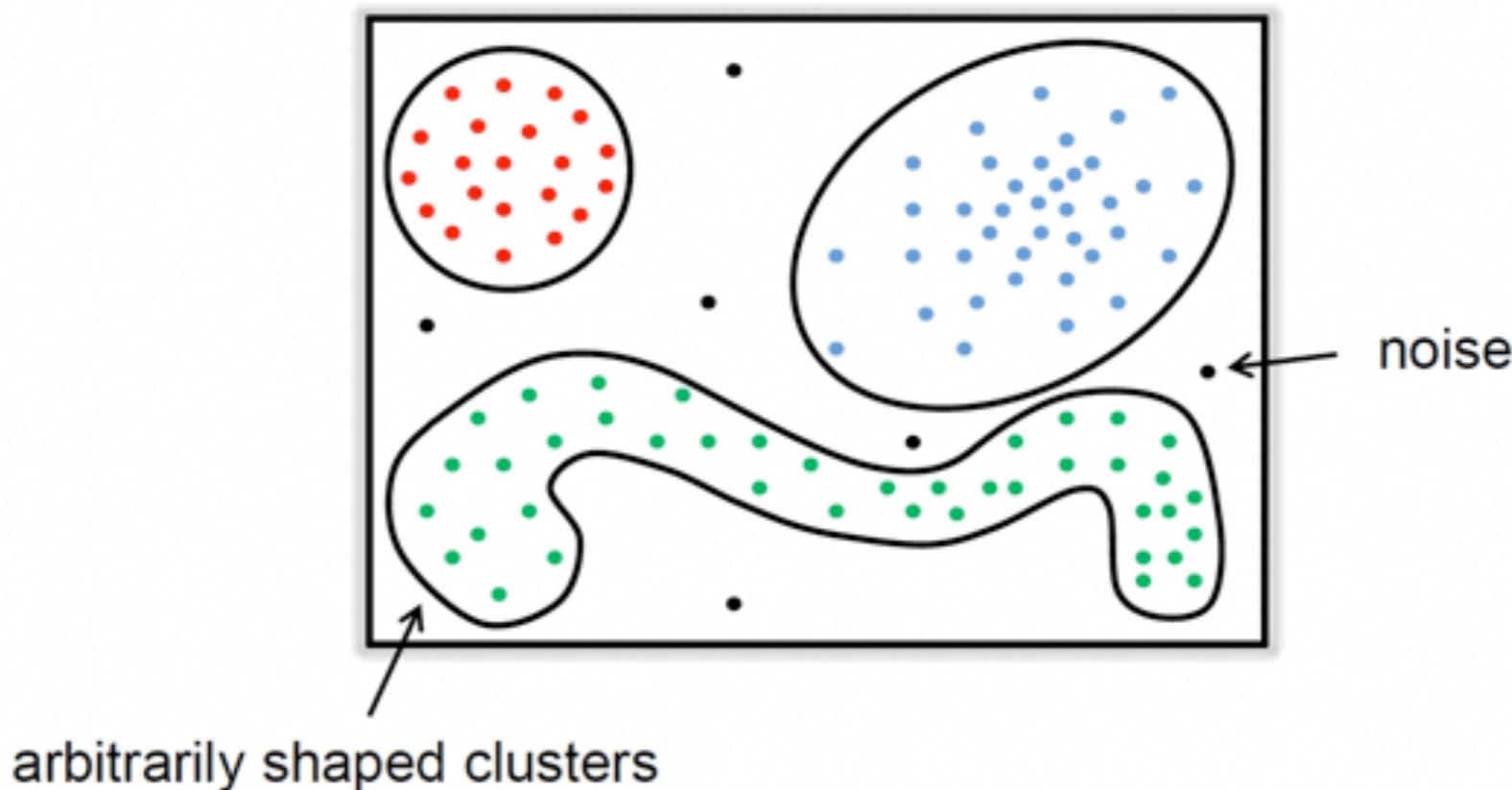
- ▶ 구형이 아닌 데이터 분포에 대해서 잘 구분할 수 있을까?
- ▶ 밀도가 다른 군집을 분할할 수 있을까?



