Undergraduate Research Internship in Affective AI LAB.

혼자 공부하는 머신러닝 & 딥러닝

5주차 : ch06. 비지도 학습

컴퓨터공학과 12201682 공지윤

목차

군집화 (Clustering)

- K-Means
- Mean Shift
- Gaussian Mixture Model
- DBScan

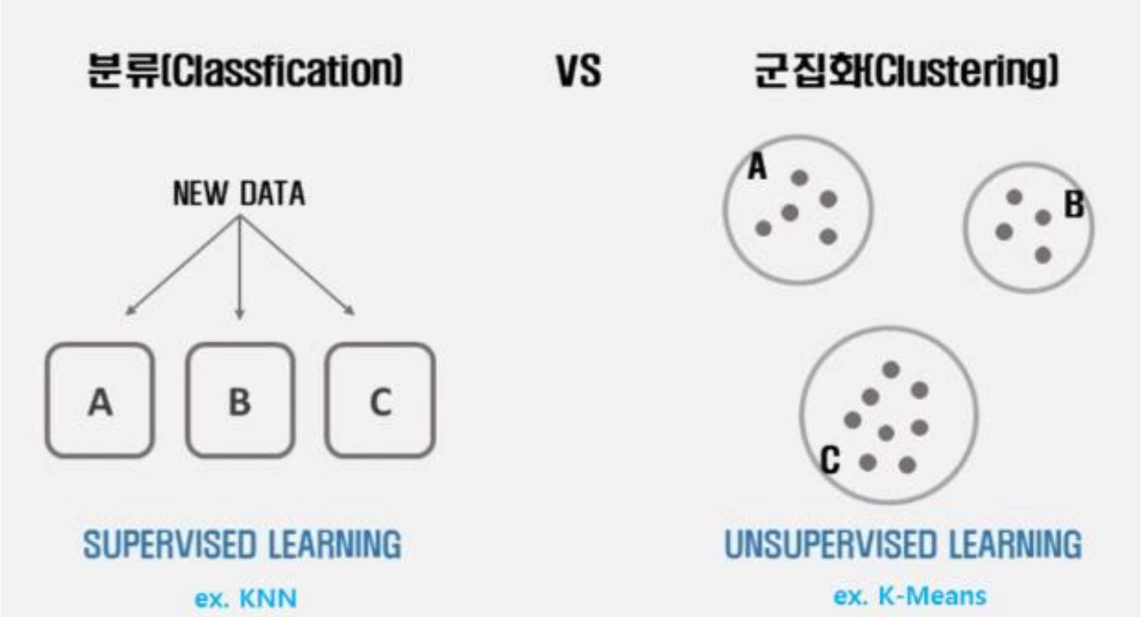
주성분 분석 (PCA)

- 차원의 저주
- 주성분 분석

1. 군집화 (Clustering)

[데이터 포인트들을 별개의 군집으로 그룹화하는 것]

- 유사성이 높은 데이터들을 동일한 그룹으로 분류하고 서로 다른 군집들이 상이성을 가지도록 그룹화한다.



(1) 군집화 활용 분야 :

- 고객, 마켓, 브랜드, 사회 경제 활동 세분화 (Segmentation)
- 이미지 검출, 세분화, tracking
- 이상 검출, outlier 검출

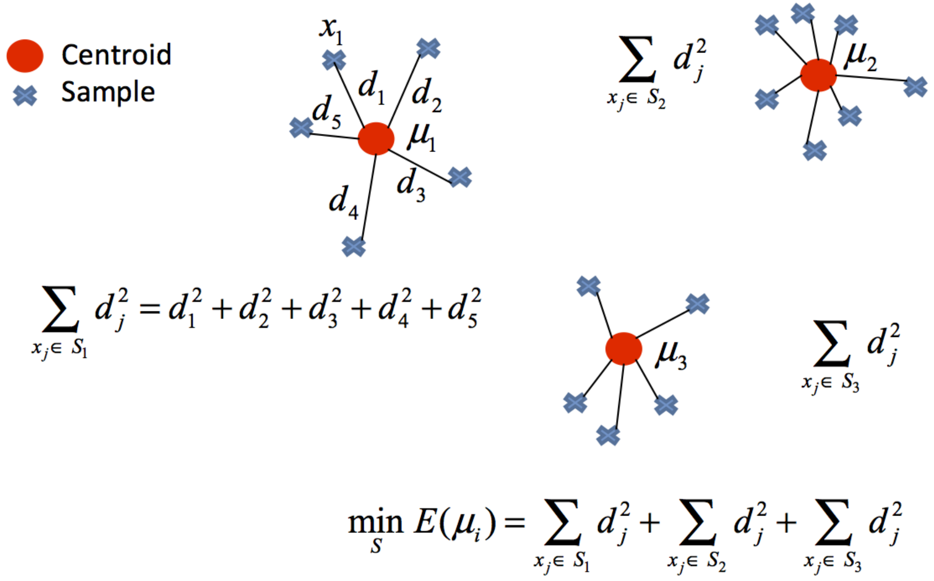
(2) 군집화 종류 :

- K-Means, Mean Shift, Gaussian Mixture Model, DBScan

1. 군집화 (Clustering)

[K-Means]

- 중심점 (centroid)과 데이터의 '유클리드 거리' 기반으로 하며, 거리 차이의 분산을 최소화하는 것을 목표로 한다.



[K-Means 과정]

1. 클러스터의 개수를 결정한다.
2. 초기 centroid를 선택한다.
3. 모든 데이터를 순회하면서 각 데이터마다 가장 가까운 centroid가 속해 있는 cluster로 assign한다.
4. Centroid를 cluster의 중심으로 이동한다.
5. Cluster에 assign 되는 데이터가 없을 때까지 3, 4번을 반복한다.

1. 군집화 (Clustering)

[K-Means]

- 중심점 (centroid)과 데이터의 '유클리드 거리' 기반으로 하며, 거리 차이의 분산을 최소화하는 것을 목표로 한다.

[K-Means 과정]

1. 클러스터의 개수를 결정한다.
2. 초기 centroid를 선택한다.
3. 모든 데이터를 순회하면서 각 데이터마다 가장 가까운 centroid가 속해 있는 cluster로 assign한다.
4. Centroid를 cluster의 중심으로 이동한다.
5. Cluster에 assign 되는 데이터가 없을 때까지 3, 4번을 반복한다.

