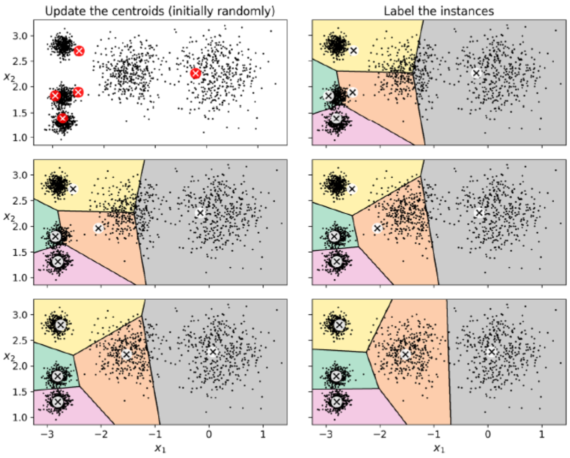
Unsupervised learning

Clustering

- The goal is to group similar instances together into clusters
- A great tool for data analysis, customer segmentation, recommender systems
- EX) K-means, DBSCAN

K-means algorithm

- A simple algorithm capable of clustering kind of dataset very quickly and efficiently, often in just a few iterations.

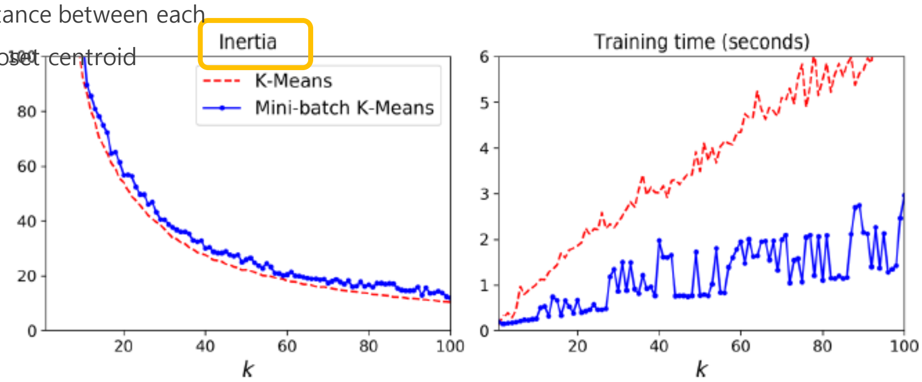


1. start by placing the centroids randomly.
2. Label the instances, update the centroids
3. so on util the centroids stop moving
4. Guaranteed to converge in a finite number of steps

Mini batch k-means algorithm

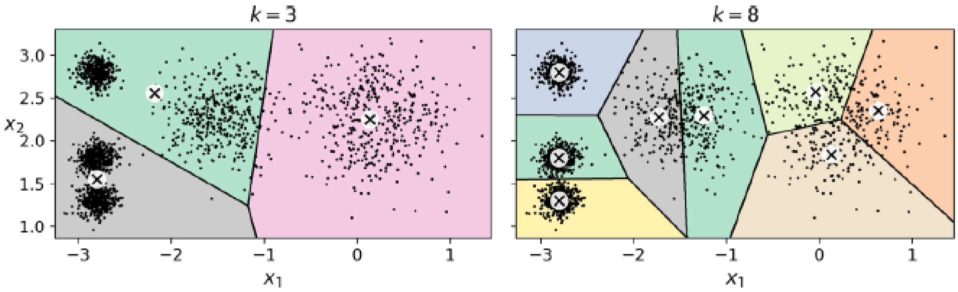
- Accelerating the algorithm by avoiding many unnecessary distance calculation
- Mini-batch algorithm is faster but slightly worse

