D&A
ML Session 9차시

Word Embedding & Clustering

2022 / 11 / 22
D&A 운영진 나요셉

CONTENTS

01. INTRO

벡터 : 실제 세계를 수학적 공간에서 표현하기 위한 도구

벡터의 내적 : 벡터 간의 유사도를 측정하는 도구 **(두 벡터는 얼마나 닮았는가?)**

컴퓨터 사이언스에서의 벡터 : 위치 정보를 갖고 있는 데이터 포인트 **(ex : X = [1.2,4.6,2.9,10])**

Machine Learning 의 학습, **Deep Learing** 의 학습 방식은?

Clustering : 데이터를 분류하기 위한 공간을 나누는 작업 **(공간을 찾아가는 것 !)**

어떻게 분류하는 것이 효율적인가?

- 데이터의 특성에 따라 다르다.**(ex:** 분자구조 설계를 위한 **Clustering,** 기하 특성을 고려한 **Clustering)**
- 문제를 정의하는 것에 따라 다르다.
- 데이터가 어떻게 클러스터링 될지 우리는 알지 못한다. **(비지도학습)**

02. WordEmbedding

컴퓨터는 단어를 어떻게 이해할까?

- 나는 머신러닝 공부를 합니다.

해당 문장을 컴퓨터가 인지하는 방식은?

세상의 단어가 4가지라면..

나는 = **[1,0,0,0]**

머신러닝 = **[0,1,0,0]**

공부를 = **[0,0,1,0]**

합니다 = **[0,0,0,1]**

각각의 단어가 의미를 갖는 이유는?
각 단어 간의 표현하기 어려운 관계가 있기 때문!

02. WordEmbedding

단어의 표현 방법

- 희소표현 (ex : One-hot Encoding)

강아지 : [0,0,0,0,0,1,0,0,0,0] 고양이 : [1,0,0,0,0,0,0,0,0,0]

- 밀집표현(ex : Word2Vec)

강아지 : [1.2,3.5,0.9] , 고양이 : [0.1,0.4,1.6]

두 표현 방법의 차이는?

02. WordEmbedding

문제의식

- 컴퓨터는 단어 각각의 뜻을 파악하지 못하더라도 단어들의 관계를 학습한다면 각 단어의 의미를 간접적으로 받아들일 수 있지 않을까?
- 왕 – 남자 = 여왕

이러한 관계를 잘 학습하기 위해 어떻게 설계해야 하는가?

- **Word2Vec** 을 이해하며 알아가자..

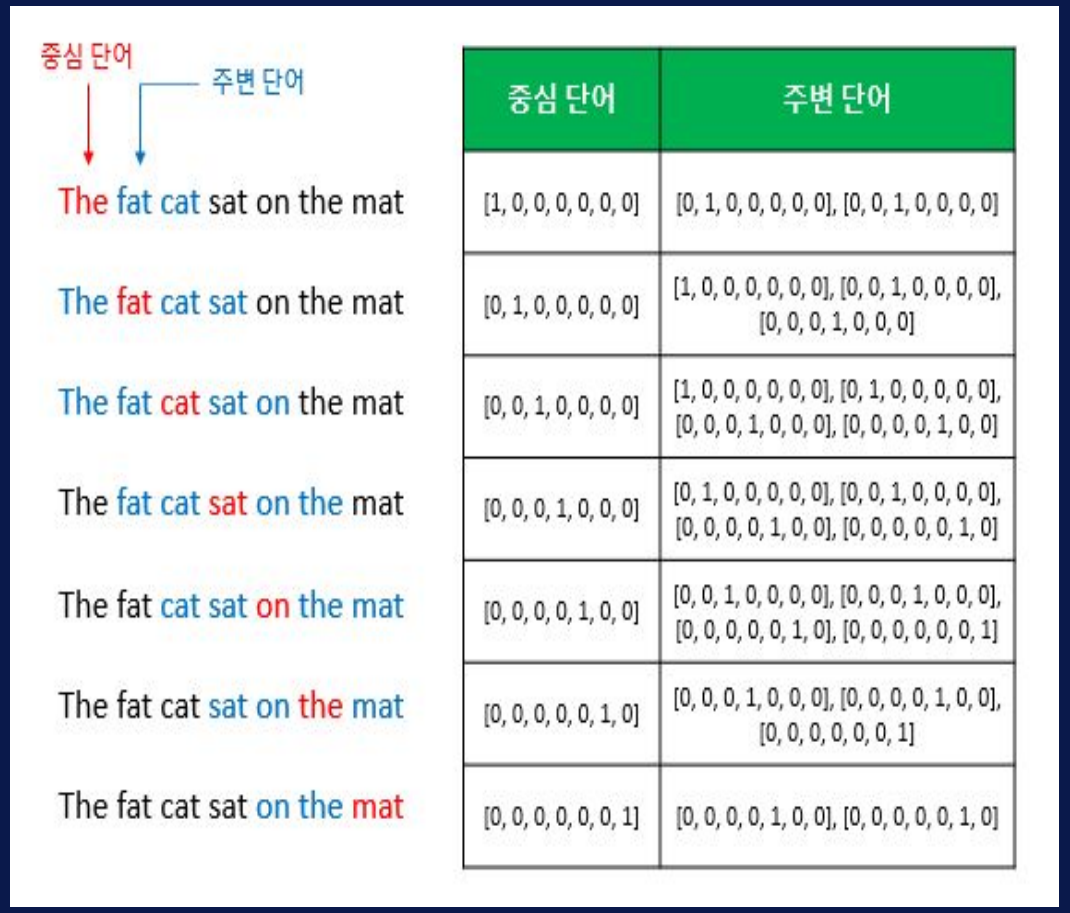
02. WordEmbedding

문제의식

- 컴퓨터는 단어 각각의 뜻을 파악하지 못하더라도 단어들의 관계를 학습한다면 각 단어의 의미를 간접적으로 받아들일 수 있지 않을까?
- 왕 – 남자 = 여왕

이러한 관계를 잘 학습하기 위해 어떻게 설계해야 하는가?