1. Rule of Thumb
- 가장 간단한 방법

$K \approx \sqrt{\frac{n}{2}}$ (n : 데이터의 개수)

2. Elbow Method
- 클러스터 개수를 순차적으로 늘려가면서 결과 모니터링.
- 하나의 클러스터를 추가했을 때 이전보다 더 나은 결과를 나타내지 않는다면, 이전의 클러스터의 수로 설정한다.

3. 정보 기준 접근법 (Information Criterion Approach)
- 클러스터링 모델에 대해 likelihood를 계산하는 것이 가능할 때 사용하는 방법.

1. 군집화 (Clustering)

[K-Means]

- 중심점 (centroid)과 데이터의 '유클리드 거리' 기반으로 하며, 거리 차이의 분산을 최소화하는 것을 목표로 한다.

[K-Means 과정]

1. 클러스터의 개수를 결정한다.
2. 초기 centroid를 선택한다.
3. 모든 데이터를 순회하면서 각 데이터마다 가장 가까운 centroid가 속해 있는 cluster로 assign한다.
4. Centroid를 cluster의 중심으로 이동한다.
5. Cluster에 assign 되는 데이터가 없을 때까지 3, 4번을 반복한다.

1. 랜덤하게 설정
2. 수동으로 설정
3. K-mean++ 방법

(1) 데이터 중 하나를 골라서 centroid로 설정한다.

(2) 나머지 데이터들과 거리를 계산한다.

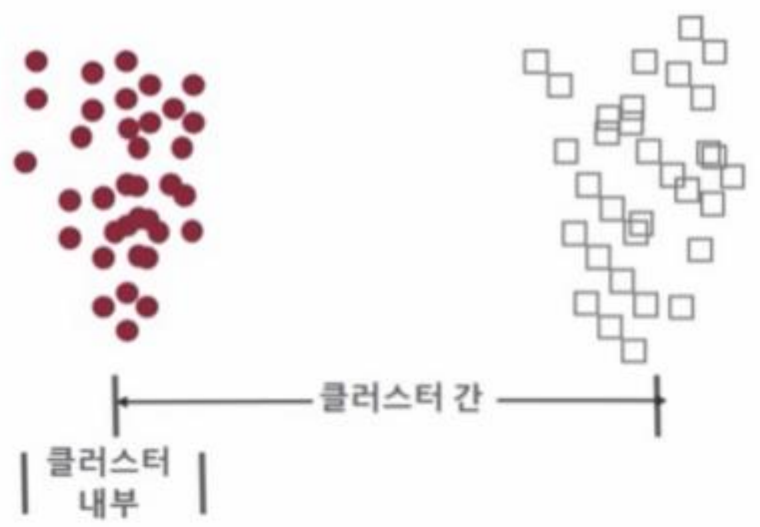
(3) (2)에서 구한 거리 비례 확률에 따라 선정한다. 즉, 거리가 멀어질수록 centroid가 될 확률이 높아진다.

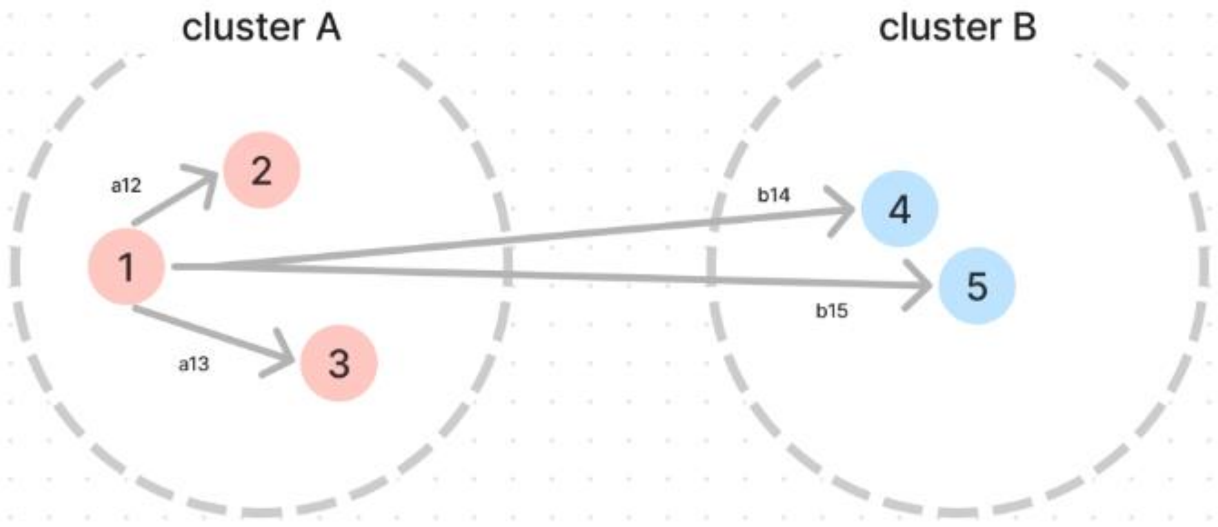
1. 군집화 (Clustering)

[K-Means]

K-Means 성능 평가 지표 : 실루엣 계수 (Silhouette Coefficient)

1. 해당 데이터가 같은 군집 내의 데이터와 얼마나 가깝게 군집화 되어 있는가?
2. 다른 군집에 있는 데이터와는 얼마나 멀리 분리되어 있는가?





