

# CLUSTERING

---

# Supervised / Unsupervised Learning

- Supervised Learning

지도 학습이라고 부르며, 데이터에 대한 Label이 주어진 상태에서 학습시키는 방법이다. classification 및 regression 으로 예측이 가능하다.

- Unsupervised Learning

비지도 학습이라고 부르며, 데이터에 대한 Label이 주어지지 않은 상태에서 학습시키는 방법이다. Clustering 알고리즘이 있다.

# K-means Clustering

- Clustering이란 특성이 비슷한 데이터끼리 묶어주는 머신러닝 기법이다.
- K-means Clustering은 다차원 입력 데이터에서 해당 데이터가 어떤 그룹에 속할지 군집하는 대표적인 알고리즘이다.
- 각 그룹은 하나의 중심점을 갖게 되며, 데이터는 가장 가까운 중심에 할당된다.

# K-means Clustering

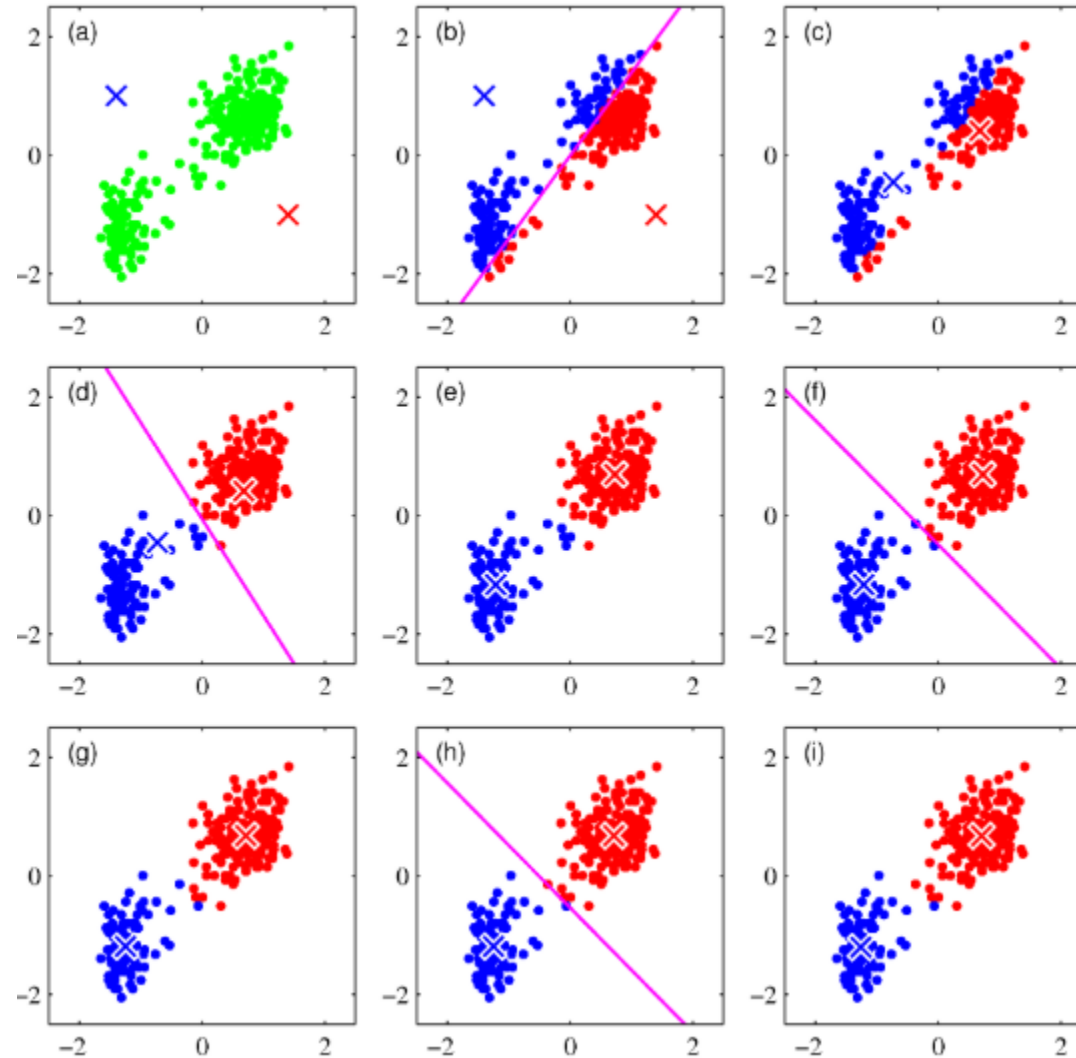
$$X = C_1 \cup C_2 \dots \cup C_K, \quad C_i \cap C_j = \phi$$