- **Word2Vec** 을 이해하며 알아가자..
- 희소표현을 잘 학습하여 밀집표현으로 바꾸기
- **C Bow, Skip gram**



중심 단어	주변 단어
[1, 0, 0, 0, 0, 0]	[0, 1, 0, 0, 0, 0], [0, 0, 1, 0, 0, 0]
[0, 1, 0, 0, 0, 0]	[1, 0, 0, 0, 0, 0], [0, 0, 1, 0, 0, 0], [0, 0, 0, 1, 0, 0]
[0, 0, 1, 0, 0, 0]	[1, 0, 0, 0, 0, 0], [0, 1, 0, 0, 0, 0], [0, 0, 0, 1, 0, 0], [0, 0, 0, 0, 1, 0]
[0, 0, 0, 1, 0, 0]	[0, 1, 0, 0, 0, 0], [0, 0, 1, 0, 0, 0], [0, 0, 0, 0, 1, 0], [0, 0, 0, 0, 0, 1]
[0, 0, 0, 0, 1, 0]	[0, 0, 1, 0, 0, 0], [0, 0, 0, 1, 0, 0], [0, 0, 0, 0, 1, 0], [0, 0, 0, 0, 0, 1]
[0, 0, 0, 0, 0, 1]	[0, 0, 0, 1, 0, 0], [0, 0, 0, 0, 1, 0], [0, 0, 0, 0, 0, 1]
[0, 0, 0, 0, 0, 1]	[0, 0, 0, 0, 1, 0], [0, 0, 0, 0, 0, 1, 0]

02. WordEmbedding

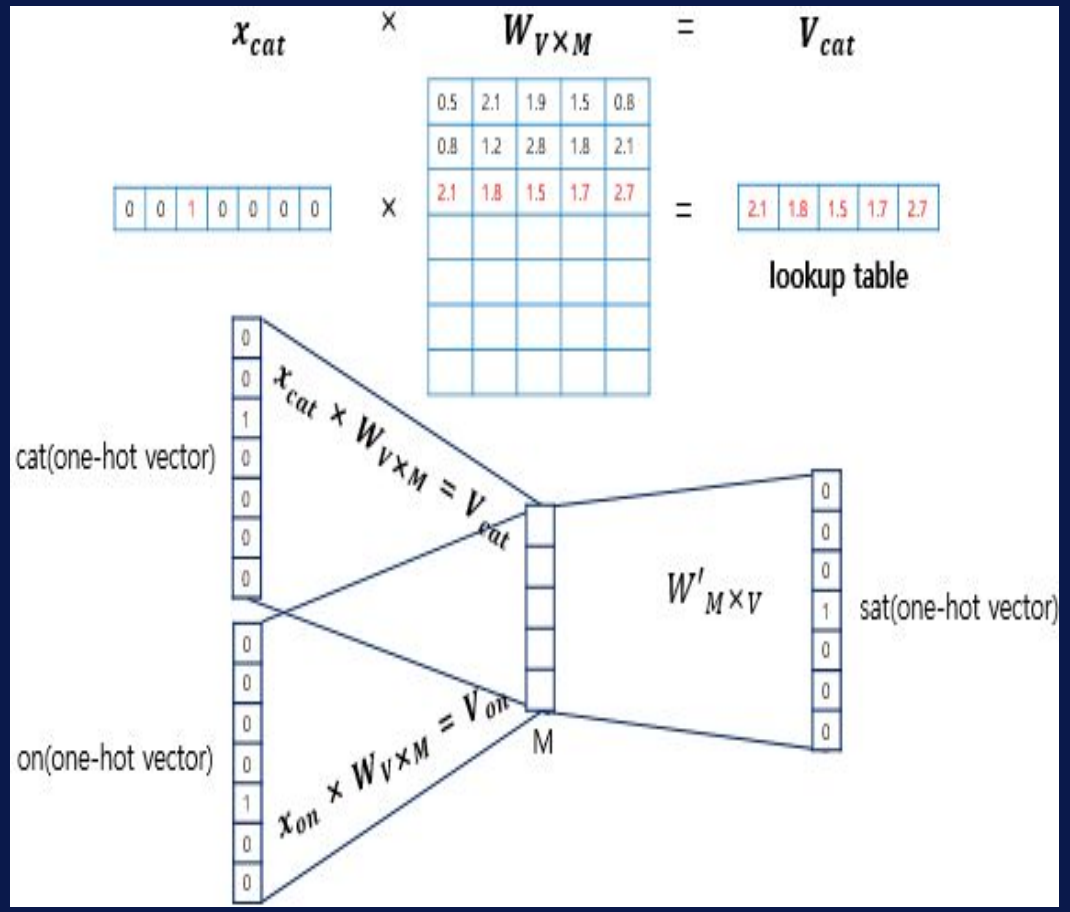
가정해봅시다

- 주변 단어로 중심 단어를 예측합시다.(Cat ??? On)
- 각 단어는 V 차원의 원-핫 벡터 (ex cat = [0,0,1,0,0,0,0]
- 가중치 행렬 : Look up table 과 같이 활성화 된 열의 정보만 담고 있음
- 단어 벡터 X 가중치 행렬 : 단어의 활성화된 곳의 가중치만을 출력

Flow

각 단어 벡터를 가중치 행렬의 차원으로 이동시킴 (V -> M으로 차원이동)

이후 또 다른 가중치 행렬을 (W') 통해 다시 차원을 이동시킴 (M -> V)



02. WordEmbedding

Flow..2

이후 나온 결과값 $\hat{y}$ 의 결과값을 SoftMax 함수 값을 구한다.

가장 확률이 높은 값을 1 나머지 값은 0으로 채운 희소 표현 벡터를 생성!

Ex: Cat ___ on -> Cat sat on

오답이라면 ?

해당 Loss 값을 줄이기 위해 **Back propagation** 을 진행합니다..

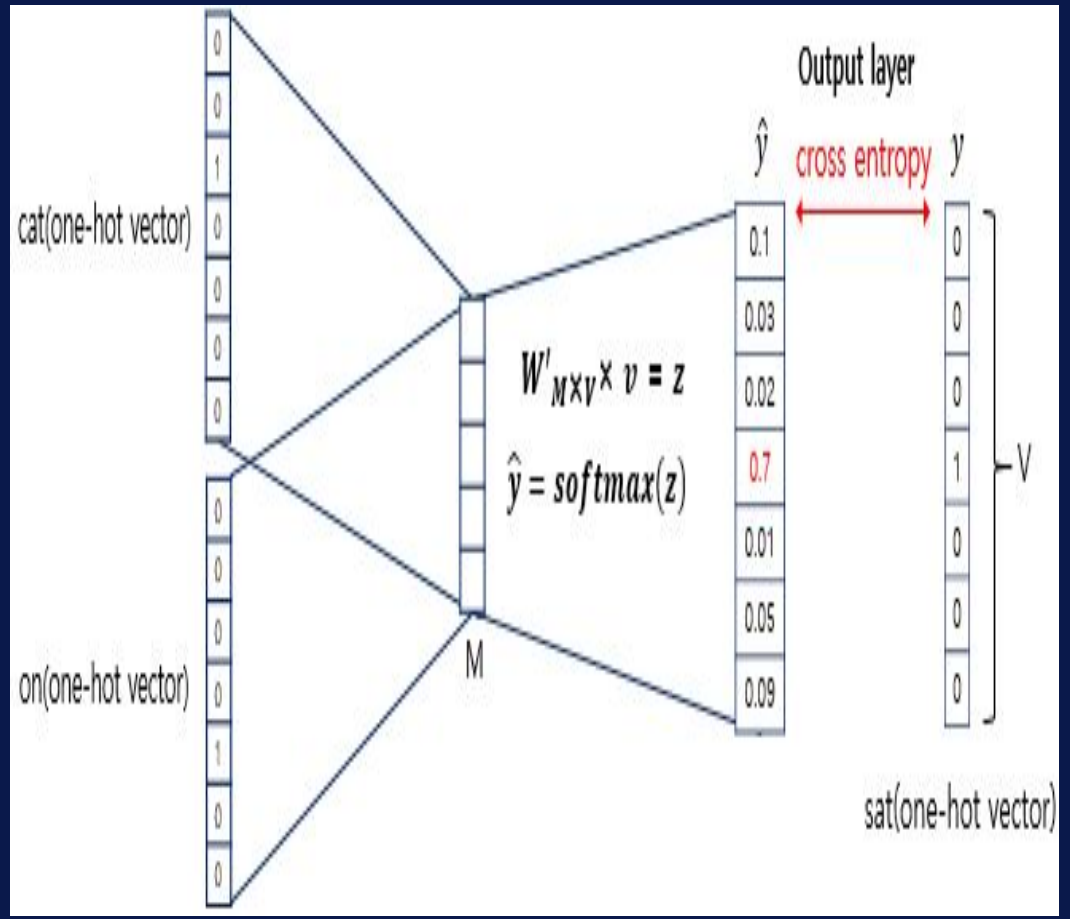
$Loss = (y - \hat{y})^2$ 해당 값을 줄이기 위해 가중치들이 조정된다.