a : 클러스터 내

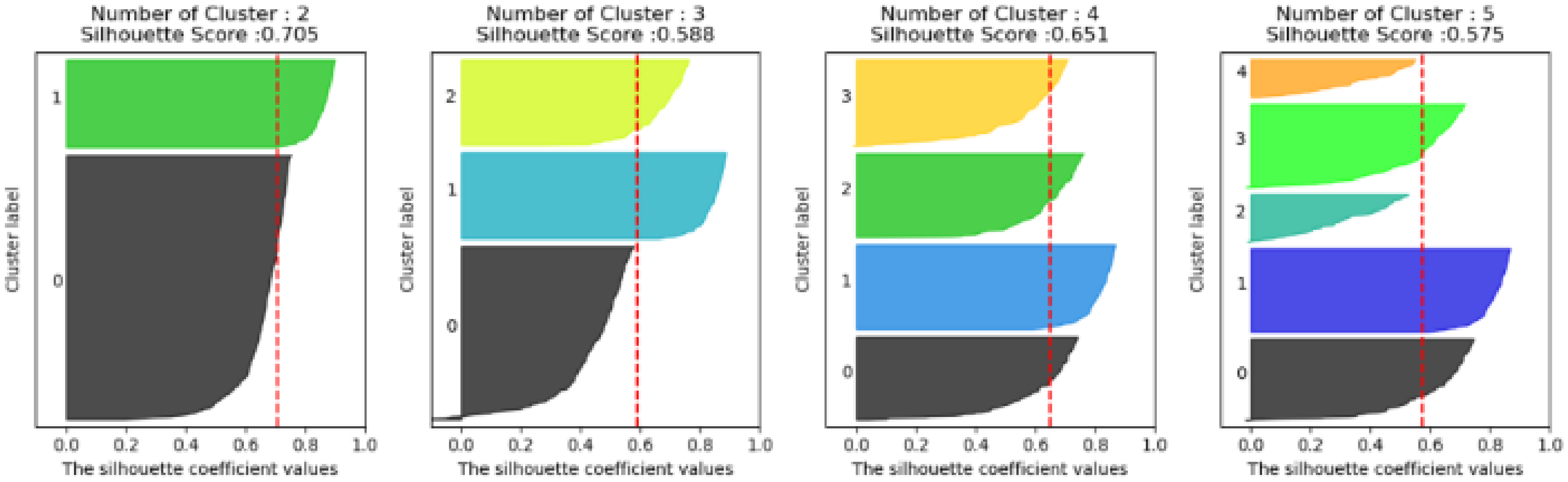
b : 타 클러스터

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

- 0 ~ 1 사이의 값을 가진다.
- 전체 실루엣 계수의 평균값과 개별 군집의 평균값의 편차가 크지 않아야 한다.
- 전체 실루엣 계수의 평균은 높지만 특정 군집의 실루엣 계수 평균만 유난히 높고 다른 군집들은 낮다면, 좋은 군집화가 아니다.

1. 군집화 (Clustering)

[K-Means]



K-Means 장점

1. 알고리즘이 쉽고 간결하다.
2. 대용량 데이터에서도 활용이 가능하다.

K-Means 단점

1. 거리 기반 알고리즘이다 보니, 속성의 개수가 너무 많을 경우 군집화 정확도가 떨어진다.
   (차원의 저주, PCA로 차원 축소 필요)
2. 반복을 수행하는데 반복 횟수가 많을 경우 수행 사간이 느려진다.
3. 이상치 데이터에 취약하다.

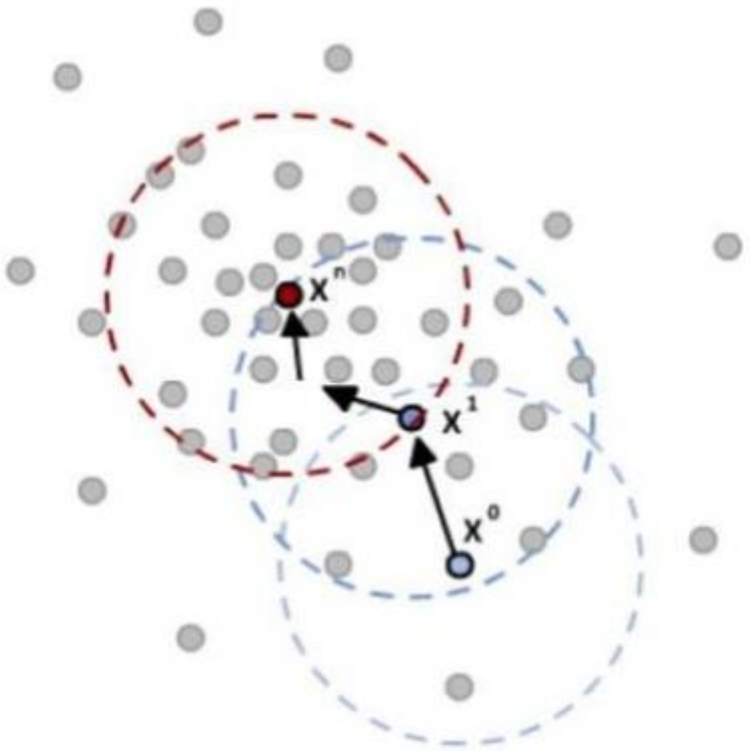
1. 군집화 (Clustering)

[Mean Shift]

- KDE (Kernel Density Estimation)를 이용하여 데이터 포인트들이 데이터 분포가 높은 곳으로 이동하면서 군집화를 수행한다.
- 별도의 군집화 개수를 지정하지 않으며 Mean Shift는 데이터 분포도에 기반하여 자동으로 군집화 개수를 정한다.

[Mean Shift 과정]

1. 개별 데이터의 특정 반경 내에 주변 데이터를 포함한 데이터 분포도 계산
2. 데이터 분포도가 높은 방향으로 중심점 이동
3. 중심점을 따라 해당 데이터 이동
4. 이동된 데이터의 특정 반경 내에 다시 데이터 분포 계산 후 2, 3 스텝을 반복
5. 가장 분포도가 높은 곳으로 이동하면 더 이상 해당 데이터는 움직이지 않고 수렴
6. 모든 데이터를 1 ~ 5까지 수행하면서 군집 중심점을 찾음



1. 군집화 (Clustering)

[Mean Shift]

[확률 밀도 추정 방법]

1. 모수적 (Parametric) 추정

- 데이터가 특정 데이터 분포 (예. 가우시안 분포)를 따른다는 가정 하에 데이터 분포를 찾는 방법
- 예) Gaussian Mixture Model

2. 비모수적 (Non-Parametric) 추정

- 데이터가 특정 분포를 따르지 않는다는 가정 하에 밀도를 추정하는 방법
- 관측된 데이터만으로 확률 밀도를 찾는 방법
- 예) 히스토그램, KDE

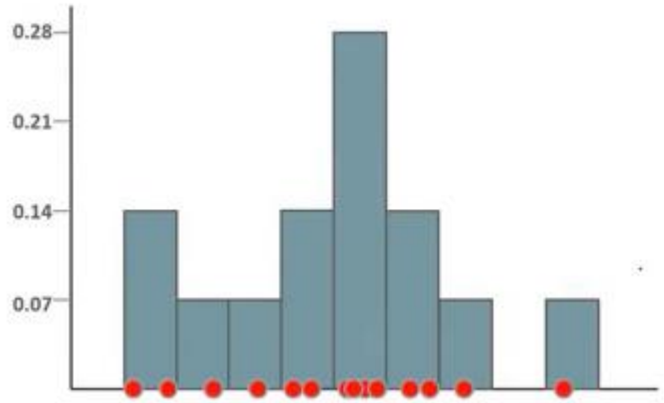
1. 군집화 (Clustering)

[Mean Shift]

[확률 밀도 추정 방법]

