



Clustering

R로 쉽게 이해하기!

조인식 조교

연령대코드	시도코드	신장(5Cm)	체중(5Kg)	허리둘레	시력(좌)	시력(우)	구강검진 수검0
10	27	175	75	86	1	1.2	0
11	41	160	65	85	1.5	1.2	0
11	43	150	55	80	1.5	1.5	1
12	48	175	70	85	1.2	1.2	1
6	11	160	50	67	0.4	0.5	0
9	46	170	55	64	1.1	1.2	1
10	31	175	70	90	1.5	1.5	0
16	48	155	60	81	0.7	0.5	1
16	27	165	60	74	0.4	0.6	1
13	48	160	60	80	0.8	0.2	0
7	46	170	80	86	1.5	1	0
6	11	155	50	66	0.9	0.6	0
7	41	170	80	92	0.6	0.4	1
15	11	155	65	81	0.8	0.5	1
15	28	160	65	79	0.7	0.2	0
11	48	165	80	93	0.7	0.8	1
9	41	170	75	81	1.2	1.2	0

독립변수(x) 종속변수(y)

각 x인자에 상응하는 y인자에 대하여 알고리즘이 학습하여 추후 새로운 x인자가 들어왔을때 해당 데이터의 y인자를 예측

연령대코드	시도코드	신장(5Cm)	체중(5Kg)	허리둘레	시력(좌)	시력(우)
10	27	175	75	86	1	1.2
11	41	160	65	85	1.5	1.2
11	43	150	55	80	1.5	1.5
12	48	175	70	85	1.2	1.2
6	11	160	50	67	0.4	0.5
9	46	170	55	64	1.1	1.2
10	31	175	70	90	1.5	1.5
16	48	155	60	81	0.7	0.5
16	27	165	60	74	0.4	0.6
13	48	160	60	80	0.8	0.2
7	46	170	80	86	1.5	1
6	11	155	50	66	0.9	0.6
7	41	170	80	92	0.6	0.4
15	11	155	65	81	0.8	0.5
15	28	160	65	79	0.7	0.2
11	48	165	80	93	0.7	0.8
9	41	170	75	81	1.2	1.2

독립변수(x)

구강검진 수검률
0
1
1
0
1
0
1
1
0
0
1
1
0
1
0
1
0
0

종속변수(y)

비지도학습을 위한 데이터는 y인자가 없다
(혹은 있어도 비지도학습에선 이를 취급하지 않음)

연령대	코드	시도코드	신장(5Cm)	체중(5Kg)	허리둘레	시력(좌)	시력(우)
10	27	175	75	86	1	1.2	
11	41	160	65	85	1.5	1.2	
11	43	150	55	80	1.5	1.5	
12	48	175	70	85	1.2	1.2	
6	11	160	50	67	0.4	0.5	
9	46	170	55	64	1.1	1.2	
10	31	175	70	90	1.5	1.5	
16	48	155	60	81	0.7	0.5	
16	27	165	60	74	0.4	0.6	
13	48	160	60	80	0.8	0.2	
7	46	170	80	86	1.5	1	
6	11	155	50	66	0.9	0.6	
7	41	170	80	92	0.6	0.4	
15	11	155	65	81	0.8	0.5	
15	28	160	65	79	0.7	0.2	
11	48	165	80	93	0.7	0.8	
9	41	170	75	81	1.2	1.2	

알고리즘이 비슷한 개체끼리
군집화(Clustering)시켜줌

알고리즘을 활용한다?

- 머리좋은 여러 학자들이 “어떻게 군집화를 시킬까?” 에 대해서 다양한 아이디어를 떠올렸고 이를 알고리즘이라는 이름으로 설명을 해놓았다.
- 이런 학자들이 꾸려놓은 알고리즘을 R에서 다수 패키지로 구현되어있어 우리는 이를 끌어다쓰면 된다. (R의 강력한 이점 중 하나)

클러스터링 알고리즘

거리기반 K-means, k-medoids

모델기반 DBSCAN, OPTICS

밀도기반 GMM(Gaussian Mixture Model)

