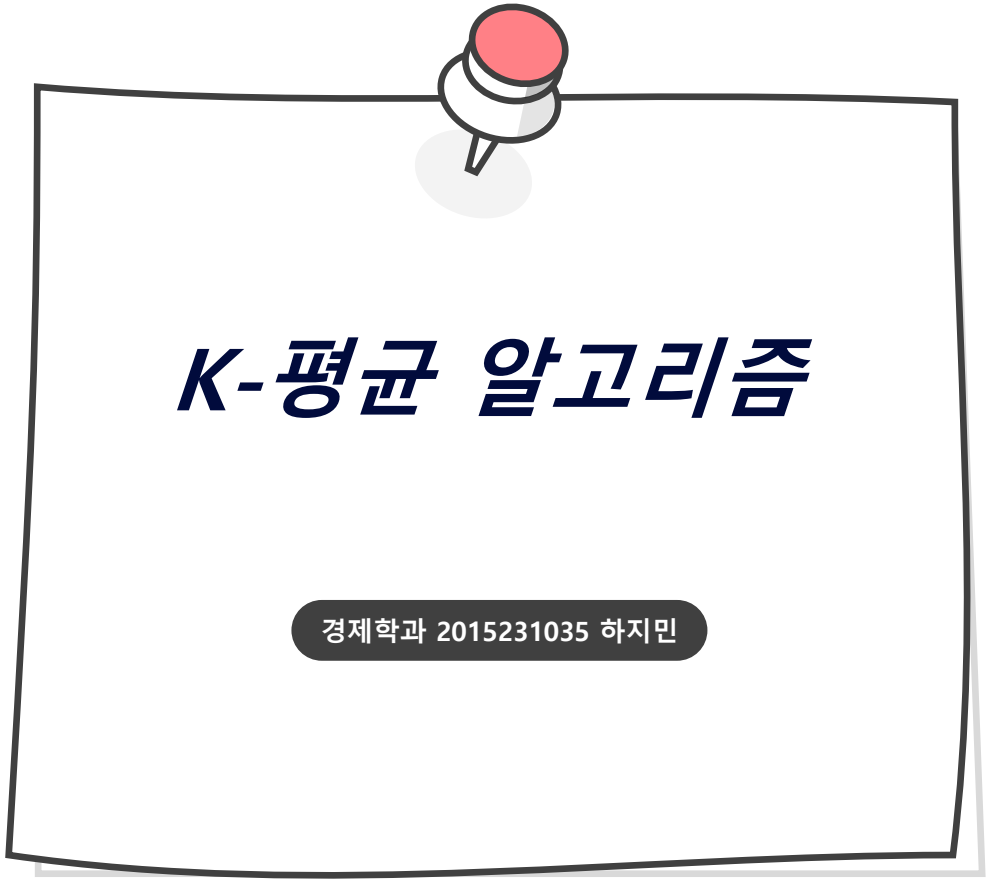


K-평균 알고리즘

경제학과 2015231035 하지민

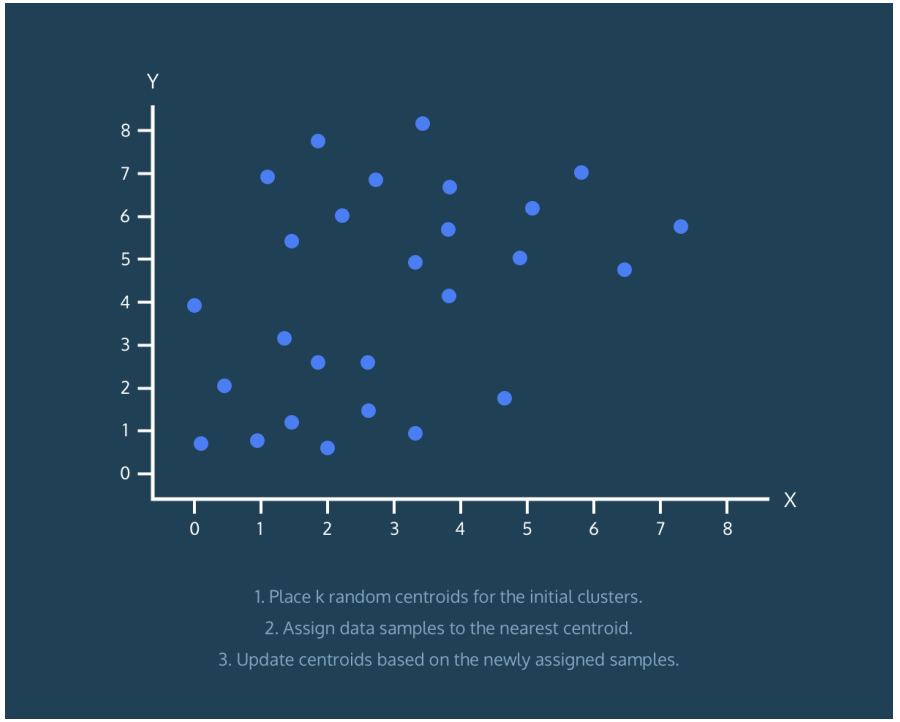
K-평균 알고리즘 소개(K-Means Clustering)

가장 유명한 클러스터링 알고리즘

“K”는 주어진 데이터로부터 그룹화 할 그룹, 즉 클러스터의 수

“Means“는 각 클러스터의 중심과 데이터들의 평균 거리를 의미

클러스터의 중심을 센트로이드(centroids)라고 부름



1. 데이터셋에서 K개의 센트로이드를 임의로 지정
2. 각 데이터들을 가장 가까운 센트로이드가 속한 그룹에 할당
3. 2번 과정에서 할당된 결과를 바탕으로 센트로이드를 새롭게 지정
4. 센트로이드에 변화가 없을 때 까지 2번으로 돌아가 반복

최적의 K

K-평균 알고리즘의 단점 중 하나는 클러스터 개수를 사전에 지정해야 한다는 것!

=> **엘보우(elbow) 방법**

이너셔(inertia) :
클러스터 중심과 클러스터에 속한 샘플 사이의 거리의 제곱 합
각 데이터로부터 자신이 속한 중심까지의 거리를 의미

이너셔가 낮을수록 군집화가 더 잘 됐다고 봄

클러스터 개수를 증가시키면서 이너셔를 그래프로 그리면 감소하는 속도가 꺾이는 지점 존재
=>이 지점이 최적의 클러스터 개수