군집화 (Clustering)

Density-Based Clustering (밀도 기반 클러스터링)

- ▶ 데이터의 분포와 밀도 (density)를 고려하여 클러스터를 구성하는 방법
- ▶ 구형이 아닌 임의의 모양으로 생긴 클러스터도 잘 찾을 수 있음
- ▶ 클러스터링 과정에서 노이즈를 제거하는 것이 가능함

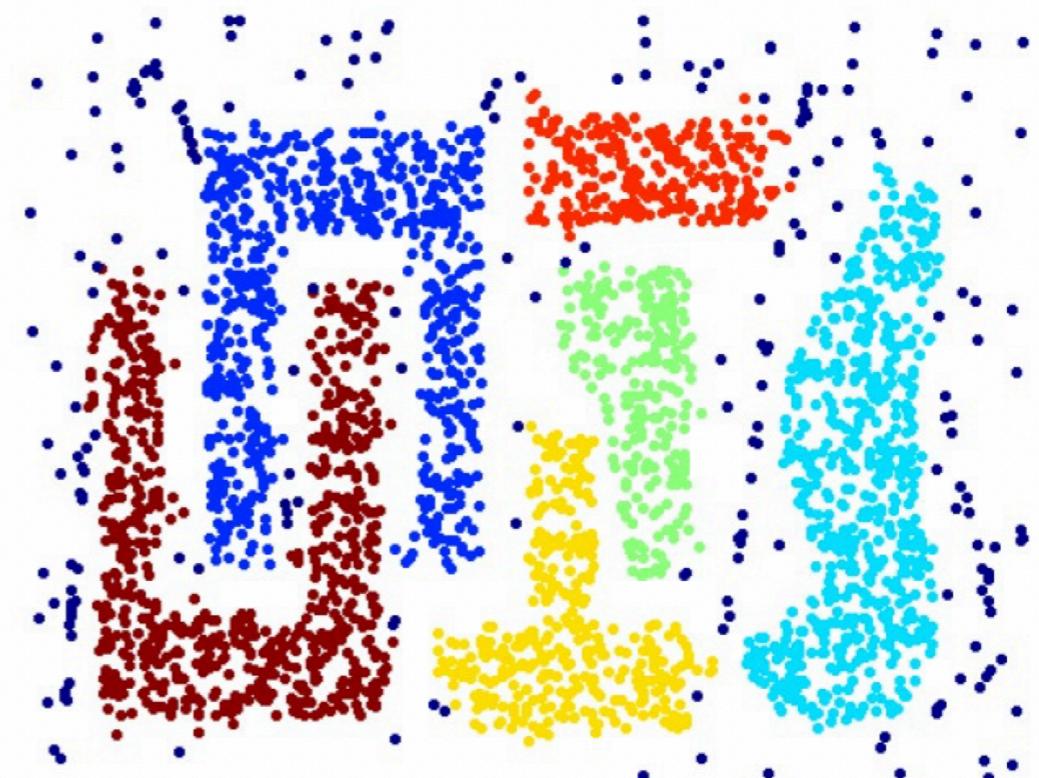


군집화 (Clustering)

Density-Based Spatial Clustering of Applications with Noise
(DBSCAN)



