<YDMS 2023-1> 2주차 과제

비지도학습

주하연

비지도 학습은 차원축소와 군집화로 나누어진다.

1. 차원축소개념 – (차원의저주, 고유벡터, 고윳값)

데이터에서의 차원 = 변수의 수

차원이 크면 시각화가 어렵고, 이해하기 어렵고, 분석하기도 어렵다. = 차원의 저주

- 고차원 데이터일수록 모델을 학습하기가 훨씬 더 어려워지고 훨씬 더 많은 데이터 양이 필요하다. 유용한 모델을 세우는 데는 차원의 크기가 오히려 걸림돌이 되기도 한다. (과적합으로 이어지게 된다)

  -> 결론 : 데이터의 용량이 커지면 성공적 모델의 학습을 저해할 수 있다.

따라서 고차원 데이터를 저차원 데이터로 변환하는 기술을 차원 축소라고 한다.

A. 차원축소의 목적

   고차원에 있는 원래 데이터를 유용한 정보의 손실을 최소화 하면서 더 낮은 차원으로 데이터를 대응 시키는 것

B. 차원축소를 하는 이유

   i. 데이터 시각화 : 고차원 데이터는 시각화가 어렵기 때문에 2차원 또는 3차원으로 시각화해서 데이터의 구조와 패턴을 파악할 수 있다.

   ii. 계산비용절감 : 고차원 데이터를 분석할 때는 비용이 많이 드는데, 데이터의 크기를 줄이면서 계산 비용을 절감할 수 있다.

   iii. 노이즈 제거 : 고차원 데이터에서는 노이즈와 불필요한 정보가 많이 포함되어 있을 가능성이 높다. 차원 축소를 통해 불필요한 정보를 제거하고 데이터에서 중요한 정보를 추출할 수 있다.

   iv. 학습시간축소 : 더 적은 계산 학습 시간

   v. 자원 활용 효율성 향상 : 차원 축소를 통해 데이터의 차원을 줄이면, 저장 공간과 전송대역폭을 줄일 수 있다. 또한 모델의 학습속도도 높일 수 있다.

➔ 차원의 저주를 피하기 위해

C. 차원축소의 단점

i. 차원 축소로 인해 일부 데이터가 손실될 수 있다.

ii. PCA 차원 축소 기술에서 고려해야 할 주요 구성 요소를 알 수 없는 경우

D. 차원축소 방법

i. 특성 선택 (Feature Selection)

기존 데이터에서 몇 개의 특성만 선택하여 새로운 데이터를 만드는 방법

상관관계가 적을 때, 정보량이 많을 때, 노이즈가 적을 때

1. Filter Methods

특성 선택을 위한 첫 번째 단계로, 특성 간의 상관 관계를 측정하고, 각 특성의 중요도를 계산하는 방법

2. Wrapper Methods

모델의 성능을 최적화하기 위해, 모델의 학습 결과를 이용하여 특성을 선택하거나 제거하는 방법 (ex. 재귀적 특성 제거)

3. Embedded Methods

모델 학습 과정에서 특성 선택을 수행하는 방법이다. 모델 내부에서 특성의 중요도를 계산하여 특성을 선택하거나 제거한다. (ex. 랜덤 포레스트, LASSO)

- 장점 : 모델이 더 간단해지며, 과적합을 방지할 수 있다. 또한 모델이 더 빠르게 학습될 수 있다.
- 단점 : 모든 특성이 중요하지 않다는 가정에 따라, 중요하지 않은 특성이 실제로 유용한 정보를 제공할 수 있다는 가능성이 있다. 또한, 선택된 특성들 간의 상호작용을 고려하지 않는다는 한계도 있다.

ii. 특성 추출 (Feature Extraction)

기존 데이터에서 새로운 특성을 추출하여 새로운 데이터를 만드는 방법

1. 주성분 분석 (PCA)

데이터의 분산을 최대한 보존하면서 서로 직교하는 축을 찾아 고차원의 공간의 표본을 저 차원으로 변환

A. 데이터 표준화

B. 공분산 행렬 구하기

C. 공분산 행렬의 고윳값(주성분)과 고유벡터 구하기

i. 고윳값 : 분산의 크기를 알 수 있다.

ii. 고유벡터 : 각 변수의 분산 방향을 알 수 있다.