결과적으로 올바른 가중치 행렬 (W , W') 이 나오게 된다.

그래서 결론은?

우리는 희소 표현된 벡터를 밀집 표현으로 바꾸고자 함.

밀집 표현의 행렬은 ? (W') 행렬을 통과한 벡터를 대신 사용하자!



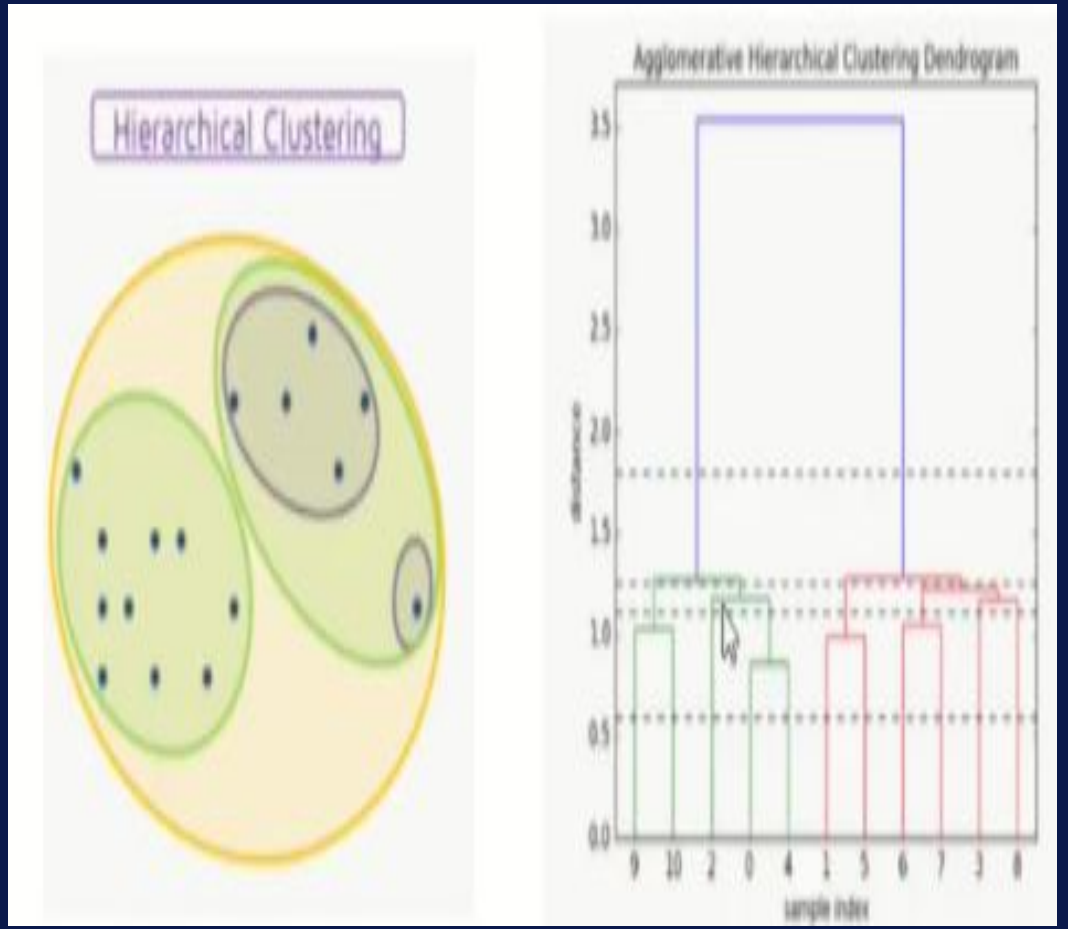
03-0. Clustering

클러스터링의 종류

- 분할적 군집화 :특정 기준에 의해 동시에 각 데이터를 미리 정해진 개수의 군집 중 하나로 군집화 하는 방법 (키, 성별, 소득분위)
- 계층적 군집화 : 가까이에 위치한 데이터들끼리 계층적으로 결합시키는 방식, 군집 개수 사전 설정 **X** (결정 트리)