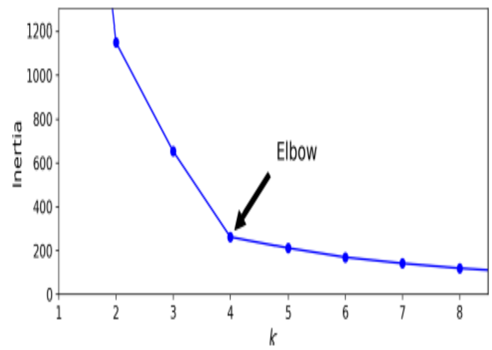
Mean squared distance between each instance and its closet centroid



Finding the optimal number of clusters

- The result might be quite bad if you set k to the wrong value





Silhouette score

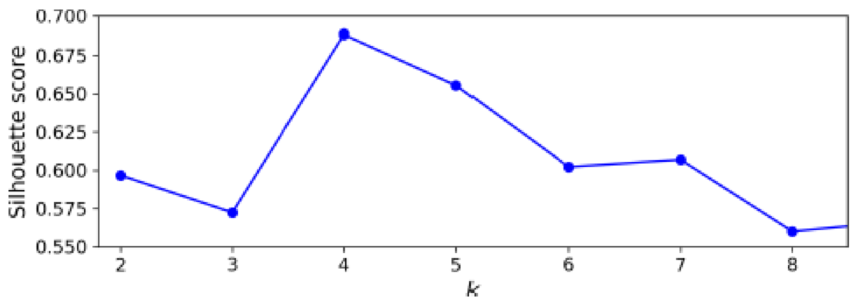
Silhouette coefficient = (b-a) / max(a,b)