D. 고윳값을 큰 순서대로 내림차순 정렬

E. 크기 순서대로 구하고 싶은 차원만큼의 고윳값 추출

F. 선정한 고윳값에 대응하는 고유벡터로 새로운 행렬 생성

G. 새로운 행렬에 새롭게 입력 데이터를 반환

$$C = P \sum P^T$$

C=공분산행렬, P=n*n 직교행렬, Σ=n*n 정방행렬

$$C = [e_1 \cdots e_n] \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \cdots & \cdots & \cdots \\ 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} e_1^t \\ \cdots \\ e_n^t \end{bmatrix}$$

e=고유벡터(Ax=ax,행렬A를 곱하더라도 방향이 변하지 않고 그 크기만 변하는 벡터), ↘=고유벡터의 크기

공분산 = 고유벡터 직교 행렬 * 고윳값 정방 행렬 * 고유벡터 직교 행렬의 전치 행렬

2. 선형 판별 분석 (LDA)

주어진 데이터를 가장 잘 구분할 수 있는 축을 찾는 방법 중 하나이다. 클래스 간의 차이를 최대화하고 클래스 내부의 분산을 최소화하여 데이터를 분리하는 데에 적용된다.

분류문제에서 PCA보다 좋은 성능을 보이고, 작은 데이터셋에서 효과적

Ex) 얼굴 인식, 의료진단, 자연어처리

A. 클래스 내부와 클래스 간 분산 행렬을 구한다. (이 행렬들은 입력 데

이터의 결정 값 클래스별로 개별 피처의 평균백터를 기반으로 구함)

B. 클래스 내부 분산 행렬을 Sw, 클래스 간 분산 행렬을 Sb라고 하면 두 행렬을 고유벡터로 분해할 수 있다.

C. 고윳값을 큰 순서대로 내림차순 정렬

D. 크기 순서대로 구하고 싶은(변환 차수만큼) 고윳값 추출

E. 선정한 고윳값에 대응하는 고유벡터로 새로운 행렬 생성

F. 새로운 행렬에 새롭게 입력 데이터를 반환

$$S_W^T S_B = \begin{bmatrix} e_1 & \cdots & e_n \end{bmatrix} \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \cdots & \cdots & \cdots \\ 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} e_1^T \\ \cdots \\ e_n^T \end{bmatrix}$$

3. t-SNE

   복잡한 데이터의 시각화 용도로 많이 쓰임

4. SVD (Singular Value Decomposition)

   PCA와 LDA가 정방행렬에 대한 작업이였다면, 이것은 일반행렬에 대한 방법이다. 행과 열의 크기가 다른 행렬에도 적용 가능

- 장점 : 데이터의 복잡도를 낮추고, 모델의 성능을 향상 시킬 수 있다. 또한, 데이터의 특성을 더욱 명확하게 이해할 수 있다.
- 단점 : 추출된 특징들이 원래 데이터의 정보를 완벽하게 대표하지 못할 수 있다. 또한, 추가적인 계산이 필요할 수 있다.

2. 군집분석개념

군집분석은 데이터를 사전에 분류하거나, 예측을 하지 않고 단순히 데이터의 유사성에 따라 그룹으로 나누는 것이 목적이다. 다른 그룹끼리는 서로 다른 특성을 가지고 있으며, 동일한 그룹 내에서는 비슷한 특성을 공유하고 있다.

군집분석은 주로 비즈니스, 마케팅, 의학, 생물학 등의 분야에서 활용된다. ex) 고객을 유사한 특성으로 가진 그룹으로 나누어 마케팅 전략을 수립

* 군집 분석의 목적

1. 데이터 탐색 : 각 군집을 유사한 특성을 가지고 있기 때문에, 군집에 속한 데이터들은 서로 비슷한 패턴을 보이고 우리는 이를 시각적으로 파악할 수 있다.

2. 데이터 분류 : 군집 분석을 통해 비슷한 특성을 가진 데이터를 같은 군집으로 분류할 수 있다. 이를 통해, 데이터를 분류하여 보다 효율적인 분석이 가능해진다.