1. 모수적 (Parametric) 추정
- 데이터가 특정 데이터 분포 (예. 가우시안 분포)를 따른다는 가정 하에 데이터 분포를 찾는 방법
- 예) Gaussian Mixture Model

2. 비모수적 (Non-Parametric) 추정
- 데이터가 특정 분포를 따르지 않는다는 가정 하에 밀도를 추정하는 방법
- 관측된 데이터만으로 확률 밀도를 찾는 방법
- 예) 히스토그램, KDE



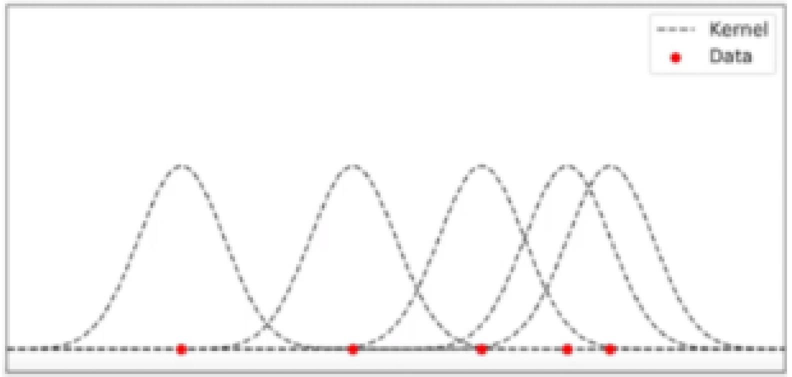
- bin의 경계에서 불연속성이 나타남
- bin의 크기에 따라서 히스토그램이 달라짐

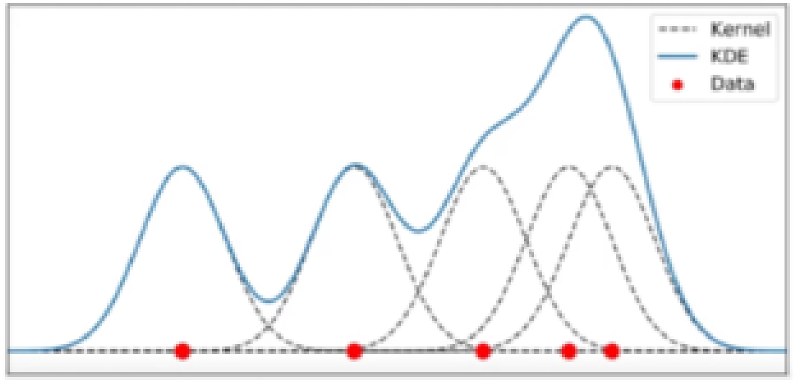
1. 군집화 (Clustering)

[Mean Shift]

[KDE (Kernel Density Estimation)]

- 개별 관측 데이터들에 커널 함수를 적용한 뒤, 커널 함수들의 적용 값을 모두 합한 뒤에 개별 관측 데이터의 건수로 나누어서 확률 밀도 함수를 추정하는 방식.
- 커널 함수로는 대표적으로 Gaussian 분포함수가 사용됨



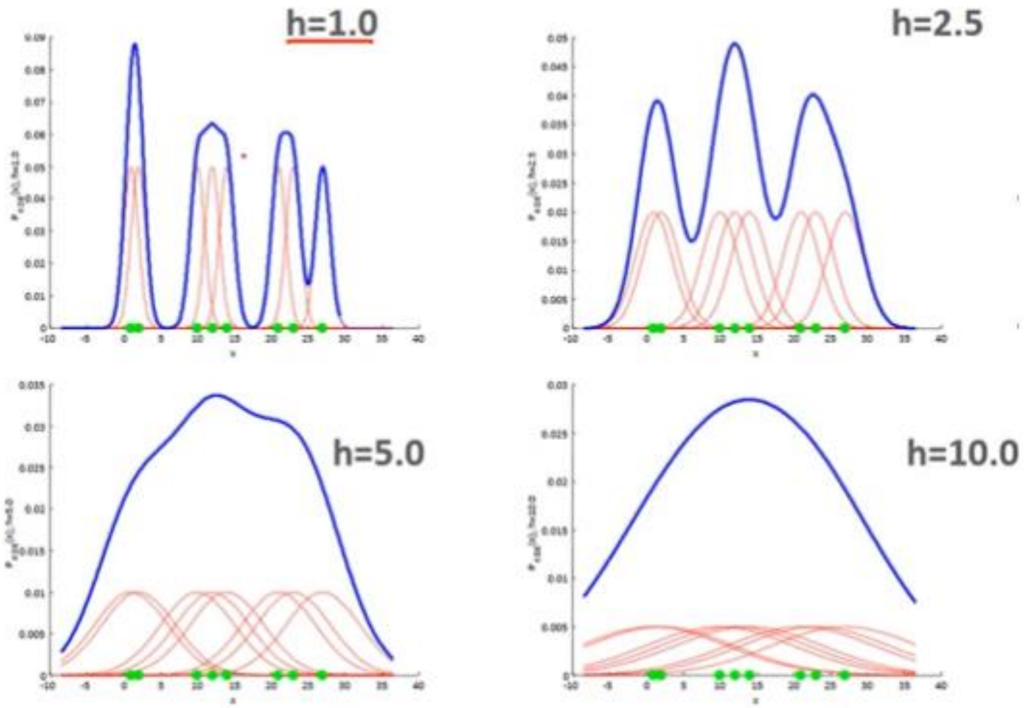


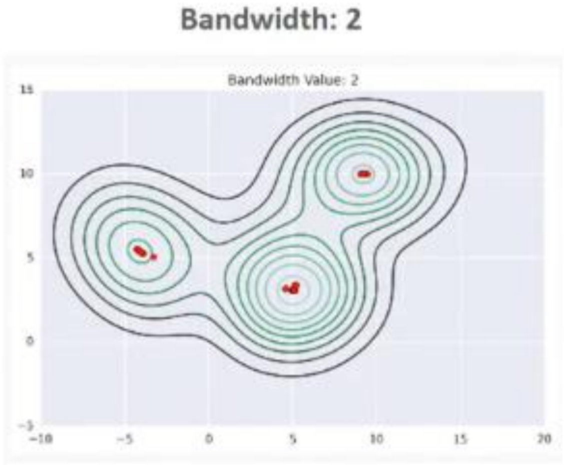
1. 군집화 (Clustering)

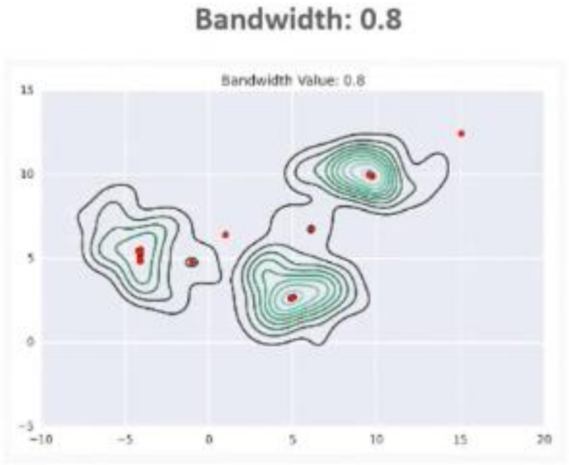
[Mean Shift]

- Mean Shift는 군집의 개수를 지정하지 않는다. 오직 Bandwidth의 크기에 따라 군집화를 수행한다.