비지도학습 클러스터링

- **K-means 군집화 : 우리가 배울 것**
- **DBSCAN : 우리가 배울 것**



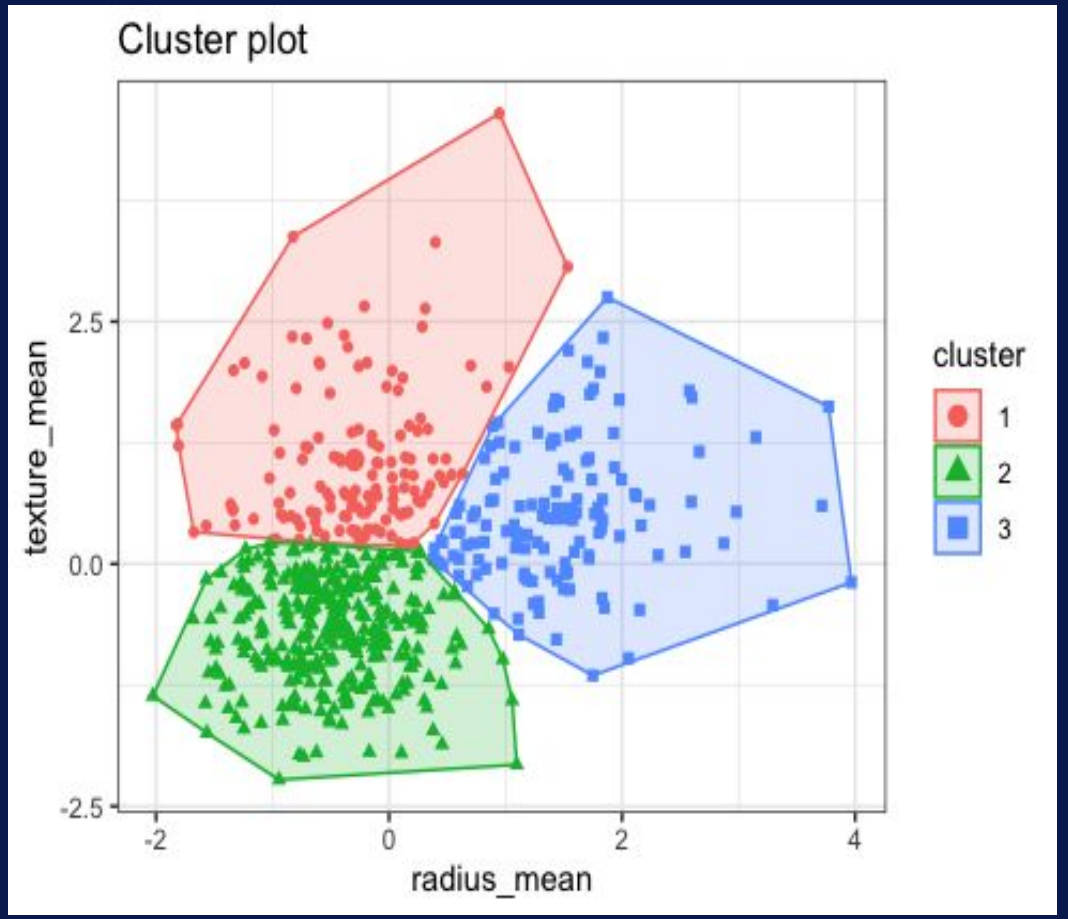
03. K-means

문제의식

- 데이터 간의 유사한 무리들을 어떻게 찾을 수 있을까? (데이터 라벨링 X)
- 실제로 우리가 다루는 데이터는 매우 광활한 영역에 펼쳐져 있다.
- 우리가 이미 알고 있는 매우 중요한 도구 : 거리계산
- 데이터 차원이 매우 크더라도 거리 계산 식의 형태는 변하지 않는다.

그렇다면 시도할 수 있는 것은?

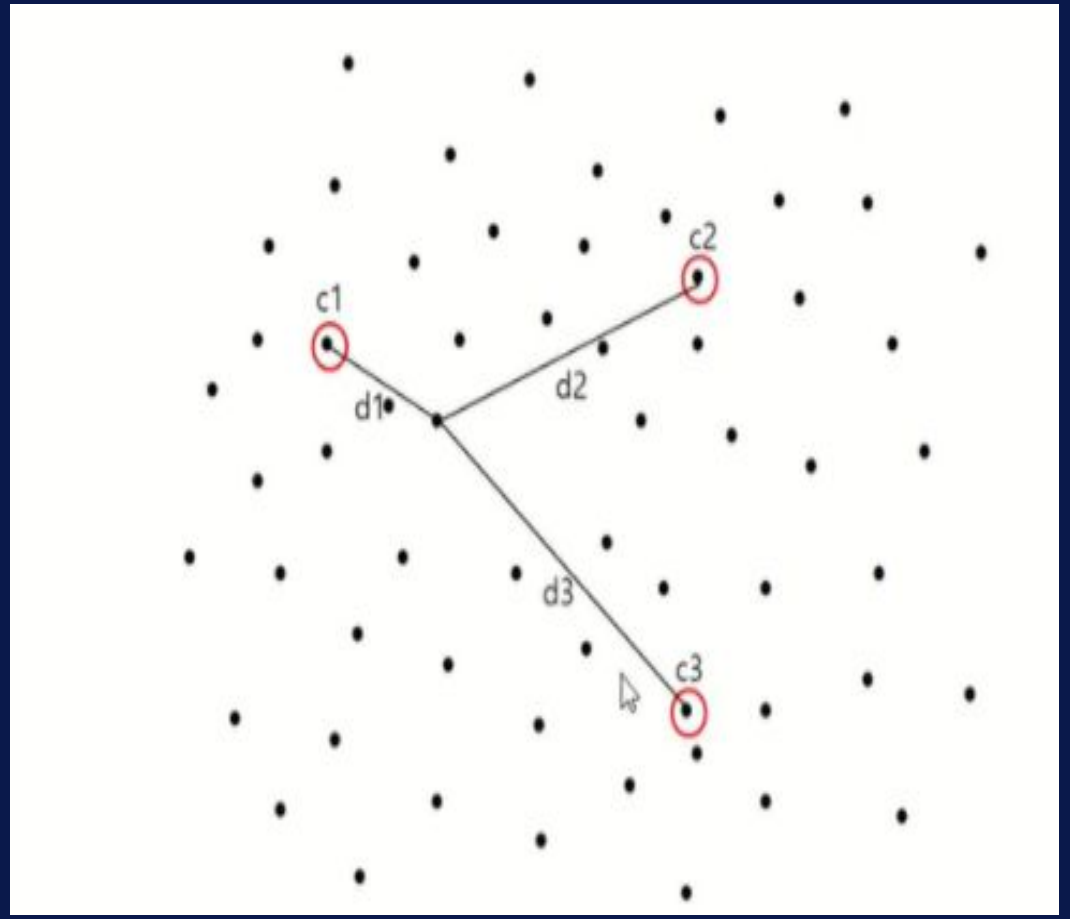
- 군집의 개수를 K 개라고 할 때 어떻게 군집을 형성하는 것이 각 데이터 거리의 합을 최소화 하는지 판단하자.
- 어떻게 ? 반복



03. K-means

K-means의 Flow

- 내가 나누고 싶은 **Cluster**의 개수 구하기 **(k)**
- 초기 데이터 분포에서 **K** 개의 중심점을 임의로 지정
- 각 데이터로부터 **K** 개의 각 중심점 까지의 거리를 계산
- 각 데이터들을 가장 가까운 중심점이 속한 군집에 할당
- **K** 개의 중심점을 다시 계산하여 갱신 (중심점은 각 군집의 데이터들의 평균)
- 중심점이 더이상 변하지 않을 때까지 **3,4,5** 과정 반복