3. 데이터 전처리 : 군집 분석을 통해 데이터를 군집별로 분류하면, 각 군집을 대표하는 중심점을 찾을 수 있다. 이 중심점을 기반으로 데이터를 전처리하여 데이터의 차원을 축소하거나 이상치를 제거하는 등의 데이터 처리 작업을 할 수 있다.

4. 예측 모델링 : 군집 분석은 예측 모델링의 기반을 제공할 수 있다. 군집 분석을 통해 데이터의 구조를 파악하고, 각 군집의 특성을 분석하여 예측 모델링에 활용할 수 있다.

A. 계층적 군집분석

데이터 간의 유사성을 이용하여 계층적으로 클러스터를 형성

클러스터 간의 거리를 기준으로 클러스터를 분할하는 방식

i. 병합적 군집분석(주로 사용)

개체 간 거리가 가까운 개체끼리 순차적으로 묶어주는 방법(전체 데이터를 하나의 클러스터로 만드는 방법)

중요한 요소 : 거리 측정 방법, 병합 기준

1. 거리 측정 방법

A. 유클리드 거리

B. 맨하탄 거리

C. 코사인 유사도

2. 병합기준

A. 최단결합법

B. 최장결합법

C. 평균결합법 ...etc

* 장점 : 계층적이므로, 클러스터 간의 유사성을 시각적으로 파악할 수 있다. 또한, 클러스터링 결과를 다양한 수준으로 분할할 수 있어 다양한 해석이 가능

* 단점 : 계산량이 많아 대규모 데이터에서는 계산 시간이 오래걸리고, 어떤

개체가 특정한 군집에 할당되면 다른 군집에 다시 할당될 수 없다. 그리고 덴드로그램으롭 분석하기 어렵다.

ii. 분할적 군집분석

개체 간 거리가 먼 개체끼리 나누어 가는 방법

클러스터를 분할하기 위해 분할 기준을 정해야 한다. 이 분할 기준은 각 클러스터의 내부 분산과 각 클러스터간 거리를 고려하여 결정된다.

* 장점 : 전체 데이터를 한 번에 처리하지 않고, 순차적으로 처리하기 때문에 계산량이 적어 병합적 군집분석에 비해 대규모 데이터에서도 적용이 가능하다.

* 단점 : 클러스터 간의 관계를 파악하기 어려워 해석에 어려움이 있다.

B. 비계층적 군집분석

군집을 형성할 때 계층적으로 형성하지 않고, 그룹화를 할 유사도 측정 방식에 따라 최적의 그룹(cluster)을 계속적으로 찾아나가는 방법

* 장점 : 계층 구조에 구애받지 않는다. 또한 계산 비용이 낮고 빠르기 때문에 대용량 데이터셋에도 적용이 가능하고 높은 효율성을 보인다. 해석하기 쉽다.

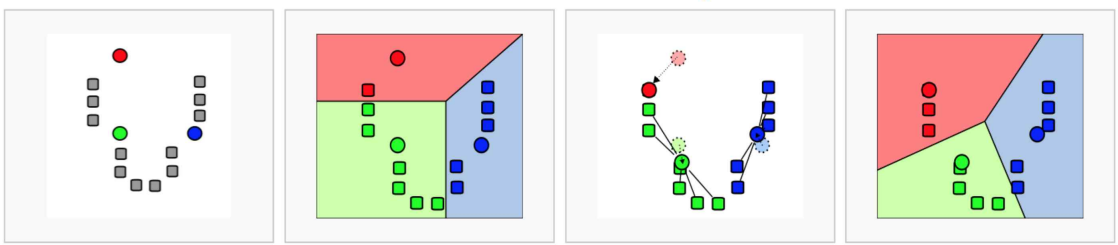
* 단점 : 초기값 설정에 민감하다. 그리고 이상치에 민감하기 때문에 이상치가 있는 데이터셋에서는 결과가 왜곡될 수 있다.

i. K-means (중심 기반)

데이터들을 k개의 그룹으로 나누는 방법이다. 각 그룹은 하나의 중심점을 가지며, 이 중심점은 해당 그룹의 데이터들과 가장 가까운 위치에 있다.

가정 : 유사한 데이터는 중심점을 기반으로 분포할 것이다.

Demonstration of the standard algorithm



1. *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).

2. *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.