(Bandwidth : kernel 함수의 뾰족한 정도로, 클수록 완만하고 단순화된 PDF를 추정한다.)







1. 군집화 (Clustering)

[GMM (Gaussian Mixture Model)]

[확률 밀도 추정 방법]

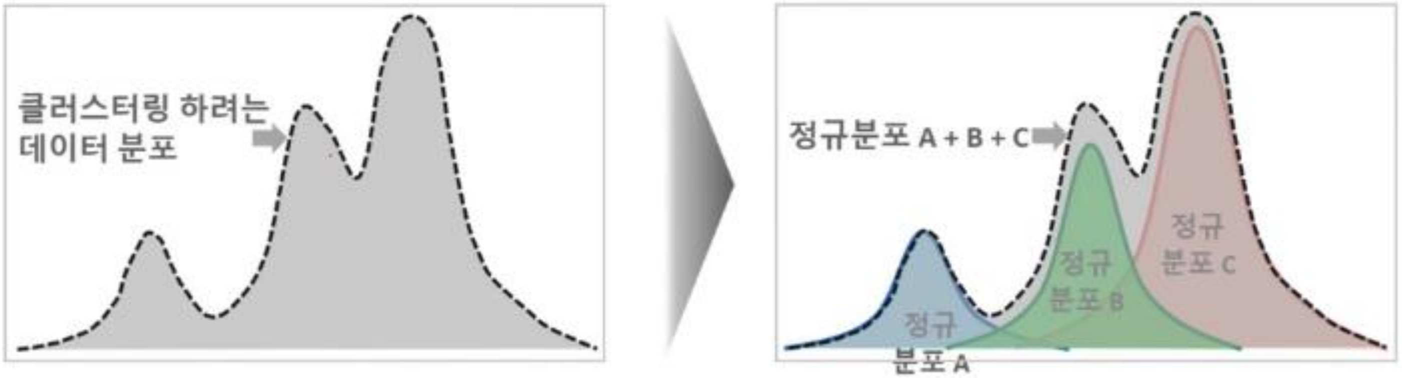
1. 모수적 (Parametric) 추정
- 데이터가 특정 데이터 분포 (예. 가우시안 분포)를 따른다는 가정 하에 데이터 분포를 찾는 방법
- 예) Gaussian Mixture Model

2. 비모수적 (Non-Parametric) 추정
- 데이터가 특정 분포를 따르지 않는다는 가정 하에 밀도를 추정하는 방법
- 관측된 데이터만으로 확률 밀도를 찾는 방법
- 예) 히스토그램, KDE

군집화를 적용하고자 하는 데이터가 여러 개의 다른 Gaussian 분포를 가지는 모델로 가정하고 군집화를 수행한다.

1. 군집화 (Clustering)

[GMM (Gaussian Mixture Model)]



- 개별 정규 분포들의 평균과 분산, 데이터가 특정 정규 분포에 해당될 확률을 추정한다.
- 예) 데이터 x에 대해 정규분포 A에 속할 확률이 30%, B에 속할 확률 30%, C에 속할 확률 40%이라고 하자.

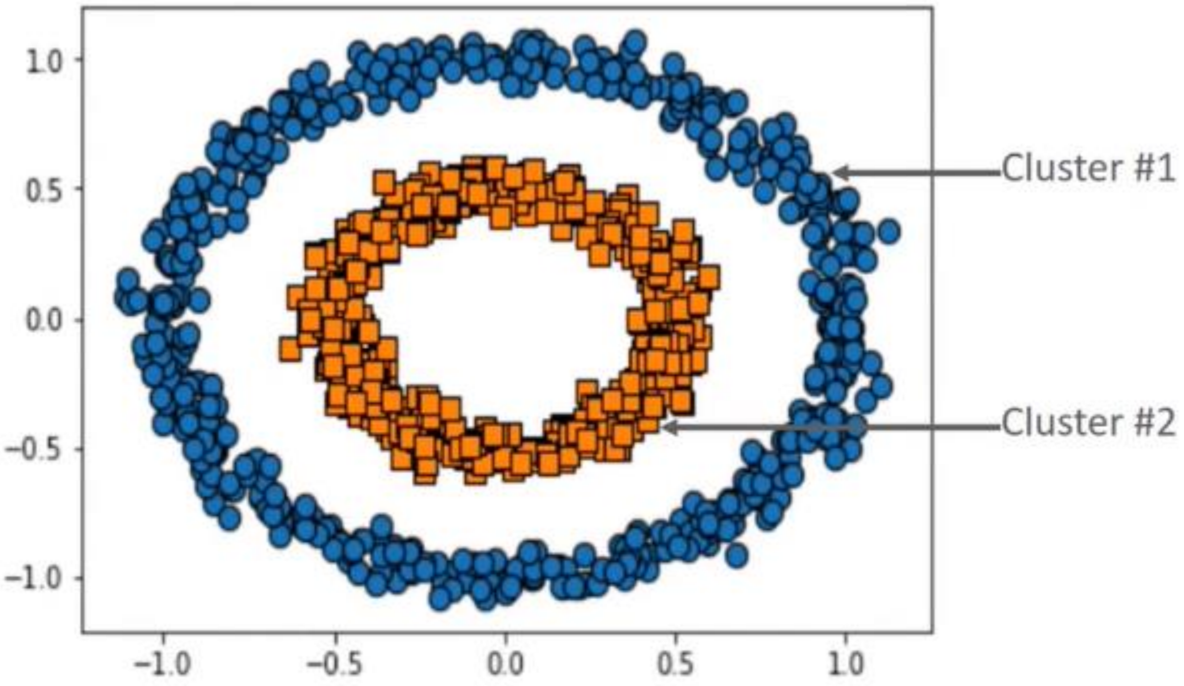
데이터 x는 정규분포 C에 속한다고 추정한다.