Etc... 개선알고리즘

클러스터링의 활용

데이터 군집화를 통해 Label 속성값이 없는 데이터를 구분화 및 그룹핑

- 언론사의 뉴스/신문 문서의 그룹핑 : 텍스트마이닝과 혼합활용
- 이미지 클러스터링을 통해 이미지의 특정 장면을 포착
- 네트워크 유해 트래픽 감지
- SNS “비슷한 사용자” 그룹핑
- 금융, 보험사 등에서 고객 차별화
- 공간데이터분석

이상치 탐지에 활용

- DBSCAN, Isolation Forest 등의 알고리즘들의 활용

EDA 과정에서 데이터를 이해하기 위하여 활용

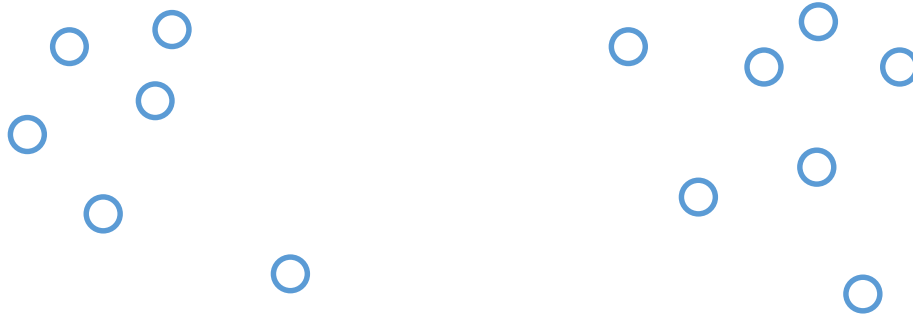
1차 데이터 분류 : 대규모 데이터셋을 클러스터로 분류하여 각 클러스터별로 데이터를 세분화 한 뒤 개별 기계학습 활용

간단하게 이해하는 K-means Algs.

1. 군집수 (k) 를 지정한다.
 2. 군집점을 랜덤한 좌표에 초기화시킨다.
 3. 데이터들과 군집점간의 거리를 계산하여 가까운 군집점에 데이터들이 속하도록 한다.
 4. 군집점을 소속된 데이터들의 중앙으로 업데이트한다.
- 3-4 과정을 반복한다.
- 종료시점 : (1)군집점의 위치가 바뀌지 않거나, (2)지정한 반복수가 끝날 경우

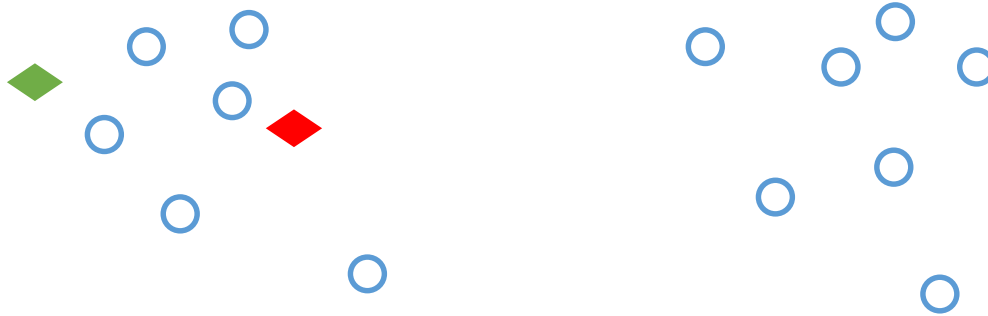
1. 군집수를 지정한다.

$K = 2$



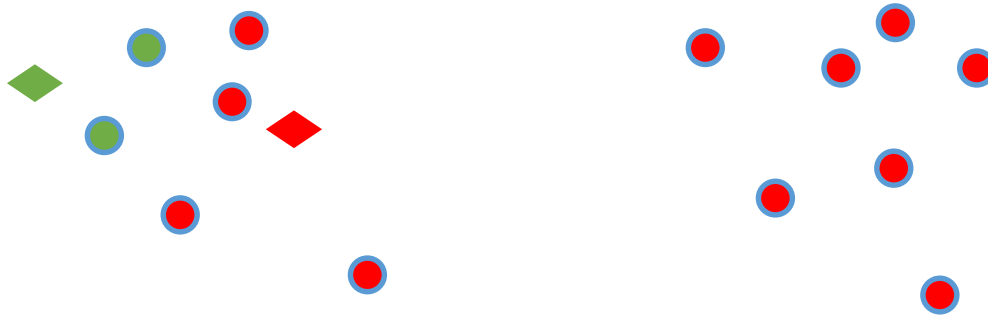
○ Data point
◆ Centers

2. 군집점을 랜덤한 좌표에 초기화시킨다.