Original Points

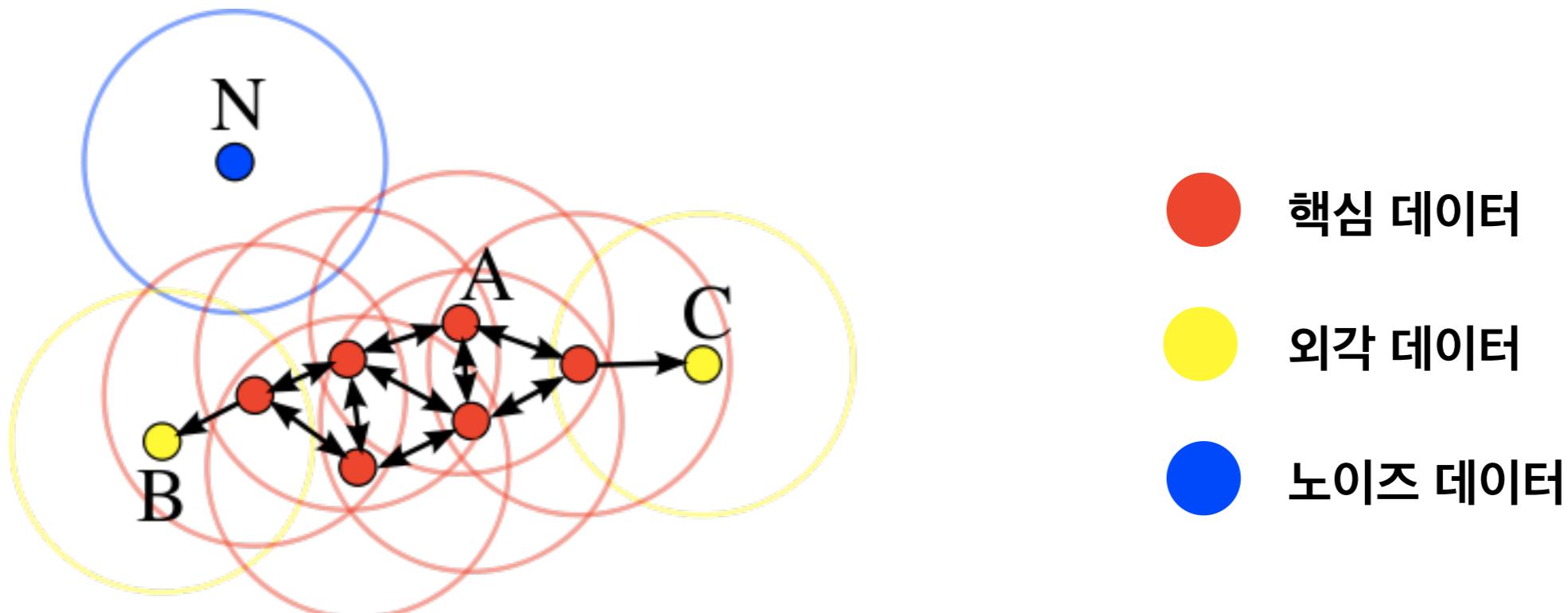


Clusters

군집화 (Clustering)

밀도 기반 클러스터링 (DBSCAN)

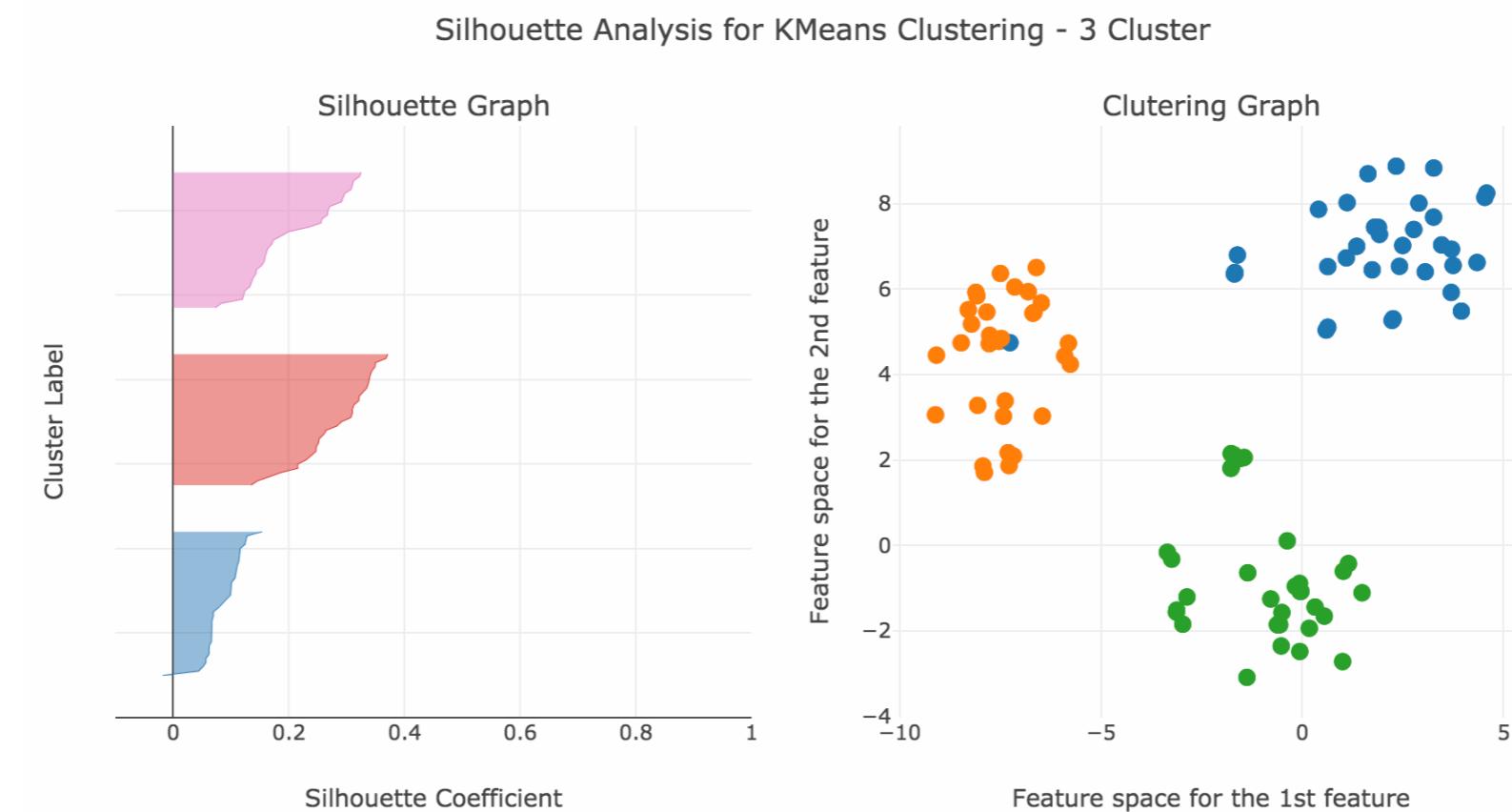
- ▶ 핵심 데이터 : 임의의 점 p에서 부터 거리 e 이하의 점이 m(=3)개 이상
- ▶ 이웃 데이터 : 임의의 점 p와의 거리가 e 이하인 데이터를 p의 이웃 데이터라고 한다.
- ▶ 외각 데이터 : 핵심 데이터는 아니지만, 임의의 핵심 데이터의 이웃 데이터인 점
- ▶ 노이즈 데이터 : 핵심 데이터도 아니고 외각데이터도 아닌 점



