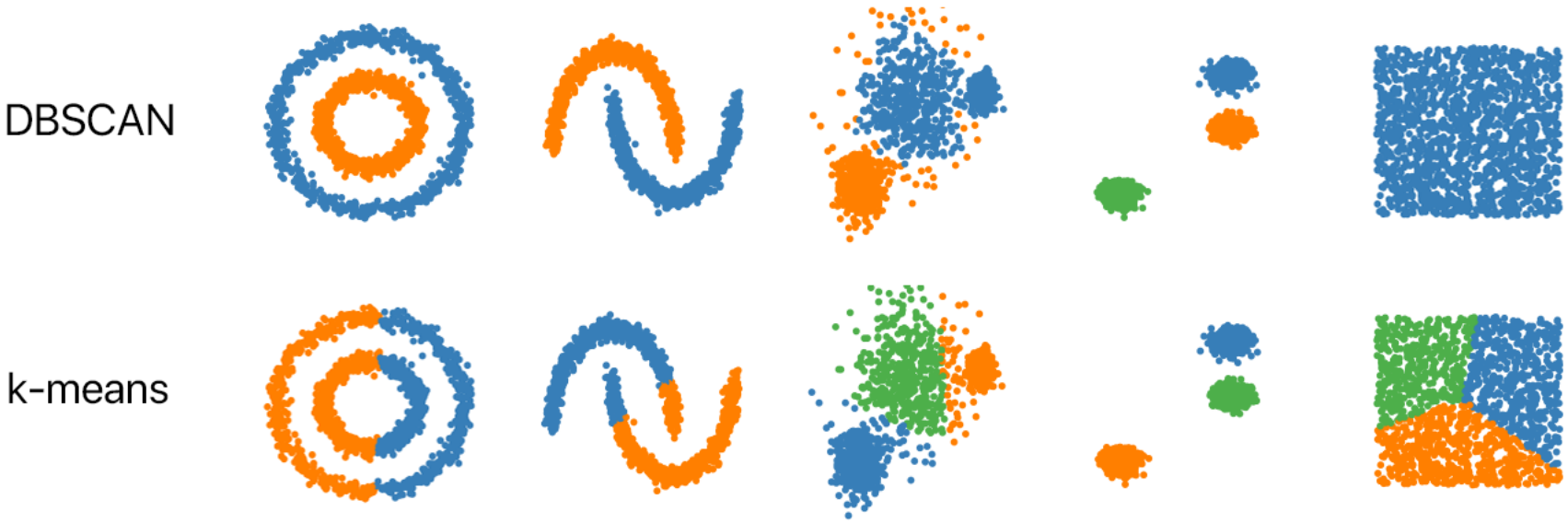
Week01 발표

DBSCAN & 가우시안 혼합 모형

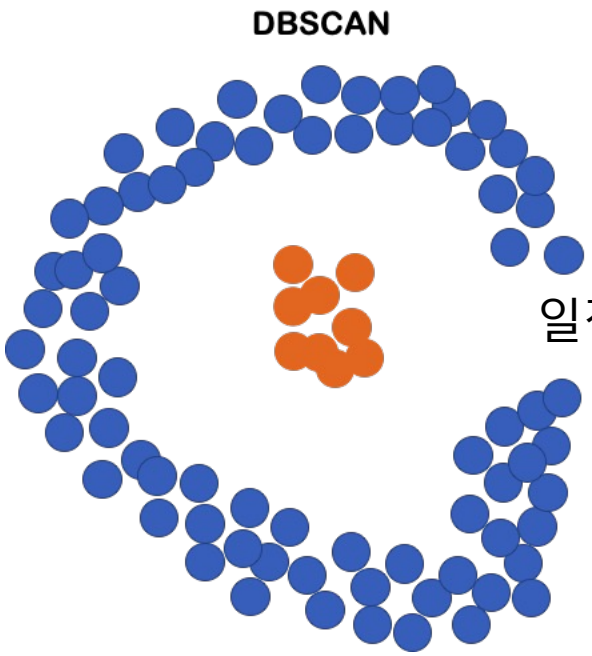
이경선

DBSCAN



Density-Based Spatial Clustering of Applications with Noise
데이터가 가장 밀집된 영역을 클러스터로 지정