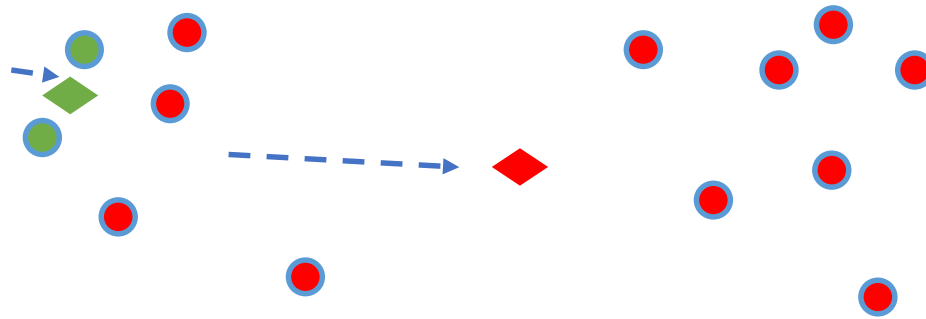
○ Data point
◆ Centers

3. 데이터들과 군집점간의 거리를 계산하여 가까운 군집점에
데이터들이 속하도록 한다.



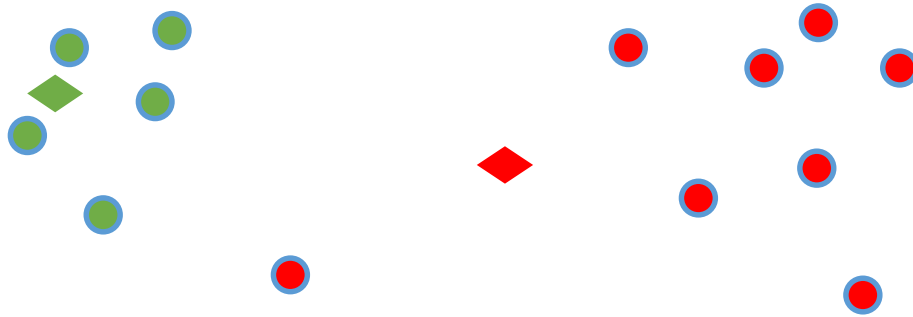
○ Data point
◆ Centers

4. 군집점을 소속된 데이터들의 중앙으로 업데이트한다.

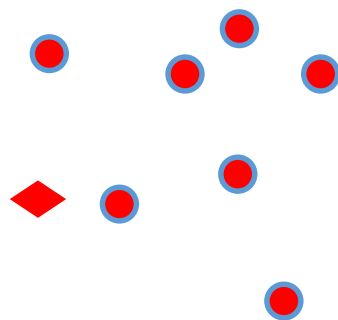
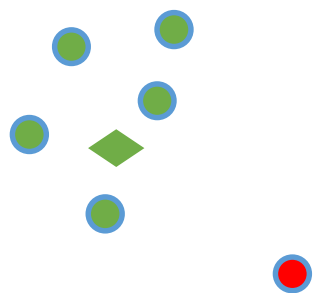


○ Data point
◆ Centers

3. 데이터들과 군집점간의 거리를 계산하여 가까운 군집점에
데이터들이 속하도록 한다.

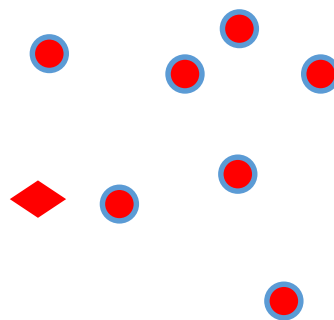
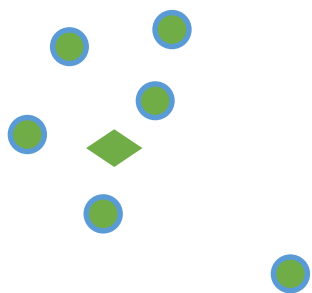


