(K-means clustering)

2-18/07/08

- 특징변수의 특성이 유사한 데이터 그룹 ⇒ 클러스터 (cluster)
- 주어진 데이터를 여러개의 클러스터로 구분하는 것 ⇒ 클러스터링 (clustering)

K-means clustering

- 가장 간단하고 쉬운 clustering Algorithm 중에 하나.
- 목적함수 값이 최소화 될 때까지 클러스터의 중심(centroid) μ_K과 각 데이터가 소속될 cluster를 반복해서 찾는 것. 이 최종적인 목적함수 값을 inertia라고 한다.
- $J = \sum_{K=1}^{k} \sum_{i \in C_k} d(x_i, \mu_K)$ $\quad d(x_i, \mu_K) = \|x_i - \mu_K\|^2$
- 세부 Algorithm은 다음과 같다.
  1. 임의의 중심값 μ_K를 고른다.
  2. 중심에서 각 Sample data 까지의 거리를 계산.
  3. 각 데이터 샘플에서 가장 가까운 중심을 선택하여 클러스터 생성
  4. 새로 만들어진 클러스터에 대해 중심을 다시 계산하여 1~4회를 반복한다.

K-means++

- k-means algorithm은 최초의 중심값을 설정하기 위한 알고리즘.
  1. 중심값을 저장할 장소 M 준비.
  2. 일단 하나의 중심 μ_0를 선택하여 M에 넣는다.
  3. M에 속하지 않는 모든 샘플 x_i에 대해 거리 $d(M, x_i)$를 계산.
     $d(M, x_i)$는 M안의 모든 샘플 μ_K에 대해 $d(\mu_K, x_i)$를 계산하여 가장 작은 값 선택
  4. $d(M, x_i)$에 비례한 확률로 다음 중심 μ_1을 선택.
  5. K개의 중심을 선택할 때까지 반복
  6. k-means Algorithm 사용.