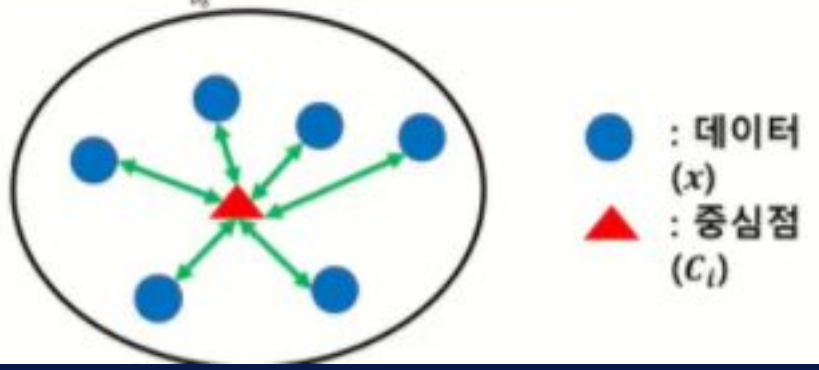


03. K-means

데이터 라벨이 없는데 평가는 어떻게..?(응집도)

- **SSE (Sum of squared error)** : 군집 내의 거리를 고려한 지표

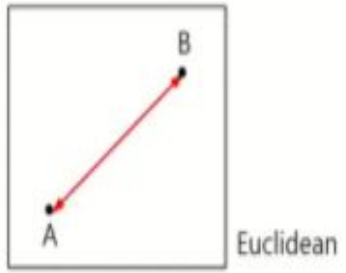
$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist(x, C_i)^2$$



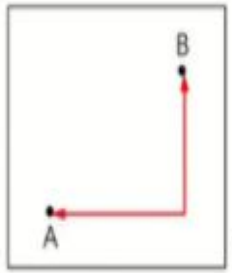
거리 계산의 종류(여러가지 거리 종류)

$$\text{Euclidean Distance (A, B)} = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \ldots + (a_n - b_n)^2}$$

$$\text{Manhattan Distance (A, B)} = |a_1 - b_1| + |a_2 - b_2| + \cdots + |a_n - b_n|$$



Euclidean



Manhattan

03. K-means

데이터 라벨이 없는데 평가는 어떻게..?(응집도)

- 실루엣 계수

- a(i) : 데이터 i로부터 같은 군집 내에 있는 다른 모든 데이터들 사이의 평균 거리 (클러스터 응집도, 작을수록 좋음)

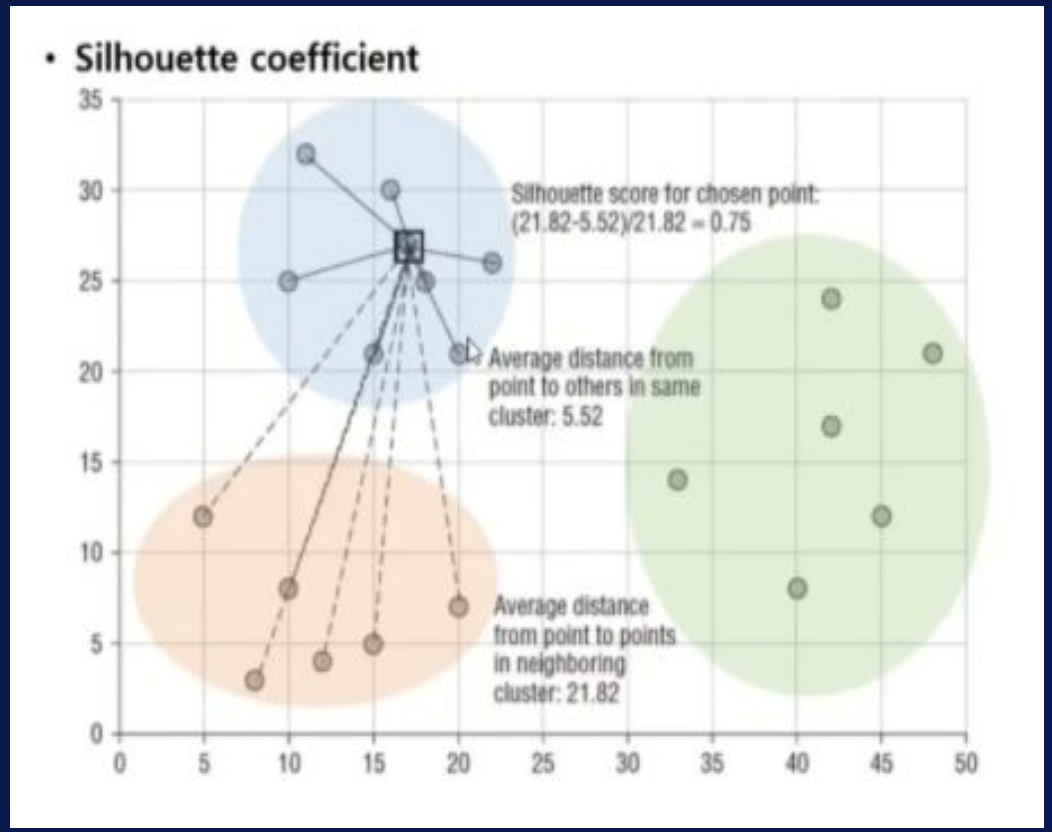
-b(i) : 데이터 i 로부터 가장 가까운 인접 군집 내에 있는 데이터들 사이의 평균 거리 중 가장 작은 값 (클러스터의 분리도 ,클 수록 좋음)