○ Data point
◆ Centers



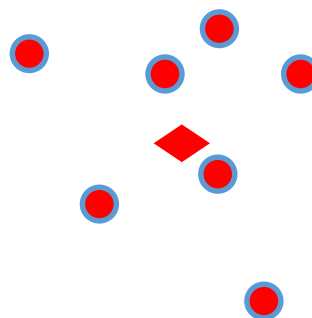
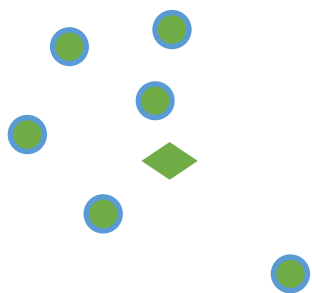
○ Data point
◆ Centers





○ Data point
◆ Centers

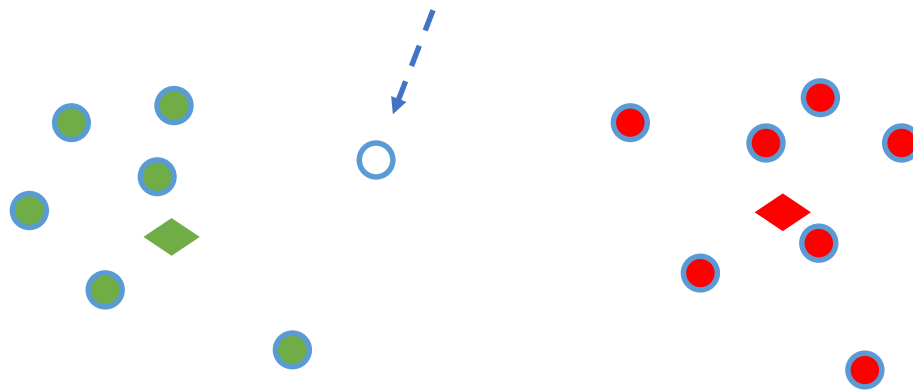




○ Data point
◆ Centers

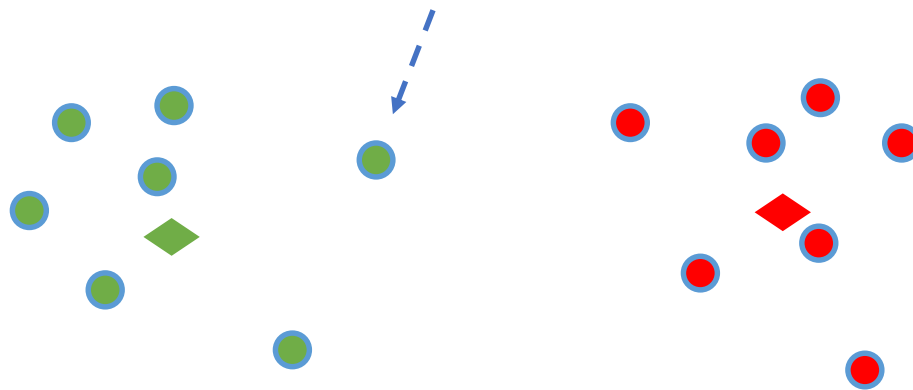


새로운 데이터가 들어오면 center값에
가장 가까운 클러스터에 포함된다.



○ Data point
◆ Centers

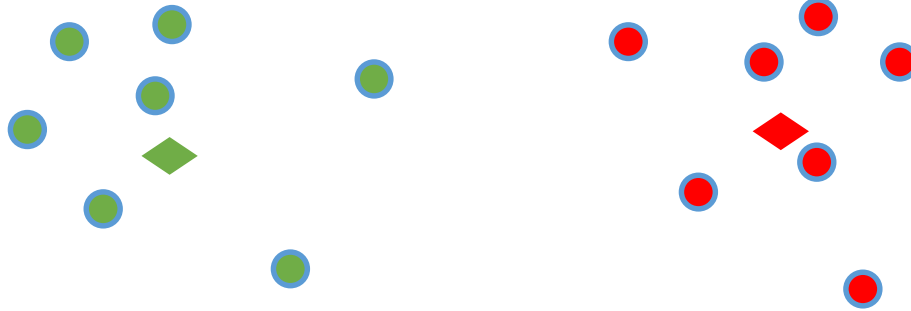
새로운 데이터가 들어오면 center값에
가장 가까운 클러스터에 포함된다.




