

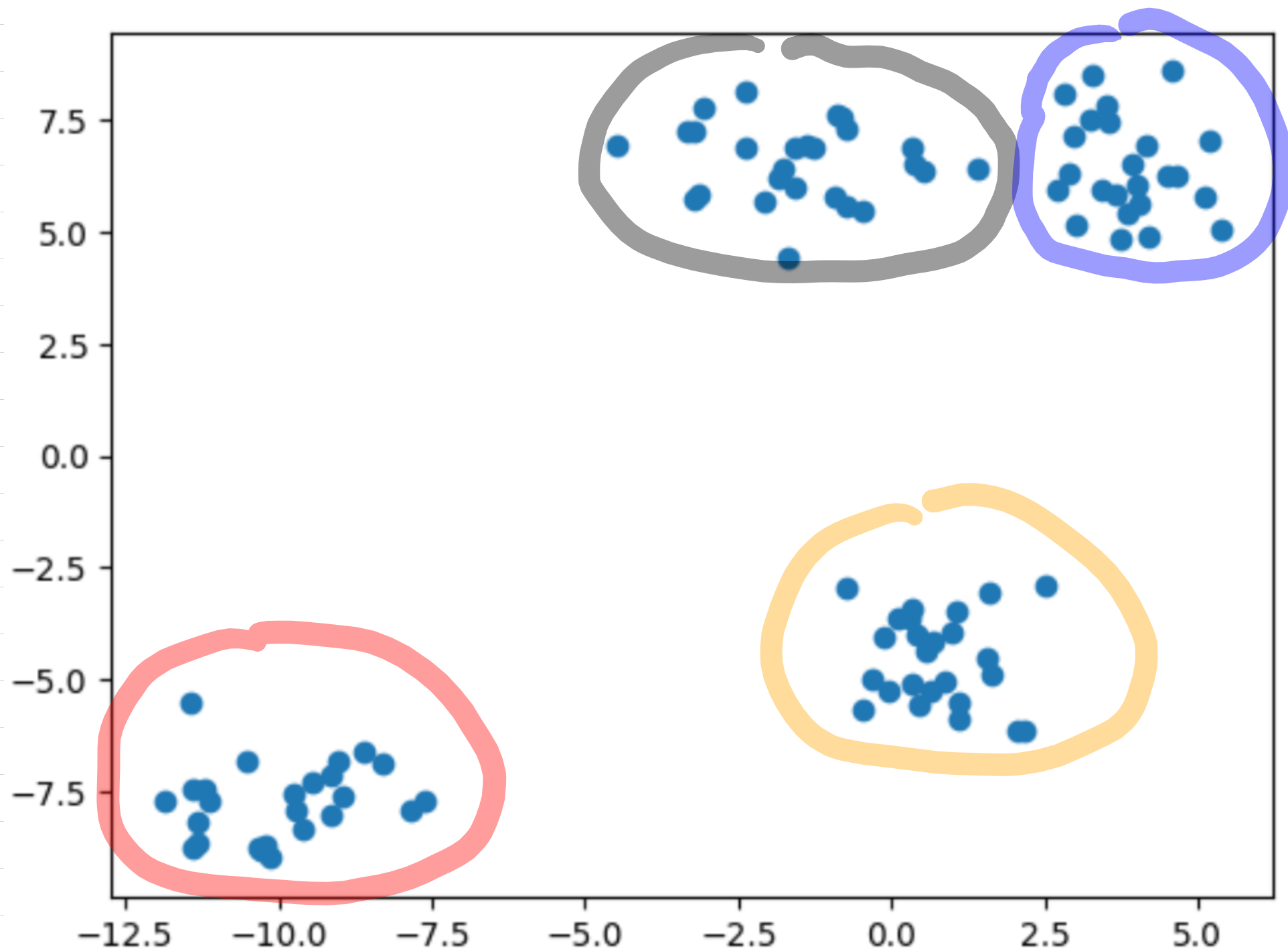
12/08 k-means ☹

① 비지도학습 → unlabeled 데이터에 숨겨진

pattern을 찾아서 구조화 (clustering)