- GMM 과정 안에 MLE (Maximum Likelihood Estimation)이 포함된다.
- 모수 추정 후 어떤 분포에서 왔는지를 확률적으로 파악한다. 각 분포에 속하는지를 판단하는 클러스터링이 수행된다.

1. 군집화 (Clustering)

[DBSCAN]

- 특정 공간 내에 데이터 밀도 차이를 기반 알고리즘
- 복잡한 기하학적 분포도를 가진 데이터 세트에 대해서도 군집화를 잘 수행한다.
- 알고리즘이 데이터 밀도 차이를 자동으로 감지하며 군집을 생성하므로, 사용자가 군집 개수를 지정할 수 없다.



1. 입실론 (Epsilon)

2. 최소 데이터 개수 (min points)

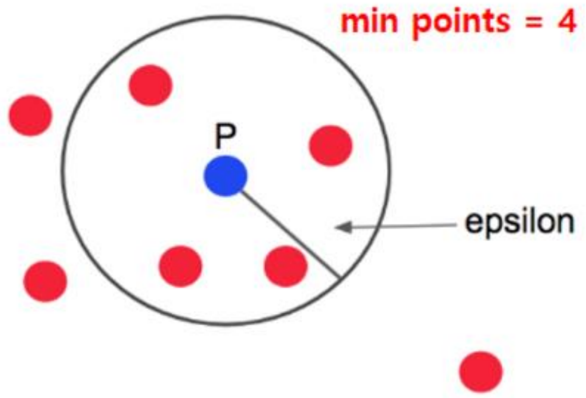
1. 군집화 (Clustering)

[DBSCAN]

점 p가 있다고 할 때,
점 p에서 부터 거리 e (epsilon) 내에 점이 m (min points)개 있으면 하나의 군집으로 인식한다고 하자.

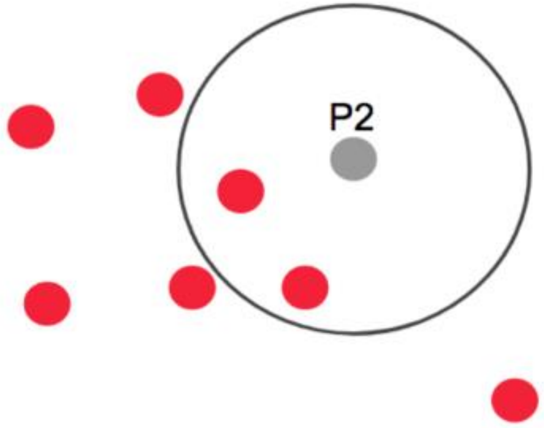
1. 핵심 포인트 (Core Point)

- 거리 e 내에 점 m개를 가지고 있는 점



2. 경계 포인트 (Border Point)

- 군집에는 속하지만, 스스로 core point가 안되는 점
- 점 P2를 기반으로 epsilon 반경 내의 점이 3개!
  min points = 4에 미치지 못하기 때문에 core point는 안된다.
  하지만, 점 P를 core point로 하는 군집에는 속한다.



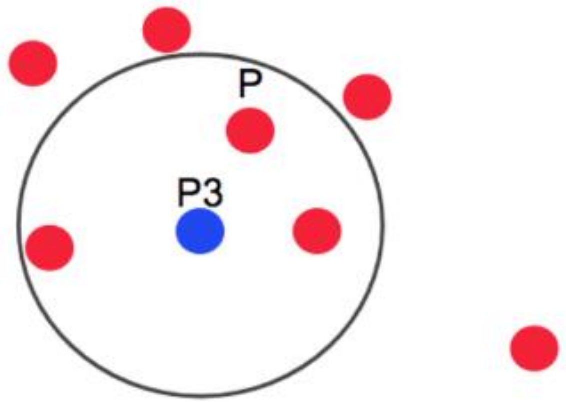
1. 군집화 (Clustering)

[DBSCAN]

점 p가 있다고 할 때,
점 p에서 부터 거리 e (epsilon) 내에 점이 m (min points)개 있으면 하나의 군집으로 인식한다고 하자.

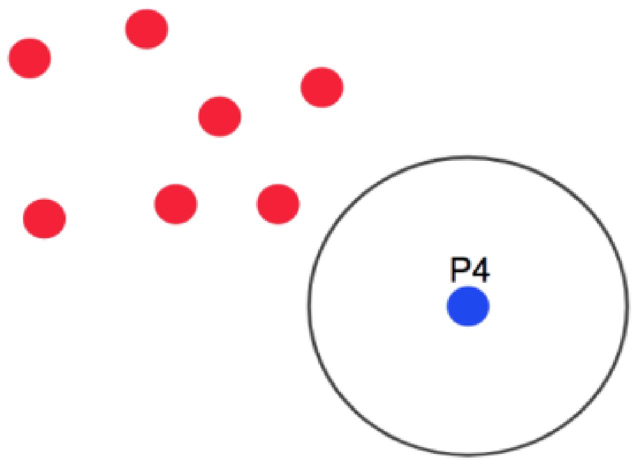
3. 이웃 포인트 (Neighbor Point)

- epsilon 반경 내에 위치한 타 데이터
- P3는 epsilon 반경내에 점 4개를 가지고 있기 때문에 core point



4. 잡음 포인트 (Noise Point)

- 최소 데이터 개수 이상의 이웃 포인트를 가지고 있지 않음



2 . 주성분 분석 (PCA)

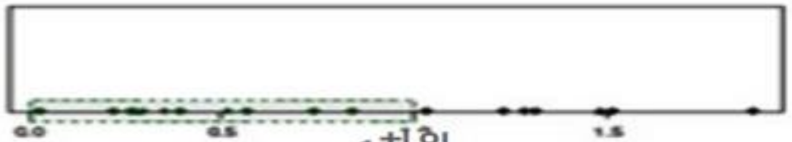
[차원의 저주]

차원이 커질수록?

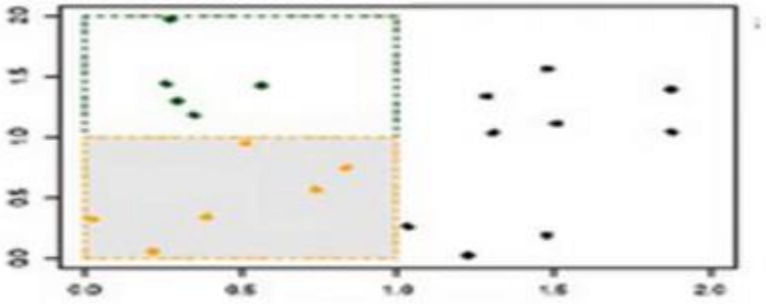
1. 데이터 포인트들간 거리가 크게 늘어남
2. 데이터가 희소화 (Sparse) 됨

차원이 커졌을 때의 문제점?