○ Data point
◆ Centers




이상치의 데이터가 들어온다면 이를 클러스터에
포함하는게 옳을까?
(Outlier Issue)

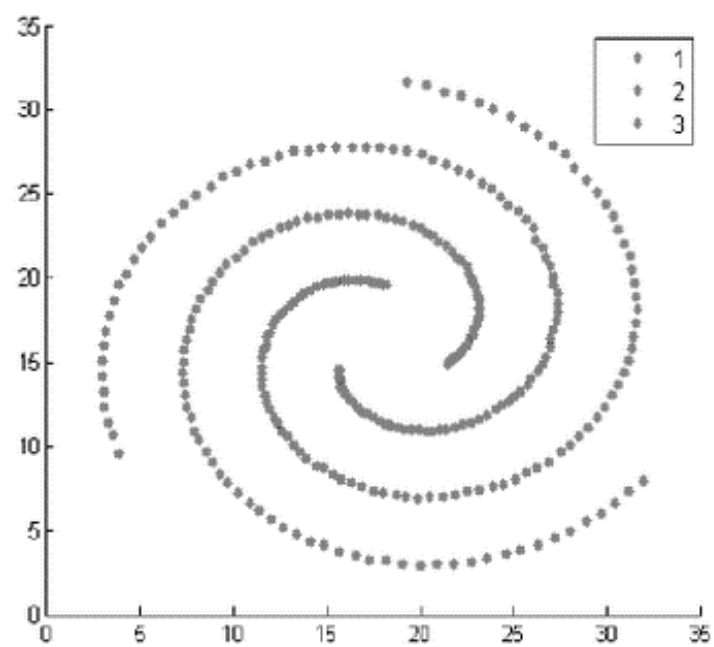


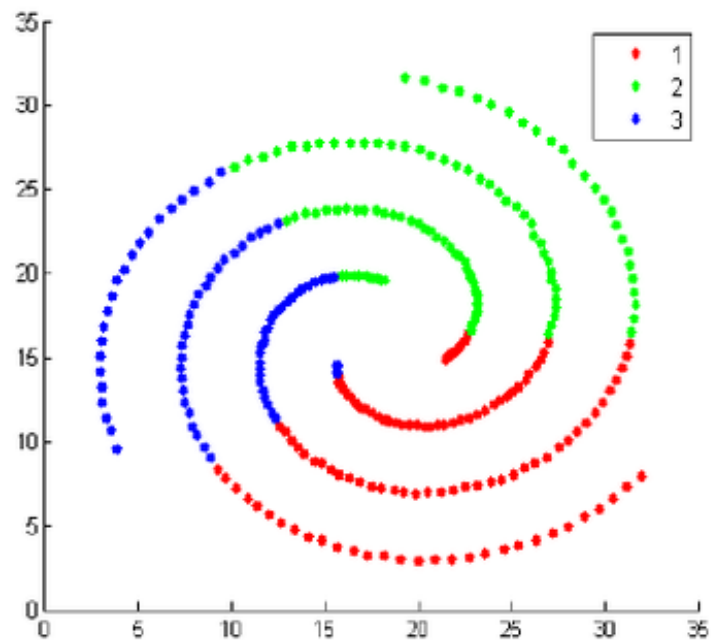
○ Data point
◆ Centers



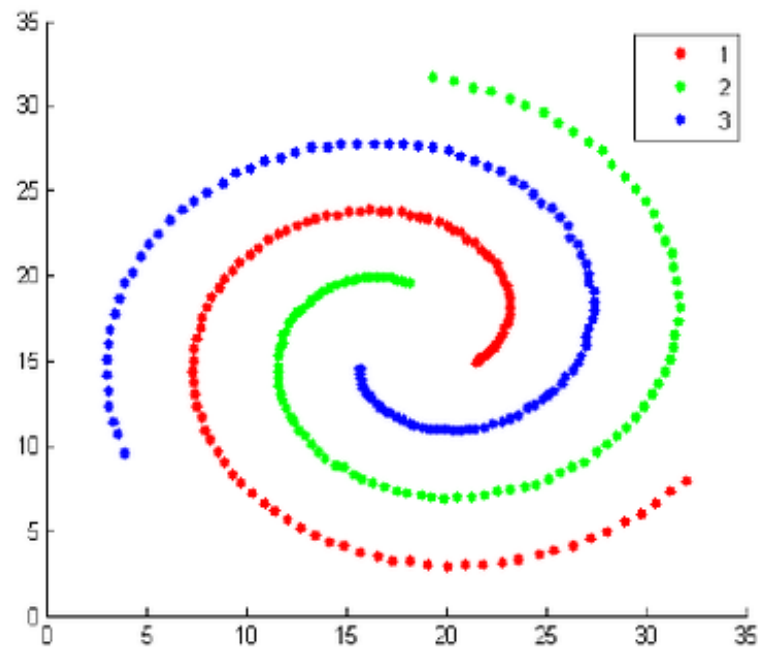
이해가 됐다면
실습으로 알아가봅시다!







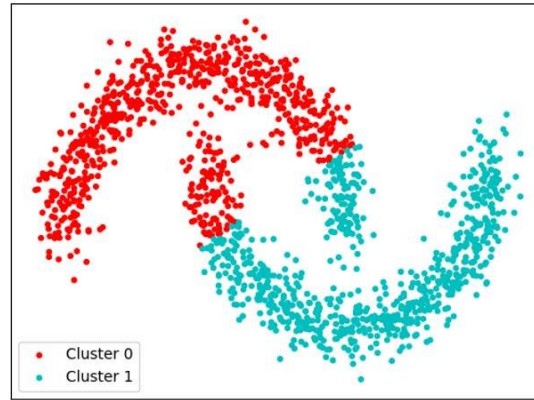
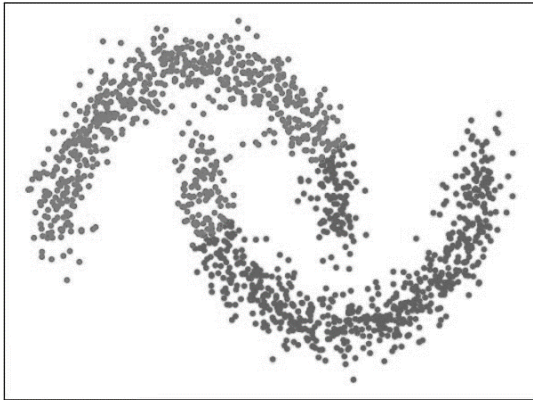
K-means : center