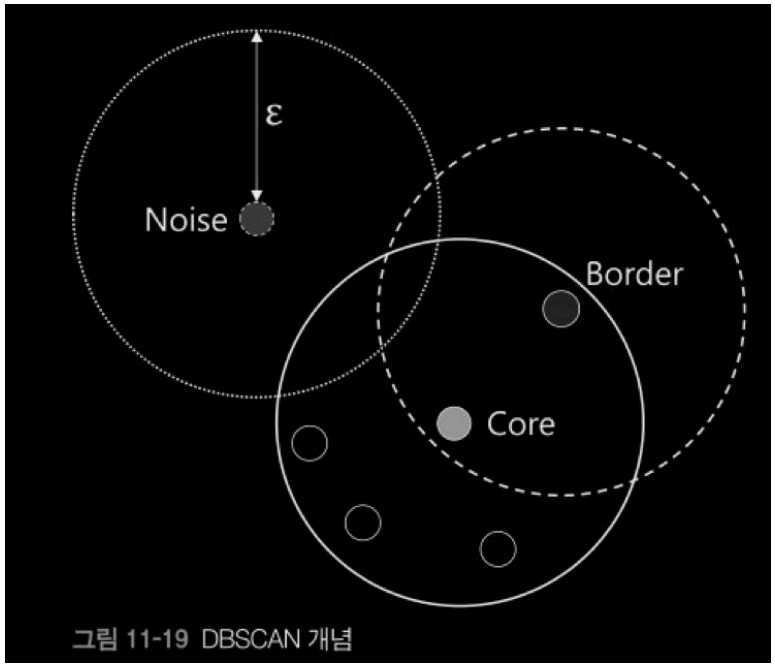
DBSCAN



일정 반경 내에 최소한의 데이터 포인트 개수 이상이 존재

Hyperparameter: min_samples, eps

DBSCAN



Noise

Eps 거리 내의 포인트
Min_sample 미만

Core

Eps 거리 내의 포인트
Min_sample 이상

Border

Core와 eps거리 내,
클러스터 지정x

DBSCAN

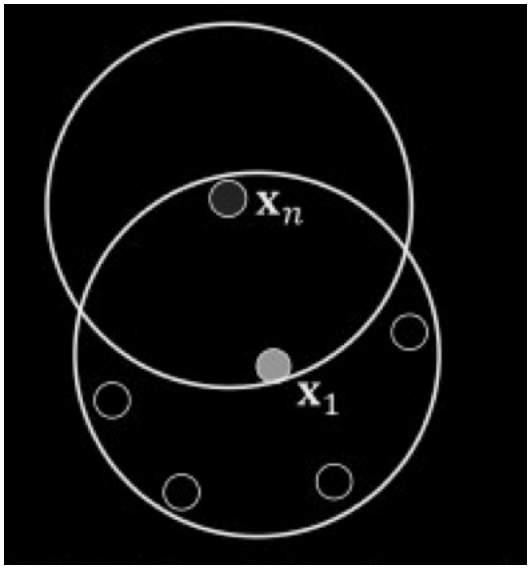
Eps-neighborhood

X1으로부터 eps내에
존재하는 Xi집합

$$N_{eps}(\mathbf{x}_1) = \{\mathbf{x}_i \in \boldsymbol{X} \mid dist(\mathbf{x}_1, \mathbf{x}_i) \leq eps\}$$