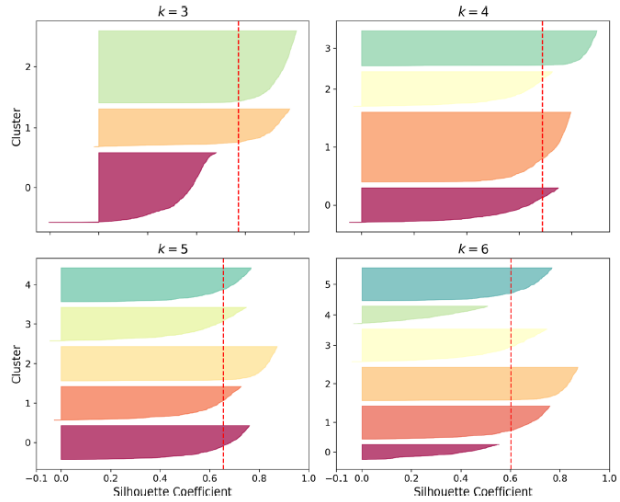
Close to -1: far from other clusters

Close to +1 : close to a cluster boundary

a = the mean distance to the other instances in the same cluster

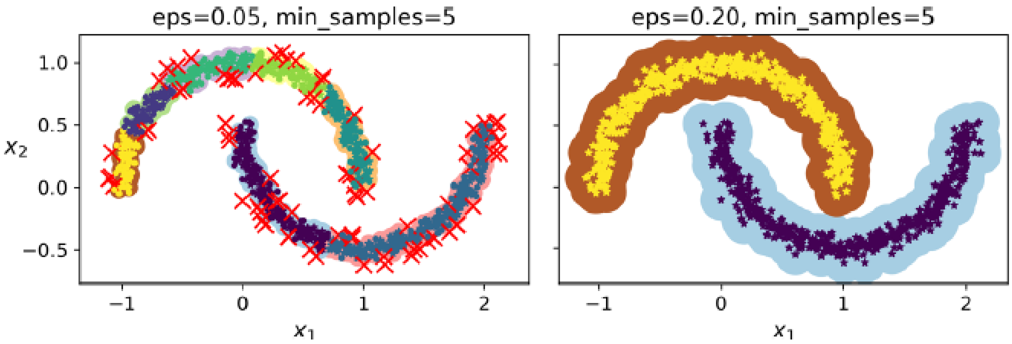
B = the mean nearest-cluster distance





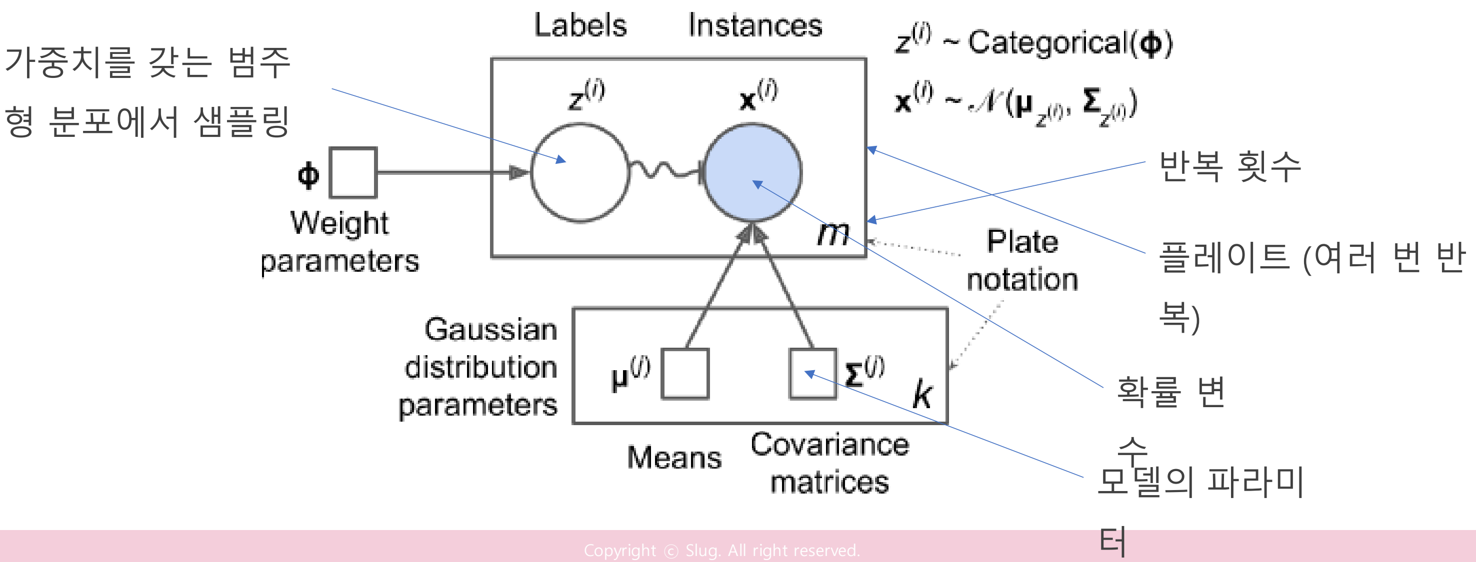
DBSCAN

- Clusters as continuous regions of high density
- Counts how many instances are located within a small distance epsilon from it
- Then it is considered a *core instance*. (in dense regions)
- If instance is not a core instance then, it is considered an anomaly.



Gaussian mixture model (가우시안 혼합 모델)

- A probabilistic model that assumes that the instance were generated from a mixture of several Gaussian distributions whose parameters are unkown