point와의 거리기반



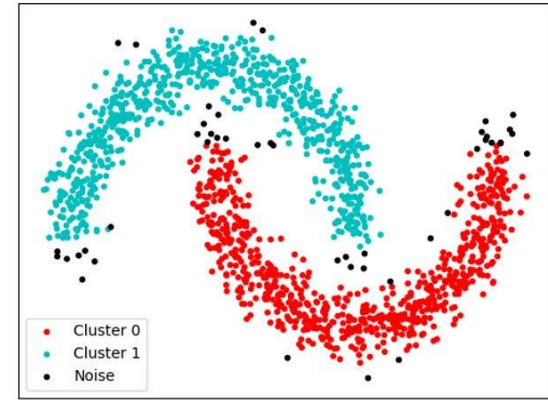
밀도기반 클러스터링

이미지출처 : ResearchGate – Results of k-means / DBSCAN on Jain dataset

비지도학습



K-means



OPTICS

DBSCAN : 밀도기반 클러스터링

DBSCAN(Density-based spatial clustering of applications with noise)

“데이터들이 얼마나 몰려있는가”를 토대로 클러스터링을 하는 알고리즘

“점의 어느 범위 내에 몇 개의 점이 있다면 이는 하나의 클러스터로 인식”

알고리즘을 이해하는데 필요한 용어!