3. The centroid of each of the *k* clusters becomes the new mean.

4. Steps 2 and 3 are repeated until convergence has been reached.

1. 자료를 K개 초기 군집으로 나눈다. (초기점 K 설정)

   K는 중심점이자, 묶일 그룹(cluster)의 수와 같다.

2. 그룹(cluster) 부여

   k개의 중심점과 개별 데이터간의 거리를 측정한다.

   가장 가까운 중심점으로 데이터를 부여한다.

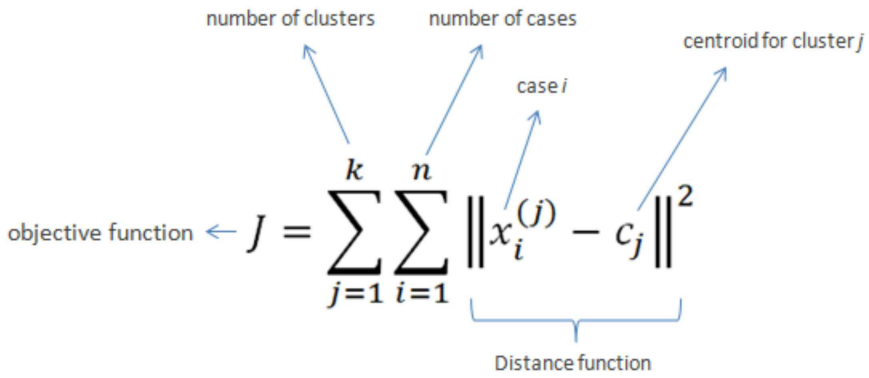
3. 중심점 업데이트

   할당된 데이터들의 평균값으로 새로운 중심점을 업데이트한다.

4. 최적화

   2,3번 작업을 반복적으로 수행한다.

   중심점에 변화가 없으면 작업을 중단한다.



목표 : 위의 목적함수를 최소화

* 문제점

- 초기 중심점을 무작위로 설정하기 때문에 결과가 초기 값에 따라 크게 영향을 받을 수 있다.

- 사전에 K개의 군집으로 되어 있다는 것을 알아도, 크기가 작은 그룹에서는 제대로 작동하지 않을 수 있다. -> 군집의 사이즈나 모양이 다를 경우 애매한 결과가 나올 수 있다.

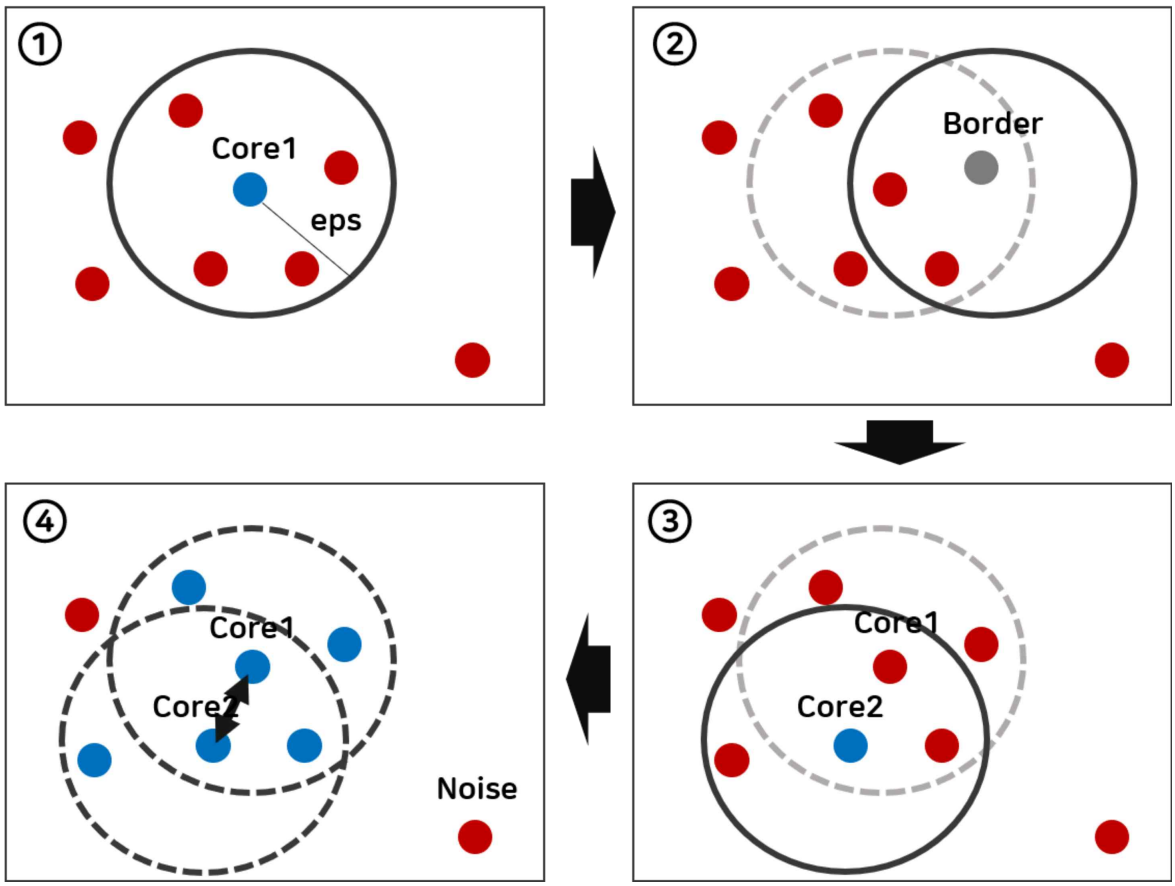
- 차원이 많아지는 경우, 유사도 측정에 어려움을 겪을 수 있다.