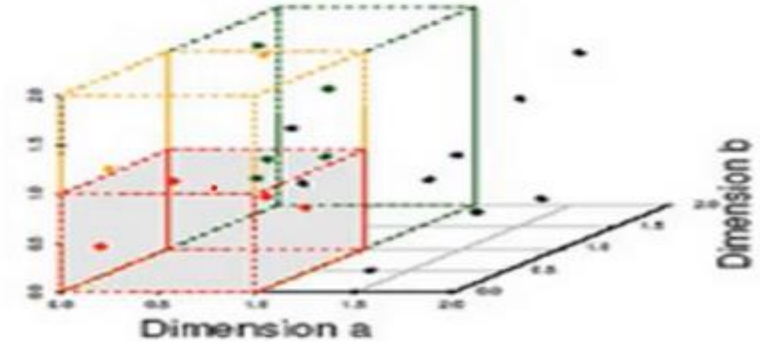
- 수백 ~ 수천 개 이상의 피처로 구성된 포인트들간 거리에 기반한 ML 알고리즘이 무력화됨
- 피처가 많을 경우 개별 피처 간에 상관관계가 높아 선형 회귀와 같은 모델에서는 다중 공선성 문제로 모델의 예측 성능이 저하될 가능성이 높음



1차원



특정 면적 공간에 6개의 데이터



특정 입체 공간에 4개의 데이터

2 . 주성분 분석 (PCA)

[차원 축소]

차원 축소의 의미

- 차원 축소는 단순히 데이터의 압축을 의미하는 것이 아니다.
- 더 중요한 의미는 차원 축소를 통해 좀 더 데이터를 잘 설명할 수 있는 잠재적 (Latent)인 요소를 추출하는 데에 있다.

활용 예시

1. 추천 엔진

2. 이미지 분류 및 변환

- 딥러닝 CNN이 나오기 전까지 가장 많이 쓰였던 방법으로 차원 축소가 있었다.
- 차원 축소 중에서도 주성분 분석 (PCA)를 많이 사용했음

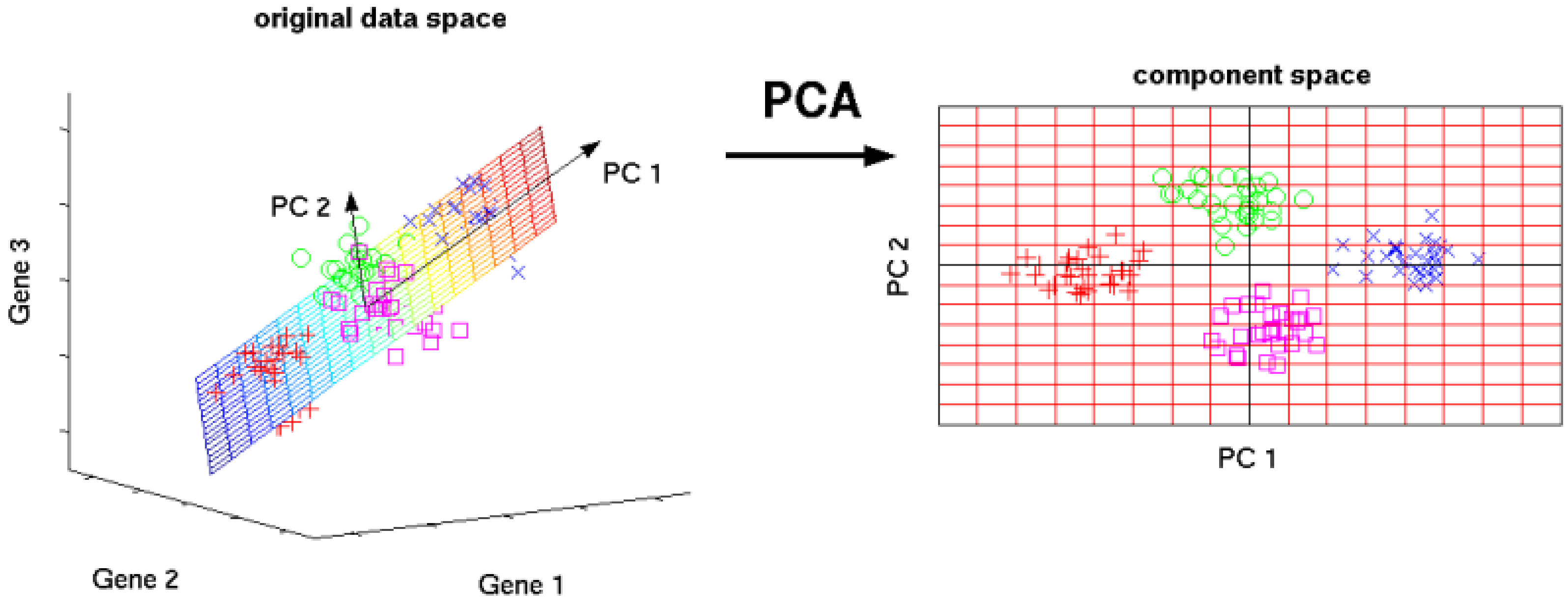
3. 문서 topic 모델링

- 수백장의 문서의 텍스트에 대해서 내용 압축

2. 주성분 분석 (PCA)