Epsilon(e) : 특정 점을 중심으로 minimum points의 개수를 셀 범위(거리)

Minimum Points(mp) : 해당 점을 판단하기 위한 거리내의 점의 수

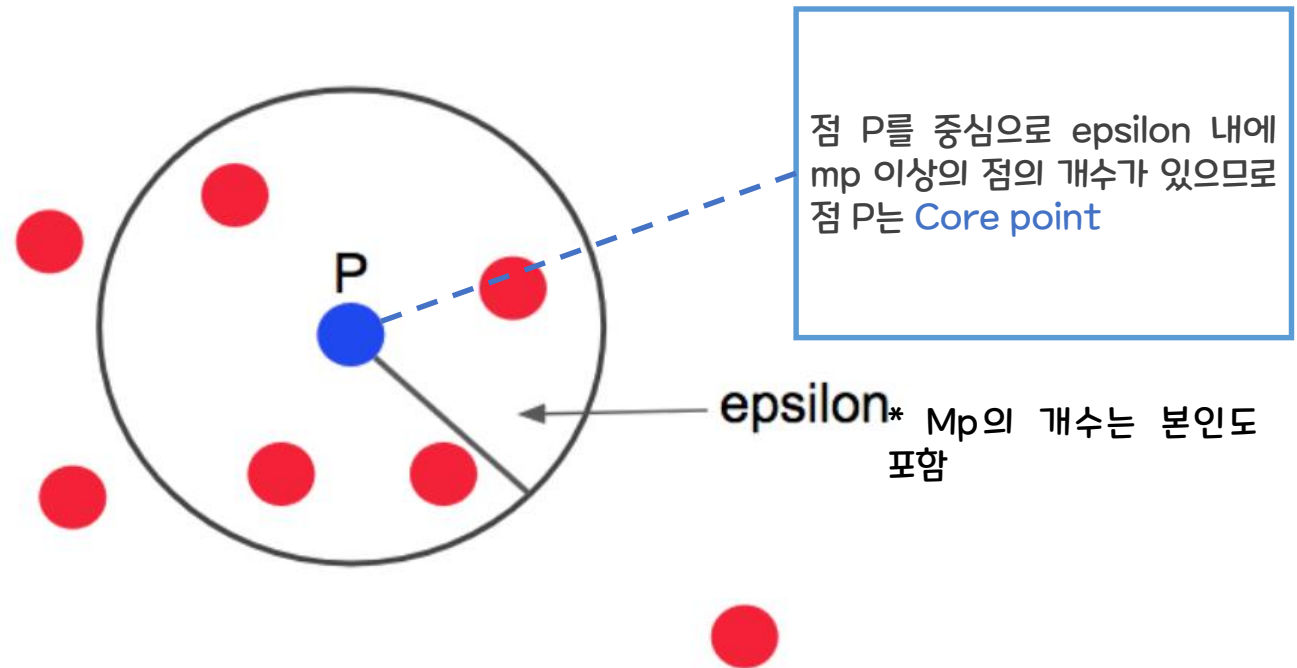
Core point : 점을 중심으로 e거리내의 점의 개수가 mp 이상인 점

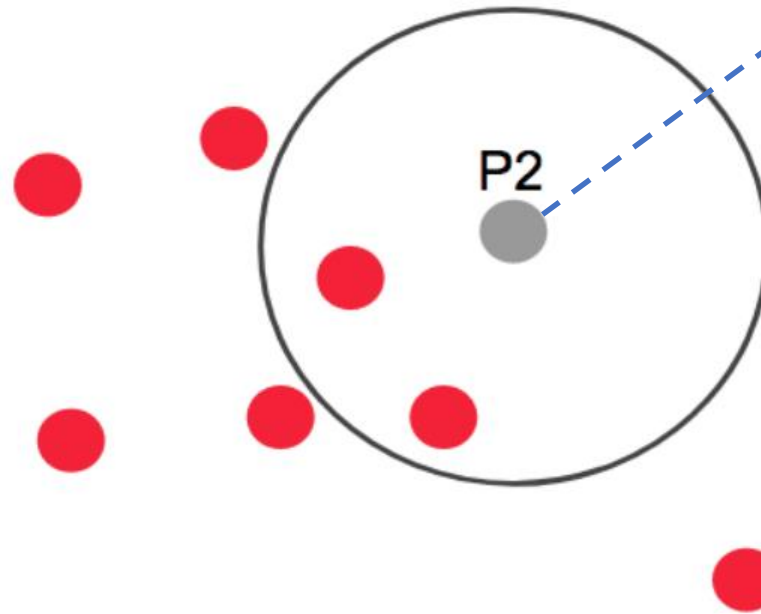
Noise point : 점을 중심으로 e거리내의 점의 개수가 mp에 미치지 못한 점