군집의 평가 (Validation)

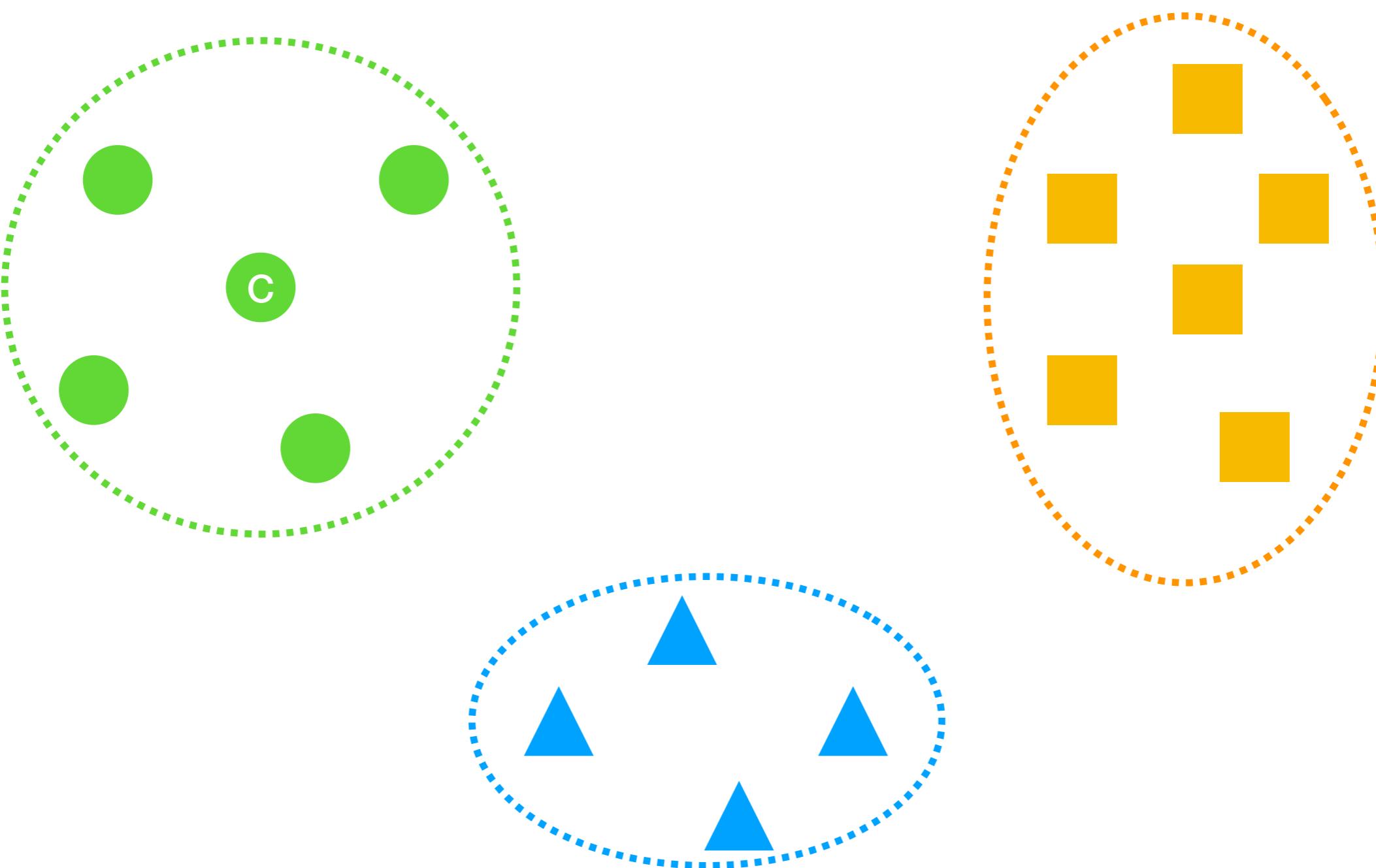
Silhouette Score

- ▶ 클러스터 내의 일관성과 유효성을 검증하는 척도 중 하나로, 각 객체가 얼마나 잘 분류되었는가를 판단하고 시각화하기 위해 활용됨