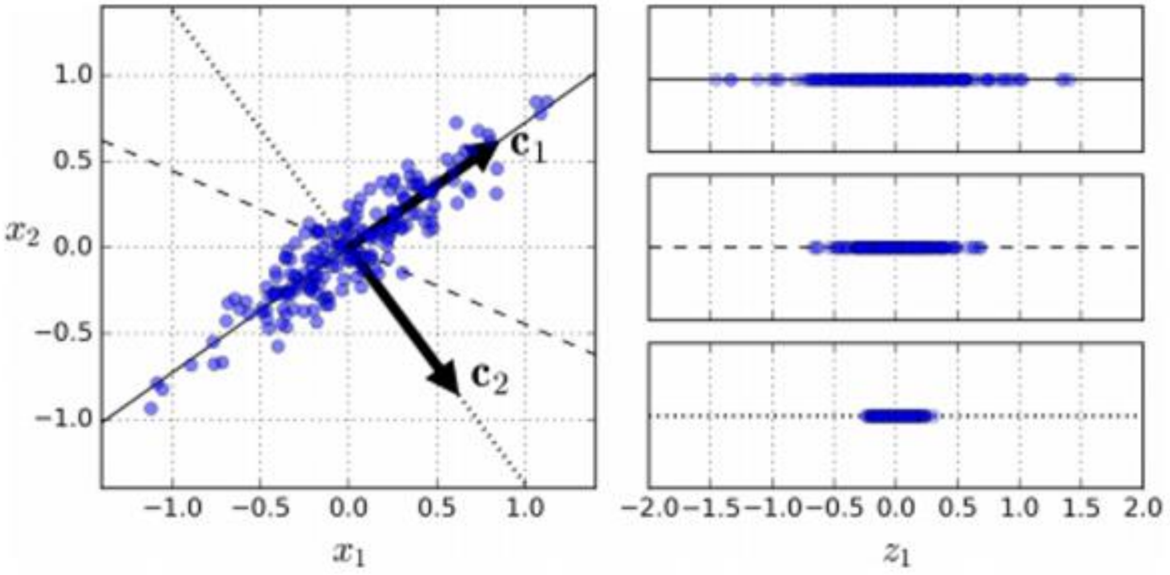
[PCA]

- 고차원의 원본 데이터를 저차원의 부분 공간으로 투영하여 데이터를 축소하는 기법
- 원본 데이터가 가지는 데이터 변동성을 가장 중요한 정보로 간주한다.
- 이 변동성에 기반한 원본 데이터 투영으로 차원을 축소
- 예) 10차원의 데이터를 2차원의 부분 공간으로 투영하여 데이터를 축소



2 . 주성분 분석 (PCA)

[PCA]



데이터 변동성

- 데이터의 분포를 나타냄에 있어 퍼진 정도를 측정
- 분산, 표준편차가 '얼마나 퍼져 있는지 ' 에 대한 척도!

1. 원본 데이터에 가장 큰 데이터 변동성을 기반으로 첫 번째 벡터 축을 생성한다.

2. 두 번째 축은 첫번째 축을 제외하고 그 다음으로 변동성이 큰 축을 설정한다.

(첫번째 축에 직각이 되는 벡터 (직교 벡터) 축으로, 다중 공선성을 해결한다.)

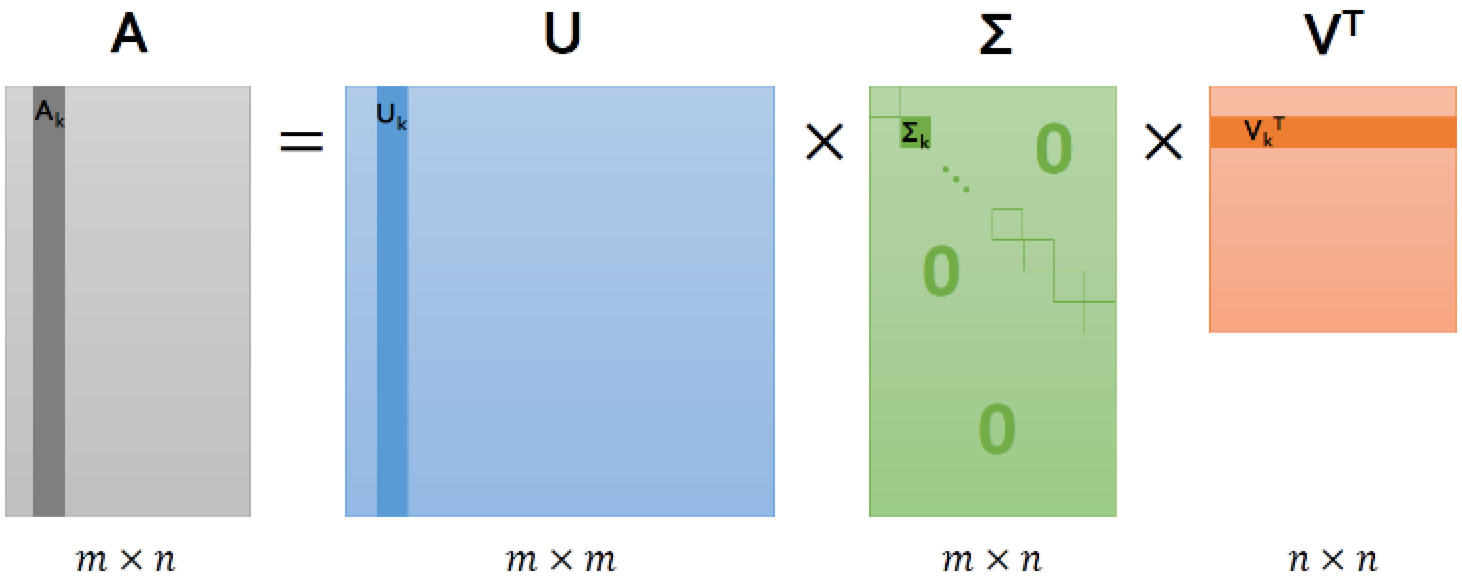
3. 세 번째 축은 다시 두 번째 축과 직각이 되는 벡터를 설정하는 방식으로 축을 생성한다.

2 . 주성분 분석 (PCA)

[PCA를 선형대수의 관점으로 해석하기]

1. 입력 데이터의 공분산 행렬 (Cov. Matrix)를 고유값 분해
2. 이렇게 구한 고유 벡터에 입력 데이터를 선형 변환

$$A = U\Sigma V^T$$



< 특이값 분해 >

2 . 주성분 분석 (PCA)

[PCA를 선형대수의 관점으로 해석하기]

1. 원본 데이터의 공분산 행렬 추출

2. 공분산 행렬을 고유벡터와 고유값 분해

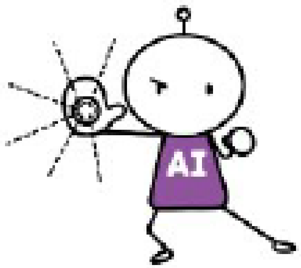
- 고유벡터는 PCA의 주성분 벡터로서 입력 데이터의 분산이 큰 방향을 나타낸다.
- 고유값 (eigenvalue)은 바로 이 고유벡터의 크기를 나타내며, 동시에 입력 데이터의 분산을 나타낸다.

3. 원본 데이터를 고유 벡터로 선형 변환

4. PCA 변환값 도출

Undergraduate Research Internship in Affective AI LAB.

감사합니다 :>



컴퓨터공학과 12201682 공지윤