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}},$$

$$-1 \le s(i) \le 1$$

$$\bar{S} = \frac{1}{n}\sum_{i=1}^{n} S(i)$$

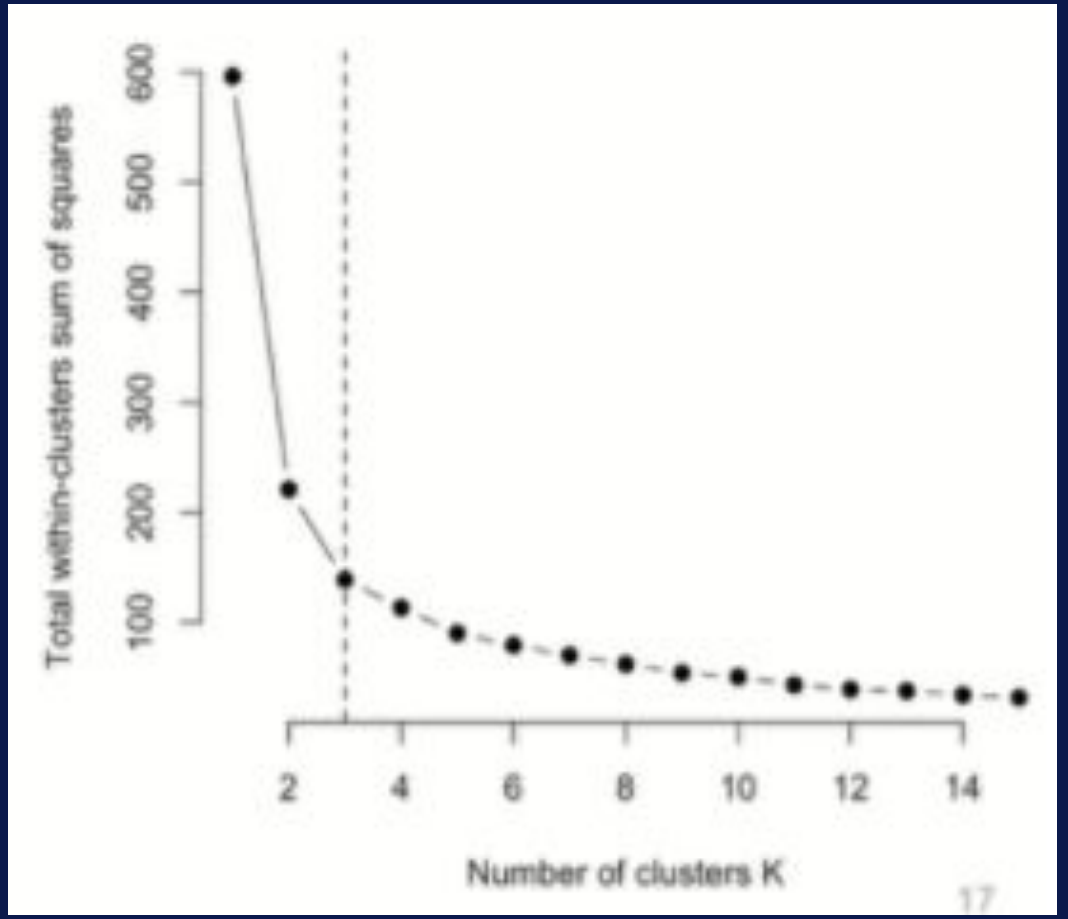
1에 근접할 수록 좋은 값



03. K-means

데이터 라벨이 없는데 평가는 어떻게..?(K의 개수)

- **K**의 값 마다 전체 오차는 달라질 것이다.
- 가장 적절한 **K** 값을 찾기 위해 **K** 마다 최적의 **SSE** 를 구한다.
- 급격하게 **SSE**가 떨어지는 지점을 파악한다. **(Elbow Point)**
- 해당 값이 적절한 **K** 값일 것 ! **(Over-fitting** 을 고려하며 최적의 효율 내기)

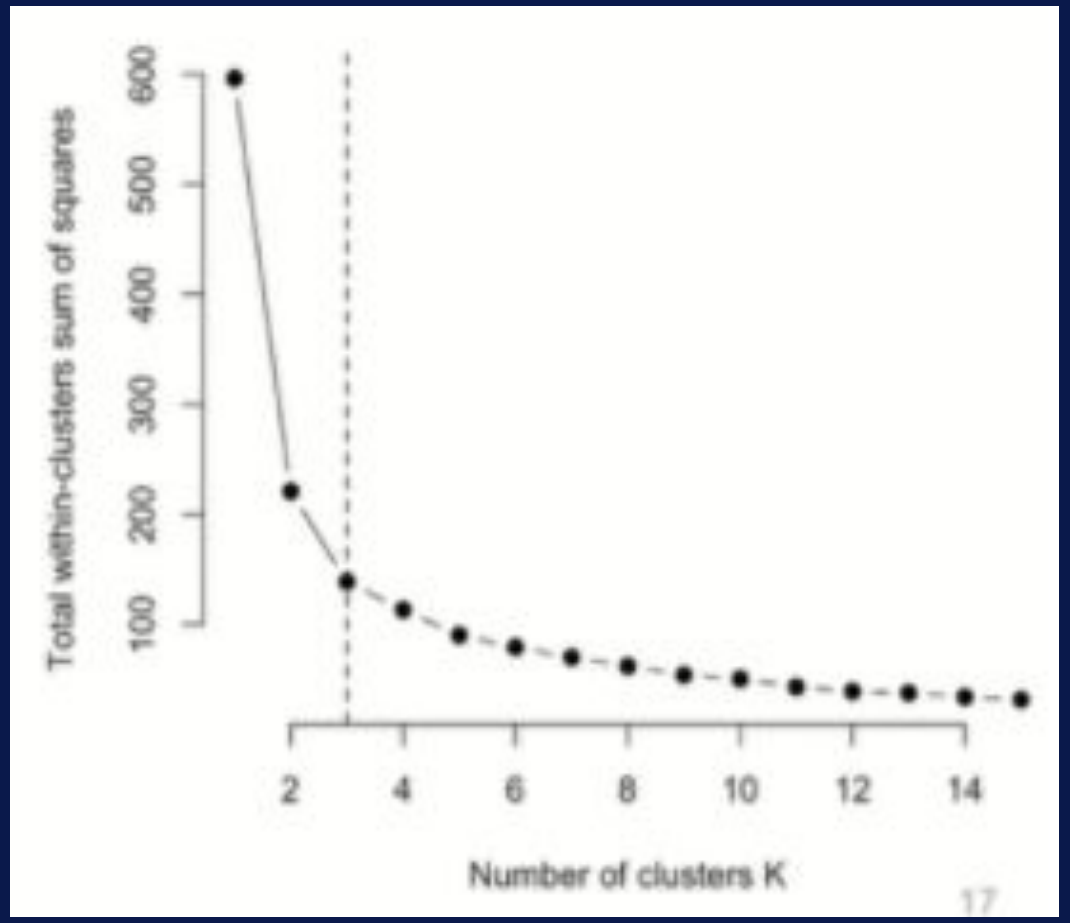


04. DBSCAN

데이터의 특성은 매우 매우 다양하다..

If 분자 구조 데이터 클러스터링 : 클러스터링 후 기하적 특성을 표현하는 것이 유효할 것!