Expectation-maximization algorithm

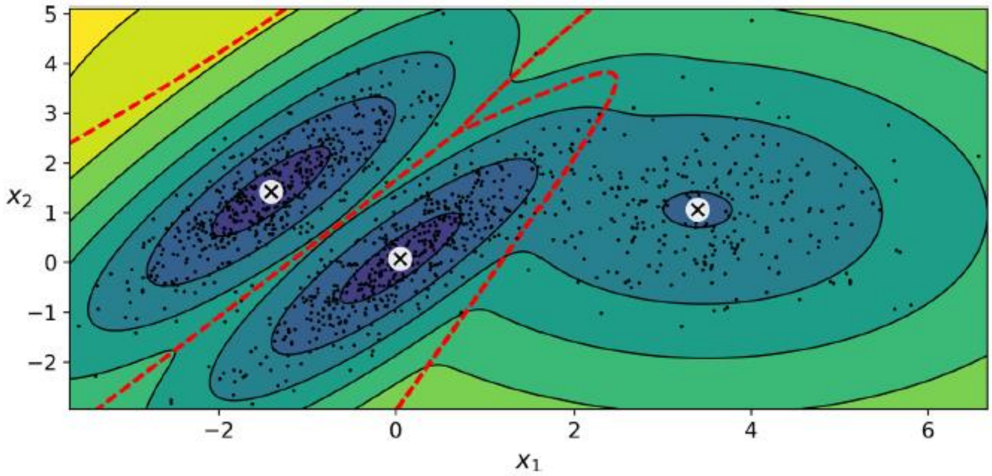
- A probabilistic model that assumes that the instance were generated from a mixture of several Gaussian distributions whose parameters are unknown

- Expectation step: assigning instance to clusters, estimate the probability(responsibility) that it belongs to each cluster

- Maximization step: update the clusters, updated using all the instances in the dataset, estimated probability that it belongs to that cluster.

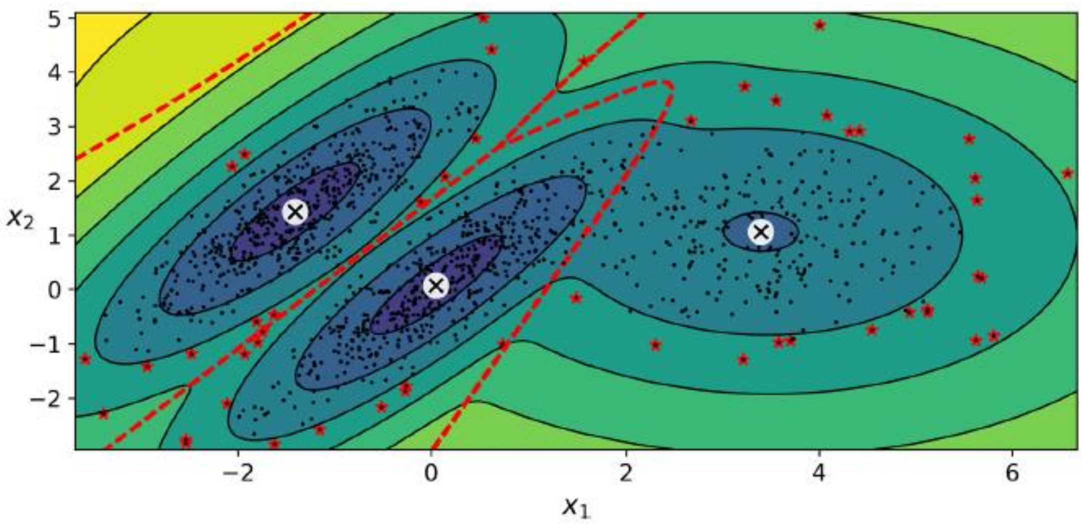
Gaussian mixture model (가우시안 혼합 모델)

- Estimate the density of the model: log of the probability density function(PDF)



Anomaly Dection using Gaussian mixture model (가우시안 혼합 모델)