- ▶ 다른 클러스터와 비교하여 객체가 자체 클러스터와 얼마나 비슷한지 (cohesion) 또는 분리되어 있는지 (separation)에 대한 척도



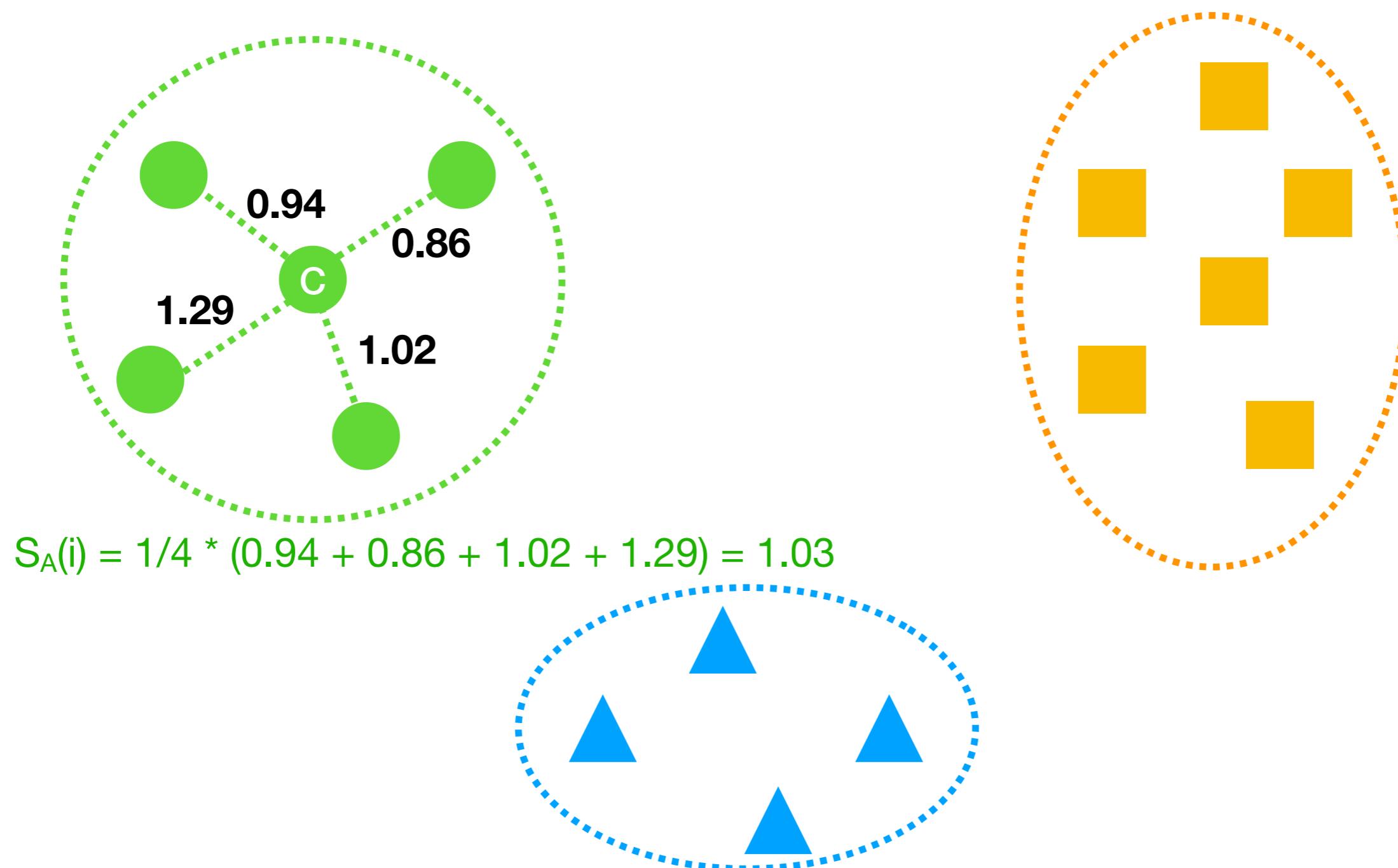
군집의 평가 (Validation)