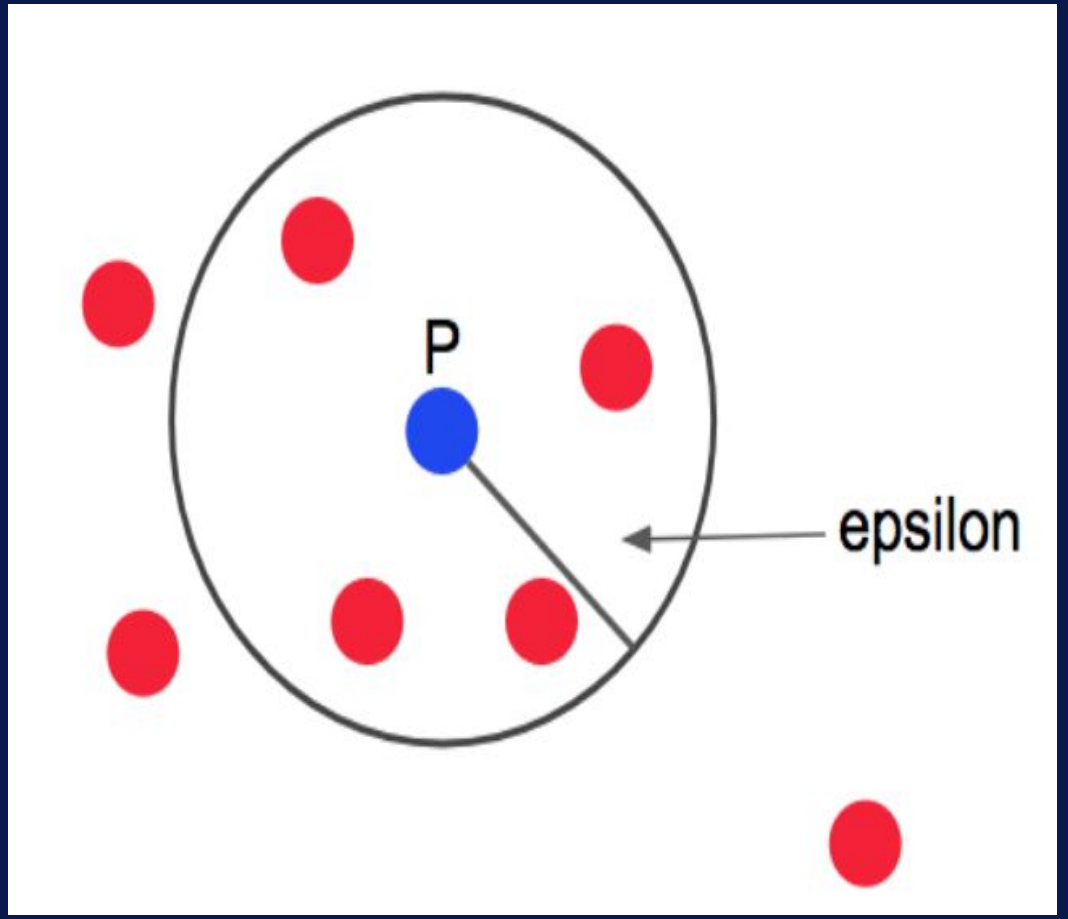
DBSCAN : K-means 와 비슷하면서도 다른 **Clustering**



04. DBSCAN

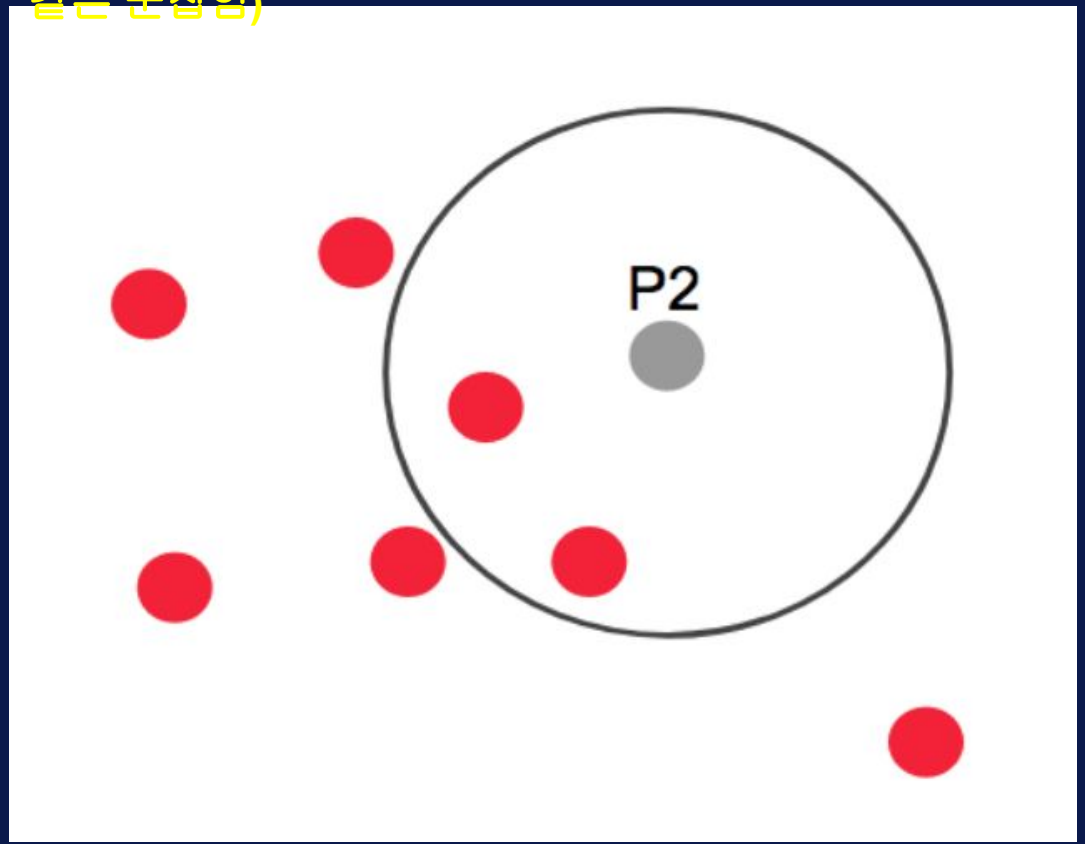
DBSCAN의 가정

- 군집을 구성하는 요소는 최소 거리 **e** 와 데이터 샘플의 수 **M** 이다.
- 점 **P** 에서 거리 **e** 안에 데이터가 **M** 개가 있다면, 하나의 군집으로 인식한다.
- 파란 점 **P** 는 두 조건을 모두 만족하므로 **Core point** 가 된다.
- 이후 모든 점에 대해 땅따먹기 진행!

