이상탐지 A to Z

5편.

차원 축소 – 비지도 학습 기반의 머신러닝 기법(2)

데이크루 2기 Team Zoo

DACON
DATA TO VALUE

목차

1. 차원 축소 기법
2. PCA
3. Isomap
4. MDS

1.차원 축소 기법

-차원 축소의 의미

- 비지도학습의 큰 축
- 원본 데이터를 저차원의 부분공간으로 투영하여 데이터 축소하는 기법
- 10차원 이상의 데이터가 주어진 경우, 2-3차원 데이터 부분 공간으로 투영하여 축소

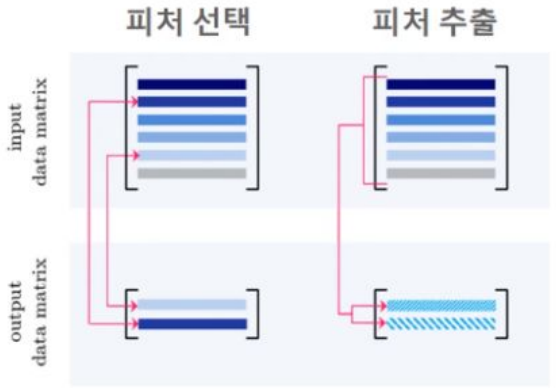
-차원 축소 하는 이유

- 차원이 증가할수록 데이터 포인트 거리가 기하급수적으로 멀어진다.
- 희소한(sparse) 구조이다.
- 피처 증가시 피처간 상관관계가 높아져 다중공선성 문제 과적합이나 예측
- 성능이 저하될 우려가 심하다

1.차원 축소 기법

-차원 축소 기법의 결과

- 기법 적용시, 다차원 피처를 피처 수로 줄이다 보면 직관적 해석 가능
- 학습 시간 감소(데이터 크기가 감소하였기 때문)
- 시각적 인지 가능(기본적으로 3차원데이터까지 시각적 구현이 가능한데, 차원 축소 기법으로 해당 결과에 대해 변수 간 차이 확연하게 구분 가능)



-차원 축소 방법

1.피처 선택(feature selection):

- 특정 피처에 종속성이 강한 불필요한 피처 제거한 후 특징 잘 살릴 수 있는 피처를 선택하는 방법이다.

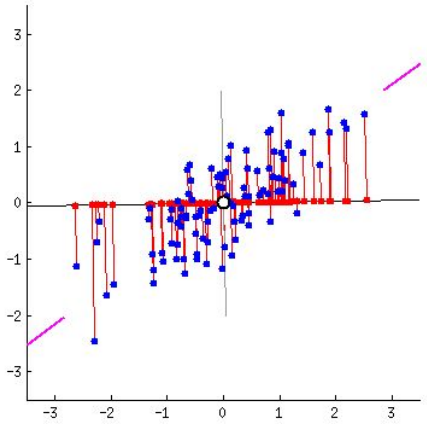
2.피처 추출(feature extraction):

- 기존 피처를 저차원으로 압축시켜 추출하는 방법으로, 새롭게 추출이 된 중요 특성은 기존 피처가 압축된 것이지만, 기존 피처와 다르게 새로운 값이 된다.
- 함축적 설명이 가능
- **또 다른 공간으로 매핑하여 추출한다는 장점**이 존재

2.PCA

1.PCA기법

고차원 공간의 표본들을 선형 연관성이 없는 저차원 공간으로 변환하는 기법이다.
분산을 최대한 보존하면서 서로 직교하는 축을 찾아 변환해야한다.



(한 주성분에 대해 직교하는 사선추기 분산을 최대한 보존할 수 있는 새 기저이다.)

2.PCA

2.변수추출

pca 기법은 변수 추출 방식을 사용. 변수 선택과 대비되는 기법이다.

