* **GMM- 거리기반 K-Means의 문제점**

**K-Means**는 **특정 중심점(각 군집의 중심점)을 기반으로 퍼져 있는 데이터 세트**에 군집화를 적용하면 효율적이지만 **이러한 데이터 분포를 가지지 않는 데이터 세트**에 대해서는 **비효율적임**.

* **GMM(Gaussian Mixture Model) 개요**

**GMM군집화** : **여러 개의 다른 가우시안분포(Gaussian Distribution)을 가지는 모델로 가정**하고

군집화 수행

* 데이터에서 여러 개의 정규 분포 곡선을 추출하고, 개별 데이터가 이 중 어떤 정규 분포에 해당되는지 결정하는 방식
* **GMM 모수(Parameter) 추정을 위한 EM(Expectation and Maximization)**

**Expectation :** 개별 데이터 각각에 대해 **특정 정규 분포에 소속될 확률**을 구하고

**가장 높은 확률을 가진 정규 분포에 소속** (최초에는 데이터들을 임의로 특정 정규 분포로 소속)

**Maximization :** 데이터들이 특정 정규분포로 소속되면 다시 해당 정규분포의 평균과 분산을 구함. **해당 데이터가 발견될 수 있는 가능도를 최대화**(Maximum likelihood)할 수 있도록 **평균과 분산(모수)를 구함**

* 개별 정규분포의 모수인 평균과 분산이 더 이상 변경되지 않고 각 개별 데이터들이 이전 정규 분포 소속이 더 이상 변경되지 않으면 최종 군집화를 결정하고 elseif 계속 **EM 반복**
* **사이킷런 GaussianMixture**

주요 파리미터 : n\_components : Mixture Model의 개수, 군집화 개수를 의미