Silhouette Score



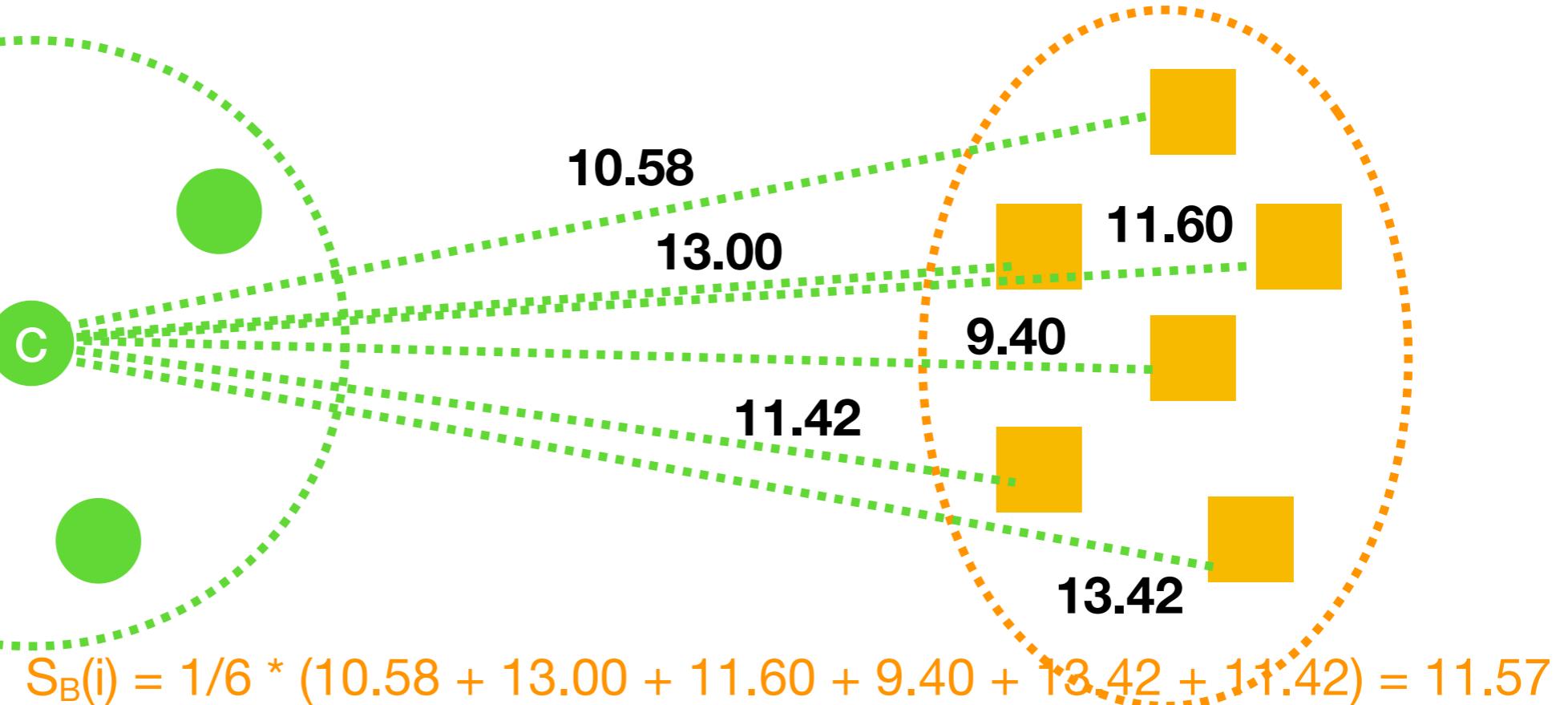
군집의 평가 (Validation)