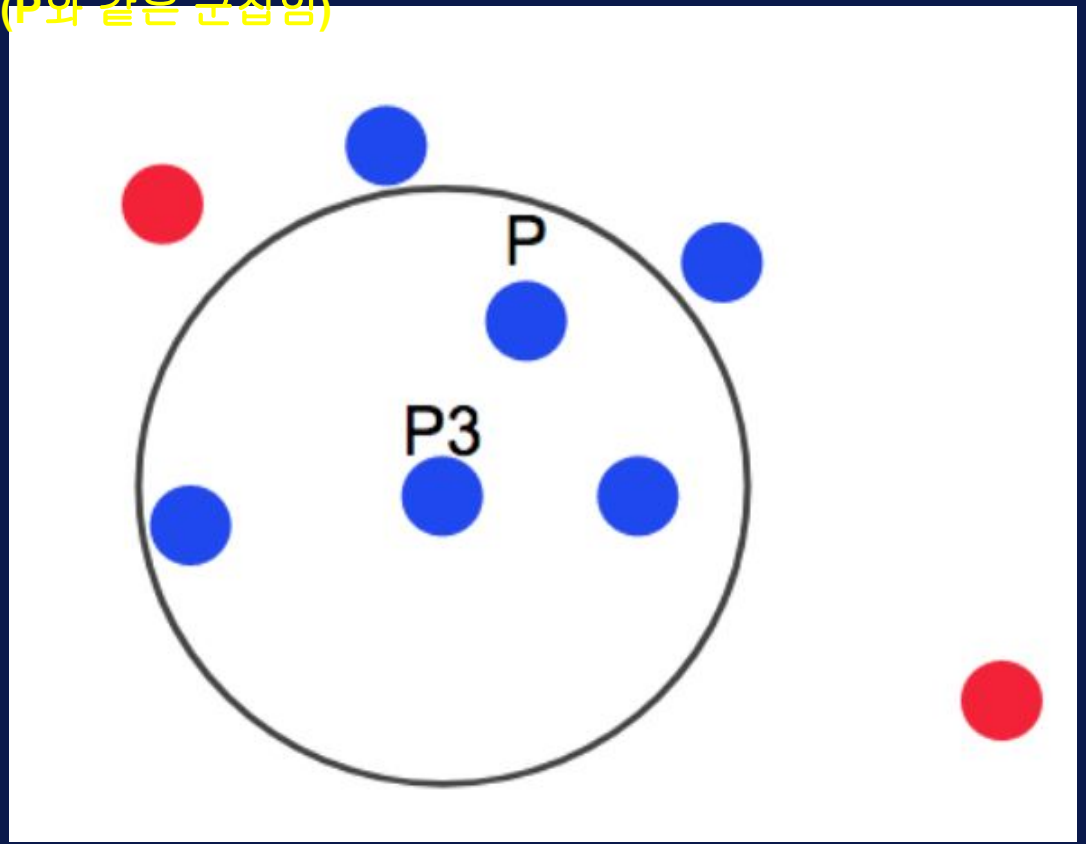


04. DBSCAN

Case1 :e 범위 안에 **P** 가 들어가지만 들어간 데이터 샘플 수가 **M** 이하이므로 코어 포인트는 아니다! **(P와 같은 군집임)**



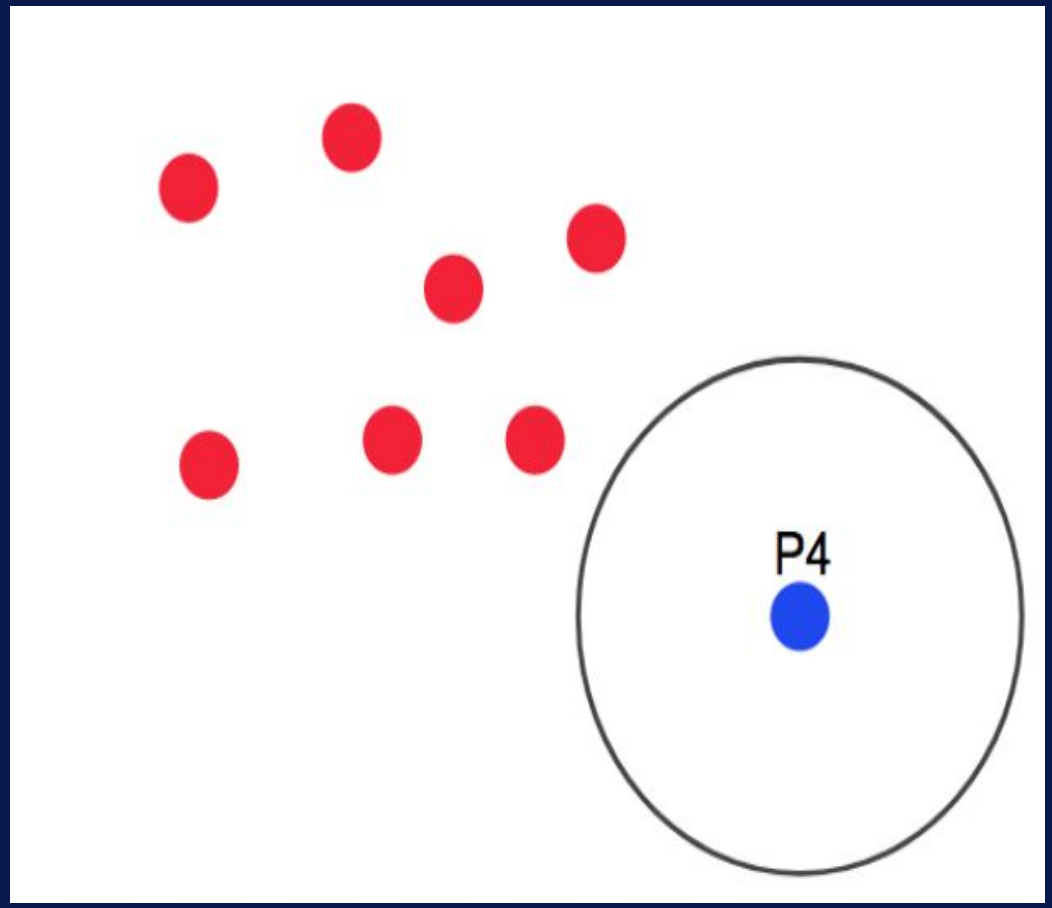
Case2 :e 범위 안에 **P** 가 들어가며 들어간 데이터 샘플 수가 **M** 개 이상이므로 **P** 와 같은 코어 포인트! **(P와 같은 군집임)**



04. DBSCAN

DBSCAN의 특징

- **P4**의 경우 **e** 거리 안에 데이터 샘플이 아무것도 없으므로 **Noise point** 로 지정
- 데이터 군집에 속하지 않은 **Noise point**를 지정함으로써 전반적인 데이터 분포의 양상을 강건하게 표현할 수 있음
- 기하적 모형을 띄는 데이터 분포를 잘 표현할 수 있다. (왜 그럴까?)



05. 추천하고 싶은 학습

비지도 학습의 종류 중 하나인 GMM(vanilla 모형)을 공부해보기

(딥러닝 머신러닝에서 확률 분포를 어떻게 활용하는지 IDEA를 이해할 수 있다. 차후 GAN, AE, VAE, GNN 등을 이해할 때 필수적)

딥러닝 초석 세우기 (Back Propagation 이해)

00. 과제

컴페티션 제출 잘 하기

첨부자료 출처

02. word embedding

https://wikidocs.net/book/2155

03. K-means

https://zephyrus1111.tistory.com/179

03. DBSCAN

https://github.com/skdytpq/skdytpq.github.io/blob/master/_posts/2022-08-15-비지도학습.md

---

폰트

네이버 글꼴 모음 _ 나눔 스퀘어 사용
출처 : https://hangeul.naver.com/font

D&A
ML Session 9차 시

Thank You .

2022 / 11 / 22
D&A 운영진 나요셉