


# ML\_07\_1. clustering

🕒 생성일	@2022년 6월 9일 오후 3:25
🏷️ 유형	머신러닝/딥러닝
👤 작성자	 동훈 오

클러스터링은 가우시안 혼합을 제외한 k-mean clustering 만 다룬다. 가우시안 혼합은 이론이 너무 어려움. 커널 관련 내용도 섞여 나와서 힘들.

내용설명은 첨부한 pptx와 ipynb 노트북으로 대부분 대체한다.

## 혼공머신: 6장 비지도 학습

비슷한 샘플끼리 그룹으로 모으는 작업을 군집(clustering) 이라고 한다. 군집은 대표적인 비지도 학습 방법 이다. 군집 알고리즘에서 만든 그룹을 클러스터라고 한다.

### k-평균 알고리즘

k-평균 알고리즘은 샘플들의 평균을 이용하여 클러스터를 형성한다. 이때 평균값은 클러스터의 중심에 위치하고 centroid 라고 표현된다. 알고리즘의 작동 방식은 다음과 같다.

- 무작위로 k 개의 클러스터 중심을 정한다.
- 각 샘플에서 가장 가까운 클러스터 중심을 찾아 해당 클러스터의 샘플로 지정한다.
- 클러스터에 속한 샘플의 평균값으로 클러스터 중심을 변경한다.
- 클러스터 중심이 더 이상 변하지 않을 때 까지 두 번째 과정으로 되돌아가며 반복한다.



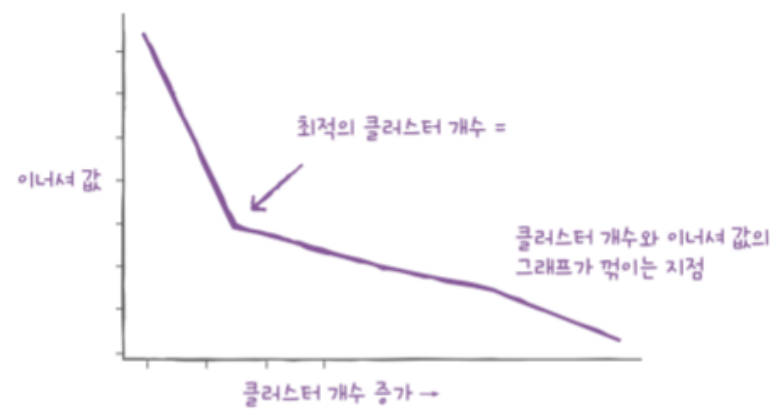
centroid 를 고정해 놓고 데이터를 마치 분류하듯 군집으로 묶는 방식이다.

### 최적의 k 찾기

k-평균 알고리즘의 단점 중 하나는 클러스터 개수를 사전에 지정해야 한다는 것이다. 실전에서는 몇 개의 클러스터가 있는지 알 수 없다. 어떻게 하면 적절한 k 값을 찾을 수 있을까.

패쵸거인 방법으로는 elbow 방법이 있다. k-평균 알고리즘은 centroid 와 클러스터에 속한 샘플 사이의 거리를 측정할 수 있다. 이 거리의 제곱합을 inertia 라고 부른다. inertia 는 클러스터에 속한 샘플이 얼마나 가깝게 모여 있는지를 나타내는 값으로 생각할 수 있다. 일반적으로 클러스터 개수가 늘어나면 클러스터 각각의 크기는 줄어들기 때문에 inertia 도 줄어든다.

elbow 방법은 클러스터 개수를 늘려가면서 inertia 의 변화를 관찰하여 inertia 가 특이적으로 감소하는 지점을 찾는 것이 목적이며, 해당 지점 이후에는 클러스터 개수가 증가하더라도 inertia 가 크게 감소하지 않는다.



[https://s3-us-west-2.amazonaws.com/secure.notion-static.com/699ecc07-6219-4d13-933e-6d176493b744/%ED%98%BC%EA%B3%B5%EB%A8%B8%EC%8B%A0\\_6%EC%9E%A5\\_%EB%B0%9C%ED%91%9C%EC%9E%90%EB%A3%8C.pptx](https://s3-us-west-2.amazonaws.com/secure.notion-static.com/699ecc07-6219-4d13-933e-6d176493b744/%ED%98%BC%EA%B3%B5%EB%A8%B8%EC%8B%A0_6%EC%9E%A5_%EB%B0%9C%ED%91%9C%EC%9E%90%EB%A3%8C.pptx)

[https://s3-us-west-2.amazonaws.com/secure.notion-static.com/53b6b3b7-3089-4a9b-8c0b-a8bc21694338/Ch\\_9\\_%EB%B9%84%EC%A7%80%EB%8F%84\\_%ED%95%99%EC%8A%B5\\_%ED%81%B4%EB%9F%AC%EC%8A%A4%ED%84%B0\\_ipynb%EC%9D%98\\_%EC%82%AC%EB%B3%B8.ipynb](https://s3-us-west-2.amazonaws.com/secure.notion-static.com/53b6b3b7-3089-4a9b-8c0b-a8bc21694338/Ch_9_%EB%B9%84%EC%A7%80%EB%8F%84_%ED%95%99%EC%8A%B5_%ED%81%B4%EB%9F%AC%EC%8A%A4%ED%84%B0_ipynb%EC%9D%98_%EC%82%AC%EB%B3%B8.ipynb)