- 이상치에 민감하다. K-means 는 중심점과 가까운 데이터들만 클러스터링하기 때문에, 이상치를 제거하거나 다른 기법을 사용하는 것이 좋다.

ii. DBSCAN (밀도 기반)

밀도가 높은 지역을 클러스터로 구분하고, 밀도가 낮은 지역은 잡음으로 처리한다. 각 중심에서 반경 eps 내에 있는 이웃 데이터의 개수가 MinPts 이상이면, 해당 데이터를 핵심 데이터로 분류하고 Cluster를 형성한다.

가정 : 유사한 데이터는 서로 근접하게 분포할 것이다.



1. 하나의 점(파란색)을 중심으로 반경(eps)네에 최소 점이 4개(minPts=4)이상 있으면, 하나의 군집으로 판단하며 해당 점(파란색)은 Core가 된다.
2. 반경 내에 점이 3개 뿐이므로 Core가 되진 않지만 Core1의 군집에 포함된 점으로, 이는 Border가 된다.
3. Border 데이터가 다른 Core의 이웃인 경우, 이를 Core가 속한 Cluster에 할당한다.
4. 그런데 반경 내의 점 중에 Core1이 포함되어 있어 군집을 연결하여 하나의 군집으로 묶인다.

3. 차원축소실습bostonhousing data

https://colab.research.google.com/drive/1P3hMTDGD1tR98FM3J_WoJrZSTFtndrWe?usp=sharing

* Discussion

PCA를 수행할 때 축소할 차원을 그냥 무작위로 결정해 진행하였습니다. 숫자를 계속 바꿔서 해보다가 n개의 주성분이 95프로 이상을 설명하고 있을 때 그 개수로(저의 실습에서는 10개) fix 하였습니다. 그런데 데이터마다 적절한 주성분 개수가 있을텐데 이를 어떤 기준으로 어떻게 구해서 진행하는지 궁금합니다.

종속변수가 있는 데이터, 종속변수가 없는 데이터가 있는데 종속변수가 있는 데이터는 종속변수를 따로 분리한 다음 독립변수들끼리만 PCA를 진행해야하는지 궁금합니다.

4. 군집분석iris data로진행(+ framingham data)

-iris

https://colab.research.google.com/drive/1EGZipdMybhZemCNy2LDWNdEKFoiuh9Gk?usp=sharing

-framingham

https://colab.research.google.com/drive/1iVL9tNIZ4DAftTzMoHTldoxfEJquZ5jO?usp=sharing

* Discussion

범주형자료들이 Kmeans하는데 오류가 생겨서 그냥 수치형 자료로만 진행하였습니다. 이럴때는 범주형 자료들을 0,1 이런식의 자료로 꼭 바꿔줘서 진행해야 하나요?