하는 작업을 clustering

★  
(  
    k-means → EM 알고리즘을 기본 동작  
    DBSCAN   "Code3 작업"



# ① k-means (거리기반)✓

② 비지도학습 알고리즘

③ hyperparameter →  $k$  클러스터의 수

④ EM 알고리즘을 기반으로 동작

⑤ 거리가 멀면서 밀집하게 연결된 data를  
성능으로 구분하기 힘들어요

⑥ 속도면에서 느림, 불안정

이걸 해결하기 위해

★ DBSCAN

✓ 밀도기반

(속도가  
느려질수도)

# ① DBSCAN

Density (밀도)  
Based  
Spatial  
Clustering  
of  
Application with  
Noise

\*\*\*

그림을 보고 알면 이해하는데 좋아요!!

특징

① k-means 처럼 클러스터의  
개수  $k$ 를 지정할 필요가  
없어요!!

② 거리가 변이 X  
조목조목 물려왔던 클러스터를 생성  
data로

③ 클러스터의 최초의 점의 중심에서  
파져나가서 생성.

# ★ DBSCAN 알고리즘이 사용하는 용어

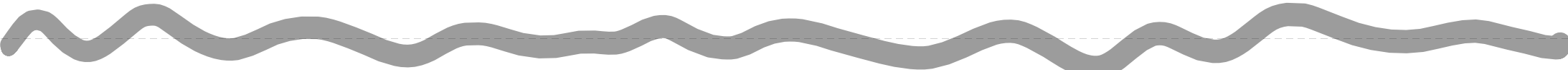
- ①  $\epsilon$  (엡실론) : 클러스터의 반경
- ② minPts : 클러스터를 이루는 최소 개체수
- ③ Core point : cluster의 중심점.
- ④ border point : cluster의 위치를 지는 Core가 아닌 점.
- ⑤ noise point : cluster의 위치를 지지 못하는 data

## 알고리즘

① 입력의 데이터 1점 설정. P로 포함해서 epsilon 이내의 데이터의 개수를 세요!

② 해당 cluster에 mpts가 이량의 data가 존재하면 P가 core point가 되고 하나의 cluster를 생성!

✓  
③ 새로운 P'가 core point가 되고 이점이 기존의 cluster에 속한다면 하나의 cluster로 묶어요!



## DBSCAN

(k-Means)

① 밀도가 높은 것이기 때문에 거리 기반 clustering 기법에서  
해결하지 못한 clustering 처리가 가능

② k (클러스터 수)를 지정하지 않아도 됨

[epsilon  
minPts] 두개를 대신 지정해야 함

③ 데이터가 많아지면 알고리즘의 수행속도가 많이  
느려져요!

# ④ 차원 축소 (Dimension Reduction)

고차원 Data를 저차원 Data로 변환

(4차원)  $\longrightarrow$  (2차원)

주의  $\rightarrow$  저차원 data 표현이 고차원 원본 data를  
잘 표현할 수 있어야 해요!!

필연적으로 문제가 발생해요!!

필연적으로 information loss (정보 손실)이 발생

하는데 이 손실을 최소화하려면 어떻게 해야  
하나요??



① 차원축소가 필요해요?

"고차원 data"

↳ 상대적으로 data가 sparse해 지요

↳ 데이터 분포도 희박하고

머신러닝 학습도 힘들어요

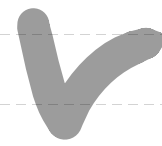
↳ Overfitting 현상 발생

○ "시각화가 가능"

○ "속도가 높고 데이터 종류를 줄일 수 있어요"

② 차원축소의 간접 → 정보 손실이 따릅니다!!