Border point : 점을 중심으로 e거리내의 점의 개수가 mp에 미치지 못하지만 mp에 속한 점들이 cluster에 속해있는 경우의 점

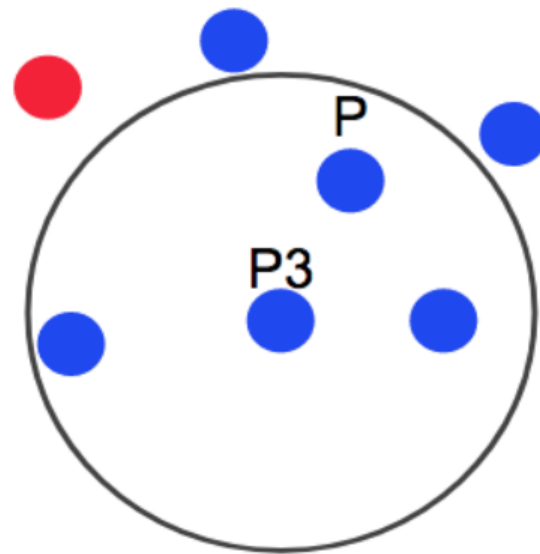
DBSCAN을 이해해보자!

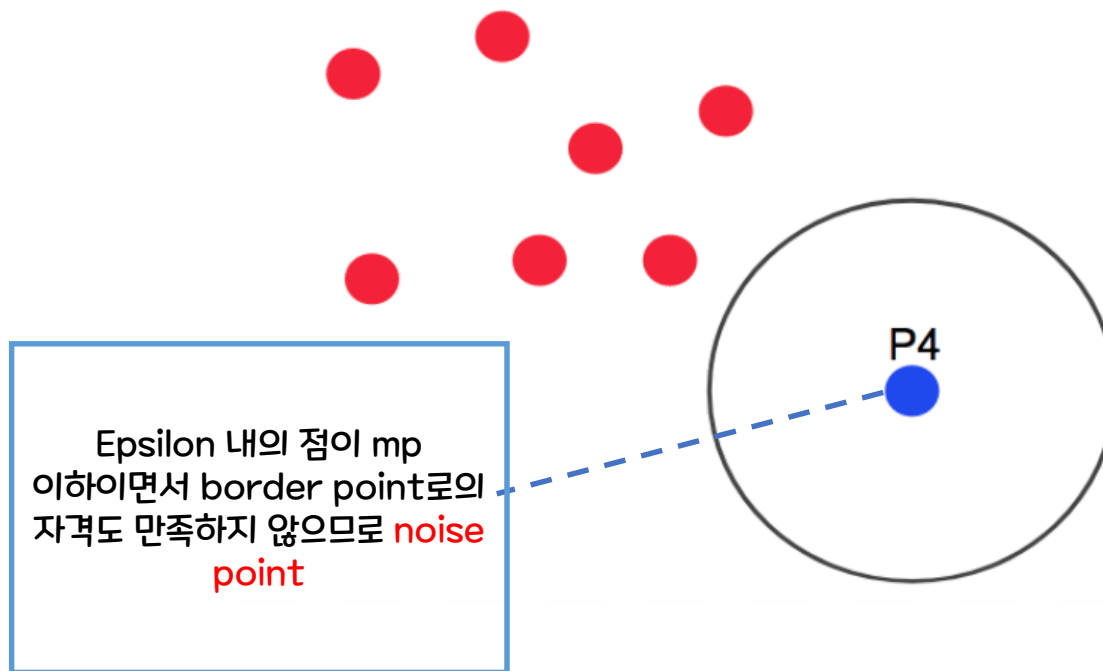
Epsilon : 사용자가 지정
Minimum points : 4라 지정






점 P2를 중심으로 epsilon 내에
mp 이상의 점의 개수가 없으나
epsilon 내의 점이 점P로 구성된
Cluster에 속한 점이 있으므로
P2는 Border point







이해가 됐다면
실습으로 알아가봅시다!