- 모든 데이터  $x$ 를  $k$ 개의 그룹으로 분류한다고 하였을 때,  $x$ 는  $k$ 개의 그룹 중 한 그룹에 속하게 된다. 그리고 그룹 간 겹치는 데이터는 존재하지 않는다.

$$\operatorname{argmin}_C \sum_{i=1}^K \sum_{x_j \in C_i} \|x_j - c_i\|^2$$

- 각 데이터는 중심점까지의 거리가 최소가 되는 그룹에 속하게 된다.

# K-means Clustering



# NEWS CLUSTERING

---

# News Data

## 대검 “서울중앙지검 등 특수부 3곳 남기고 특수부 폐지”

한겨레 · 1시간 전

- 대통령 연이은 경고에 윤석열 “서울중앙지검 등 3곳 빼고 전국 특수부 폐지”

뉴스플러스 · 1시간 전

 [전체 콘텐츠 보기](#)



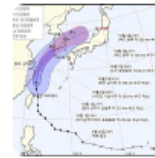
## 태풍 ‘미탁’ 한국 상륙 시점 앞당겨져...3일 0시께 전남 도착

매일경제 · 6시간 전

- [날씨] 이동 거리 줄어든 태풍 '미탁'...내일 자정 전남 해안 상륙 / YTN

 YTN NEWS · 4시간 전


 [전체 콘텐츠 보기](#)



## 의원 자녀 입시 전수조사 찬성한다더니...“입법사안” 한발 뺀 한국당

한겨레 · 5시간 전

- 고위공직자 자녀 특혜 전수조사, 나경원 “차관급 이상으로 확대”

 Nocut V CBS · 어제

 [전체 콘텐츠 보기](#)



뉴스 제목, 날짜, 내용, 키워드, 조회수 등이 나와있는 데이터로

뉴스 내용을 기반으로 비슷한 뉴스끼리 그룹화하여

뉴스 게시물 아래에 비슷한 뉴스를 보여주거나, 카테고리 분류하는데 사용된다.

# BOW(Bag of Words)

	it	is	puppy	cat	pen	a	this
it is a puppy	1	1	1	0	0	1	0
it is a kitten	1	1	0	0	0	1	0
it is a cat	1	1	0	1	0	1	0
that is a dog and this is a pen	0	2	0	0	1	2	1
it is a matrix	1	1	0	0	0	1	0

열은 단어를 나타내고, 행은 문서를 나타냈을 때

해당 단어가 문서내에 n개 있으면 있으면 n, 없으면 0으로 나타낸다.

(문서를 숫자 벡터로 변환하는 가장 기본적인 방법이다.)