Silhouette Score



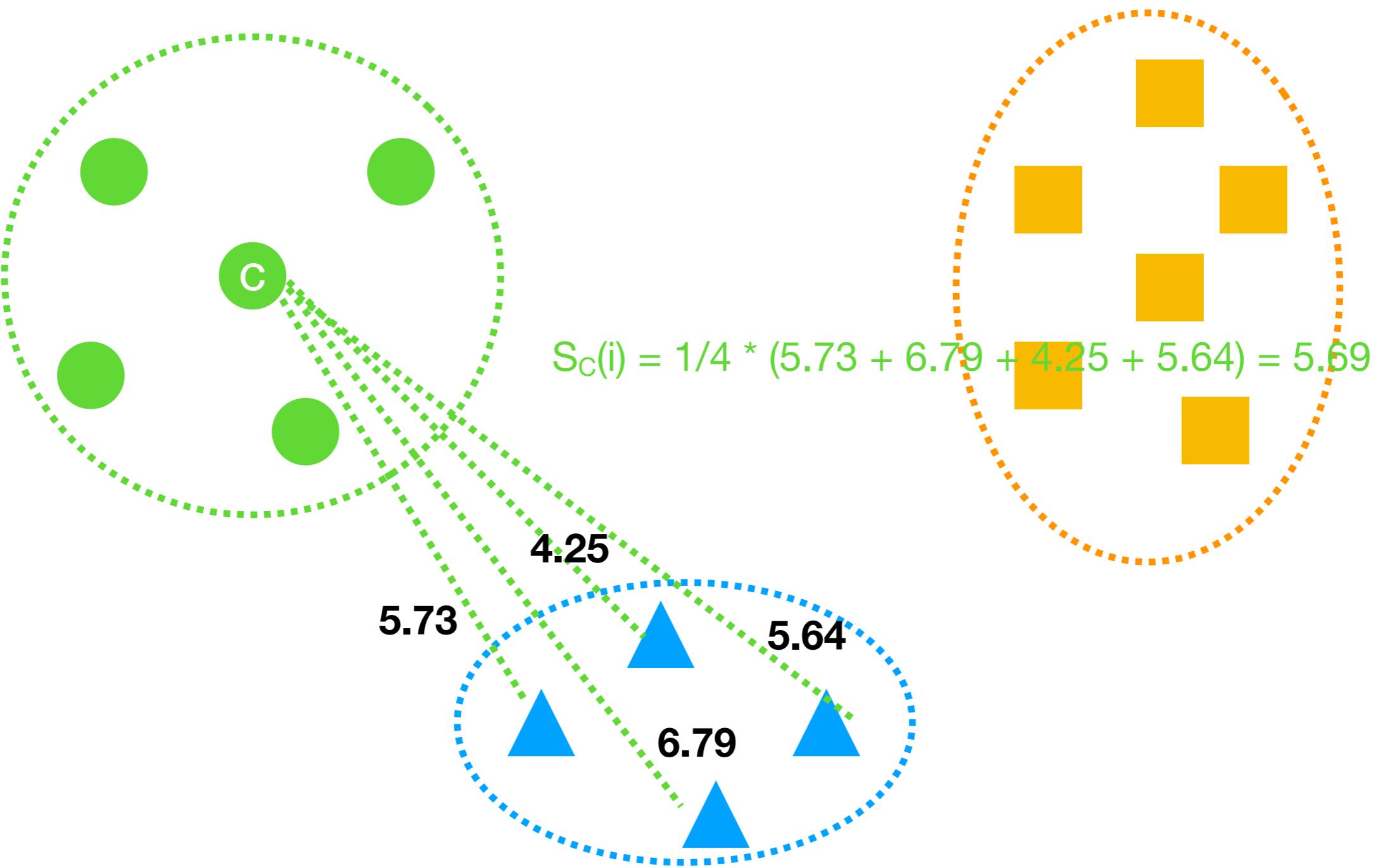
균집의 평가 (Validation)