# 차원 축소가 가능한 알고리즘

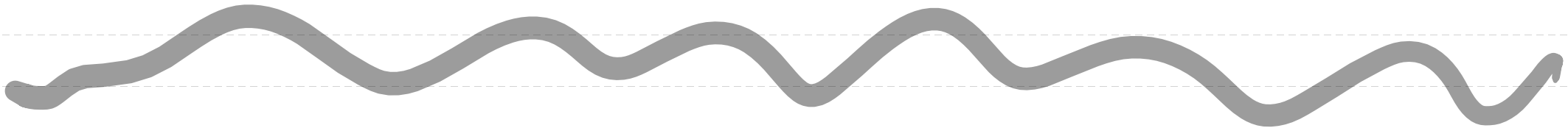


선형 변환 (projection)

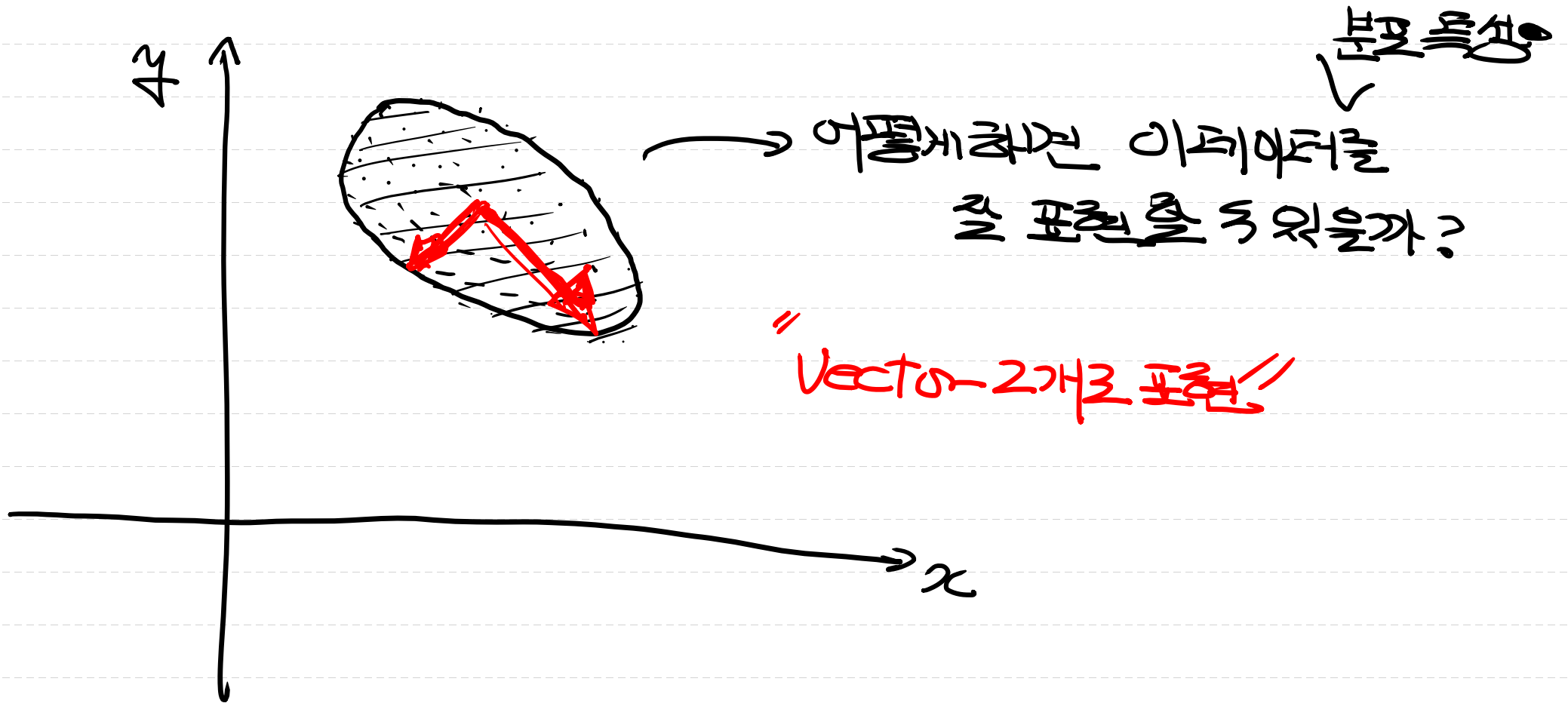
- 주성분 분석 (PCA, Principal Components Analysis)
- 특이값 분해
- SVD

비선형 변환 (Manifold)

- LLE
- Isomap
- ...



# ① PCA (주성분 분석)



② code