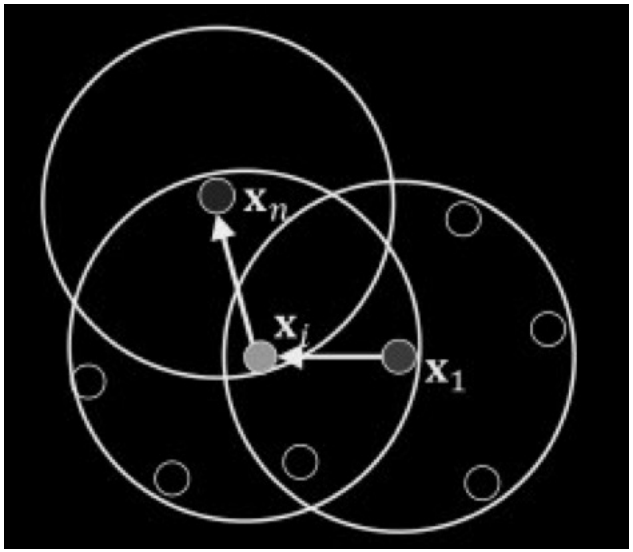
DBSCAN

Directly density-reachable



Xn d-d-r from X1

Density-reachable



Xn d-r from X1
클러스터에 속함

가우시안 혼합 모형

여러 개의 분포로부터 나온 데이터

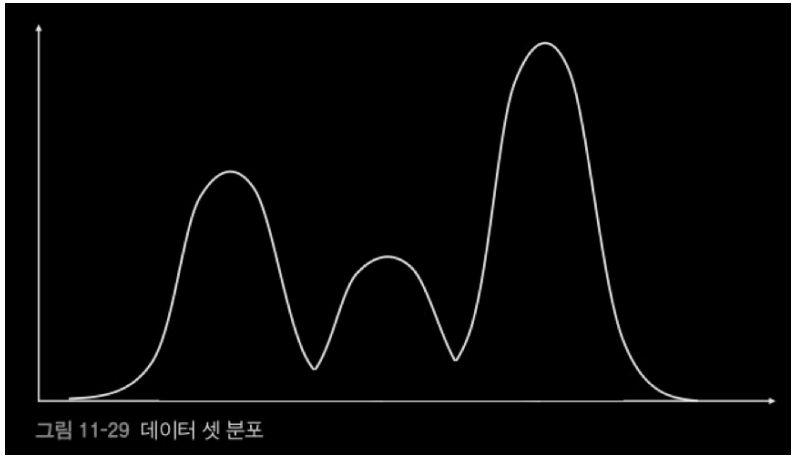


그림 11-29 데이터 셋 분포

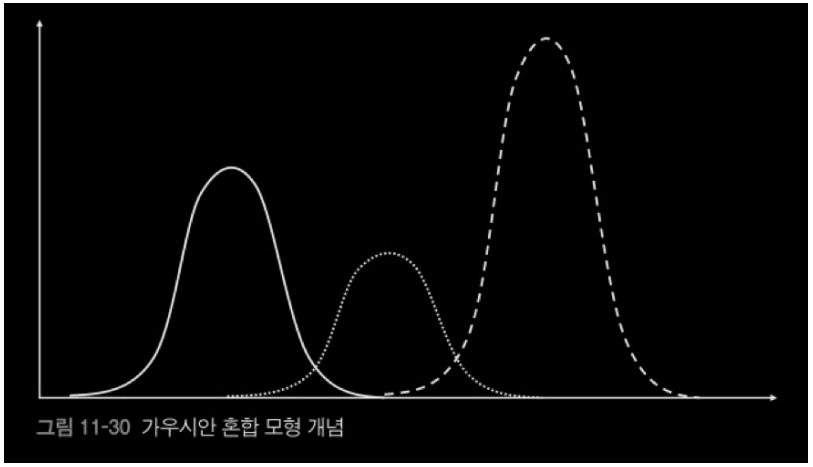


그림 11-30 가우시안 혼합 모형 개념

우리가 알아야할 것: 각 분포의 평균, 분산
Hyperparameter: 처음 평균, 분산, 분포 개수

가우시안 혼합 모형

■ EM 알고리즘

(1) 추정하려는 파라미터 μ_c, σ_c, π_c를 임의 방법으로 초기화합니다.

(2) E-step

$$r_{ic} = \frac{\pi_c N(x|\mu_c, \sigma_c)}{\sum_c \pi_c N(x|\mu_c, \sigma_c)}$$

(3) M-step

$$\pi_c = \frac{n_c}{n}$$

$$\hat{\mu}_c = \frac{1}{n_c} \sum_i r_{ic} x_i$$

$$\hat{\sigma}_c = \frac{1}{n_c} \sum_i r_{ic} (x_i - \hat{\mu}_c)^T (x_i - \hat{\mu}_c)$$

(4) 파라미터가 수렴할 때까지 (2)~(3)단계를 반복합니다.

$$r_{ic} = P(z = c|x)$$

$$= \frac{P(z = c, X = x)}{P(x)}$$

$$= \frac{P(x|z = c)P(z = c)}{\sum_c \pi_c N(x|\mu_c, \sigma_c)}$$

$$= \frac{\pi_c N(x|\mu_c, \sigma_c)}{\sum_c \pi_c N(x|\mu_c, \sigma_c)}$$

Ric: i번째 데이터가 그룹 c에서 추출되었을 확률