Silhouette Score



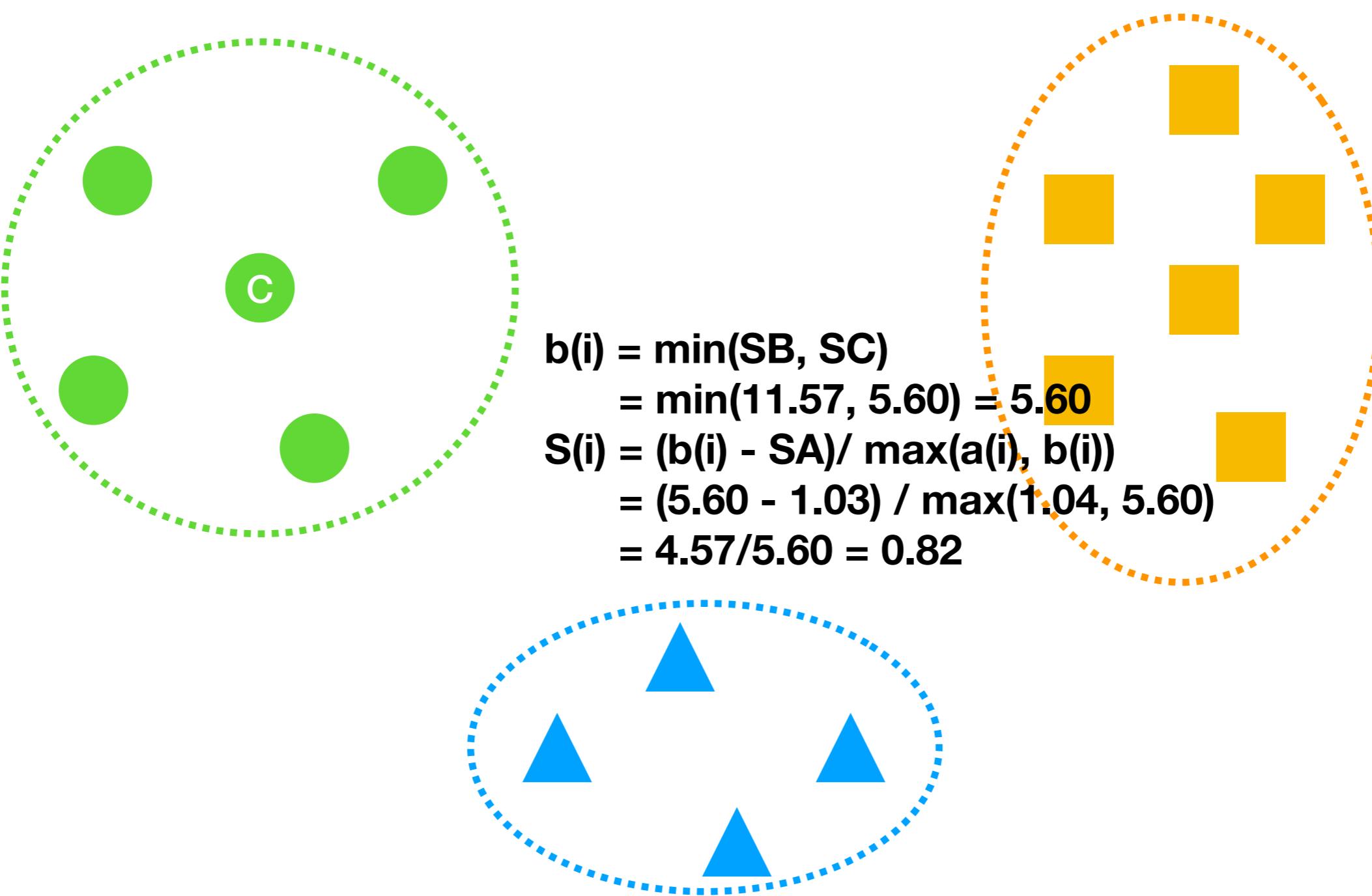
군집의 평가 (Validation)

Silhouette Score



군집의 평가 (Validation)

Silhouette Score



차원 축소

- Feature Selection
 - Pearson Correlation
 - Chi-Squared
 - Recursive Feature Elimination
 - Lasso: SelectFromModel
 - Tree-based: SelectFromModel
- Feature Projection (extraction)
 - Principal component analysis (PCA) : 주성분 분석
 - Linear discriminant analysis (LDA) : 선형판별 분석
 - Autoencoder
 - T-distributed Stochastic Neighbor Embedding (t-SNE)

E.O.D