- Set the density threshold

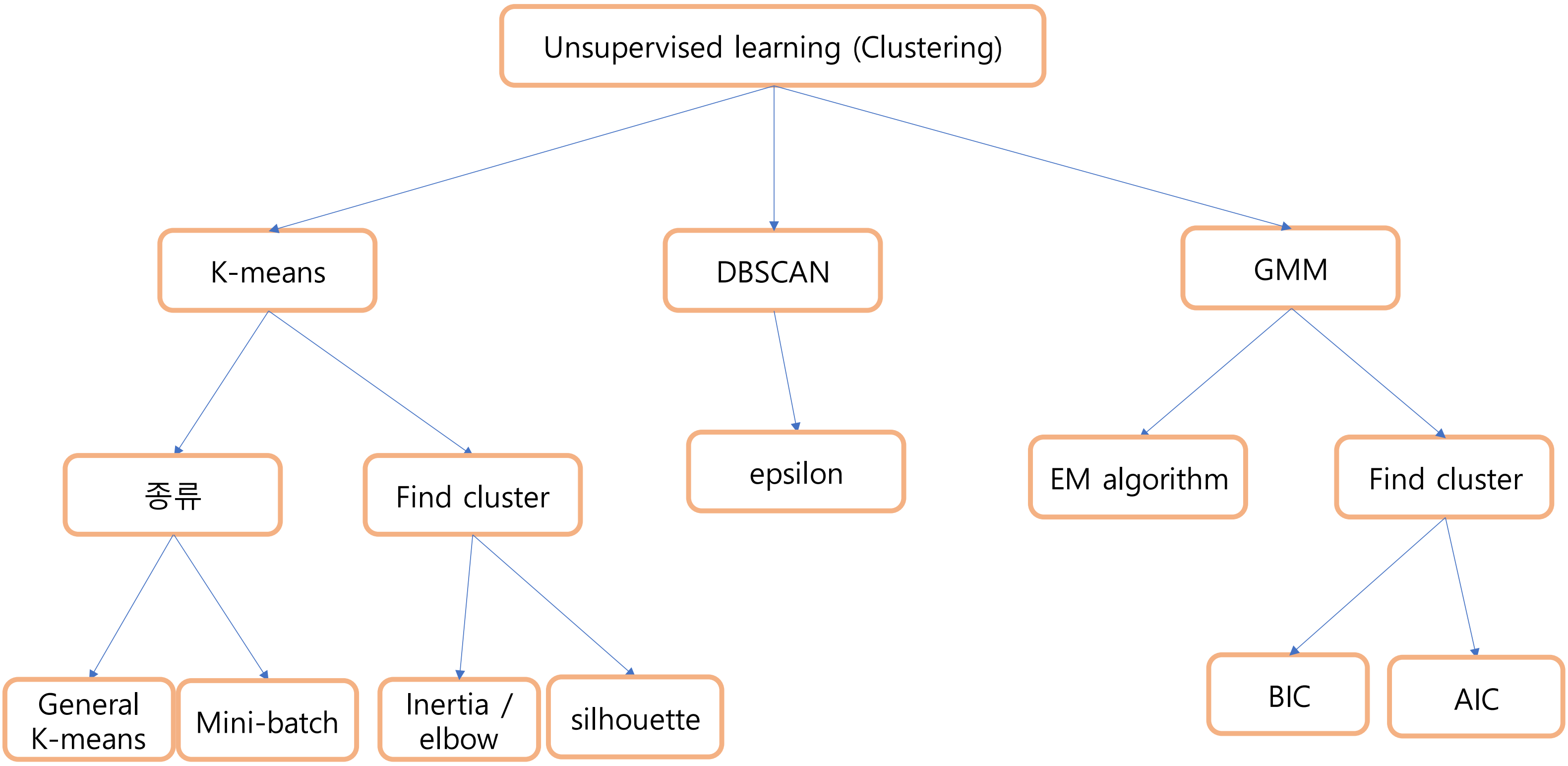


Finding the optimal number of clusters

- Not possible to use silhouette score because clusters are not spherical or have different sizes
- Minimize a *theoretical information criterion*
- Penalize models that have more parameters to learn
- reward models that fit the data well

$$BIC = \log(m)p - 2\log(\hat{L})$$
$$AIC = 2p - 2\log(\hat{L})$$

- m is the number of instances, as always.
- p is the number of parameters learned by the model.
- $\hat{L}$ is the maximized value of the *likelihood function* of the model.


Unsupervised learning (Clustering)
K-means
DBSCAN
GMM
종류
Find cluster
epsilon
EM algorithm
Find cluster
General K-means
Mini-batch
Inertia / elbow
silhouette
BIC
AIC