# News Clustering

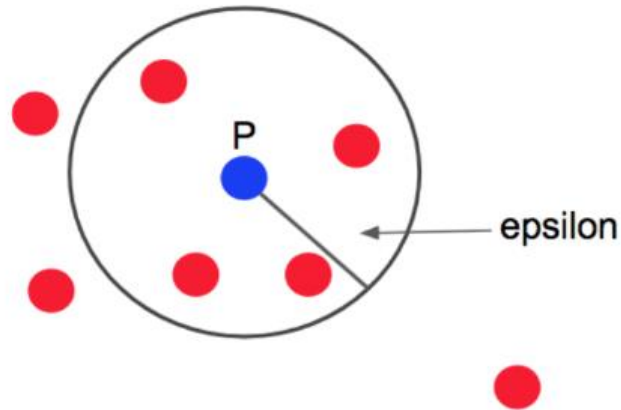
- BOW로 인코딩된 뉴스 기사를 하나의 좌표로 보고 Clustering할 수 있다.
- Scikit-learn은 회귀, 클러스터링, 차원축소, 모델 선택, 전처리 등의 다양한 알고리즘을 제공하는 파이썬 프레임워크다.
- Scikit-learn에서는 문서를 BOW로 인코딩하는 CountVectorizer 함수와 K-means Clustering 알고리즘을 제공한다.

# DBSCAN

---

# DBSCAN

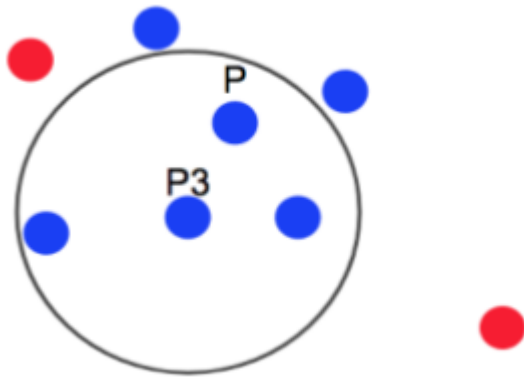
- 밀도 기반 클러스터링 방법으로 데이터가 세밀하게 몰려 있어 밀도가 높은 부분을 클러스터링 하는 방식이다.



- 점 P를 기준으로 epsilon(거리) 내에 있는 점이 n개 이상 있으면 하나의 군집으로 판단한다.

# DBSCAN

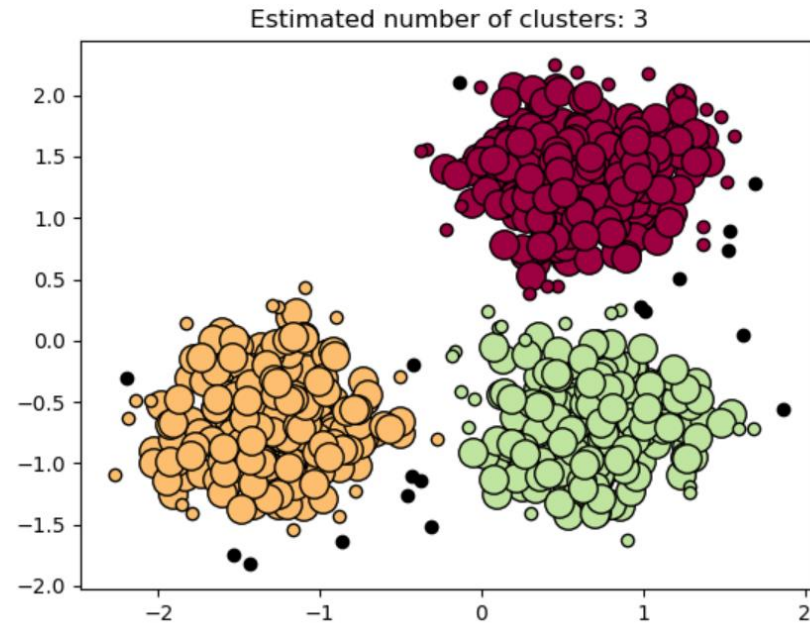
- epsilon(거리) 내에 있는 점이 4개 이상 있으면 하나의 군집으로 판단한다고 했을 때 아래와 같이  $P_3$ 가  $P$ 의 군집에 포함되어있으면  $P$ 의 군집과 하나의 군집으로 묶인다.



# DBSCAN

## Demo of DBSCAN clustering algorithm ¶

Finds core samples of high density and expands clusters from them.



Kmeans-Clustering에서 K값을 정하기 어려울 때 사용할 수 있다.

**LET'S DO IT**

---