- 변수 선택(feature selection):단순 일부 중요 변수 빼내기
- 변수 추출(feature extraction): 기존 변수 조합해 새 변수 만들기

변수 추출 방식 :

1) 기존 변수 가운데 일부만 활용
2) 기존 변수 모두 활용 - >PCA기법

-활용 방식:

선형결합하여 새 변수 만듦

선형 변환은 선형 변환으로 이해되는데 벡터가 새로운 축에 사영시킨 결과물이라는 의미에서 선형 변환이라고 한다.

(선형 변환: 선형성을 가지는 삼수로 벡터 공간의 성질을 보존하면서 벡터공간에서 벡터 공간으로 가는 준동형 사상이다

$$\vec{z_1} = \alpha_{11}\vec{x_1} + \alpha_{12}\vec{x_2} + \ldots + \alpha_{1p}\vec{x_p} = \vec{\alpha_1}^T X$$
$$\vec{z_2} = \alpha_{21}\vec{x_1} + \alpha_{22}\vec{x_2} + \ldots + \alpha_{2p}\vec{x_p} = \vec{\alpha_2}^T X$$
$$\cdots$$
$$\vec{z_p} = \alpha_{p1}\vec{x_1} + \alpha_{p2}\vec{x_2} + \ldots + \alpha_{pp}\vec{x_p} = \vec{\alpha_p}^T X$$

2.PCA

3.PCA의 목적과 solution

PCA의 목적: 원데이터 행렬 X 의 분산을 최대한 보존하는 동시 저차원 공간으로 변환
z의 분산 역시 최대화 되어야한다.

$$\begin{aligned}\max_{\alpha}\{Var(Z)\} &= \max_{\alpha}\{Var(\vec{\alpha}^{T}X)\} \\ &= \max_{\alpha}\{\vec{\alpha}^{T}Var(X)\vec{\alpha}\} \\ &= \max_{\alpha}\{\vec{\alpha}^{T}\Sigma\vec{\alpha}\}\end{aligned}$$

여기서 z을 최대화 시킬 수 있는 방법은 다양하고 알파의 크기가
클수록 z의 분산도 커진다.
무작정 키우면 안되기 때문에 alpha 값을 제한된다.

$$\|\alpha\| = \vec{\alpha}^{T}\vec{\alpha} = 1$$

여기서 최댓값을 구하기 위해 미지수 alpha로 미분한 식을 0으로 두면
고유벡터의 정의에 의해 alpha는 고유벡터 λ는 고유값이 된다.
Σ의 고유벡터를 주성분이라고 한다.

2.PCA

공분산 행렬의 서로 다른 고유벡터끼리는 서로 직교한다.
why?Σ(시그마)는 대칭행렬이기 때문이다.

$$\Sigma = cov(X) = \frac{1}{n-1} XX^T \propto XX^T$$

why? 공분산행렬은 비특이행렬로 고유값과 고유벡터의
개수가 차원수만큼 존재하기 때문에 식에 의해
정방행렬이자 대칭행렬이 된다

대칭행렬은 말 그대로 대칭이라 전치행렬의 값 과 행렬의 값이 같다.
A=A^T 이며 이는 내적이 0이 된다는 점에서 직교한다고 본다.

$$\Sigma^T = (A^{-1})^T \Lambda A^T$$
$$= A\Lambda A^{-1} = \Sigma$$

$$\therefore A^{-1} = A^T$$
$$A^T A = I$$

(직교 조건:내적 =0)

따라서 서로 다른 고유벡터끼리 **서로 직교하는 특징**을 갖는
이유가 바로 이 것 때문이며, 변수 간 연관성이 있더라도
PCA변환에 의해 바뀐 데이터들은 서로 **무상관**해진다.

2.PCA

4.PCA의 수행 절차

-1. 기존 데이터 x의 공분산 행렬을 계산한다.
(PCA의 목적은 '분산'을 최대화하여 최대한 잘 보존하는 것임)

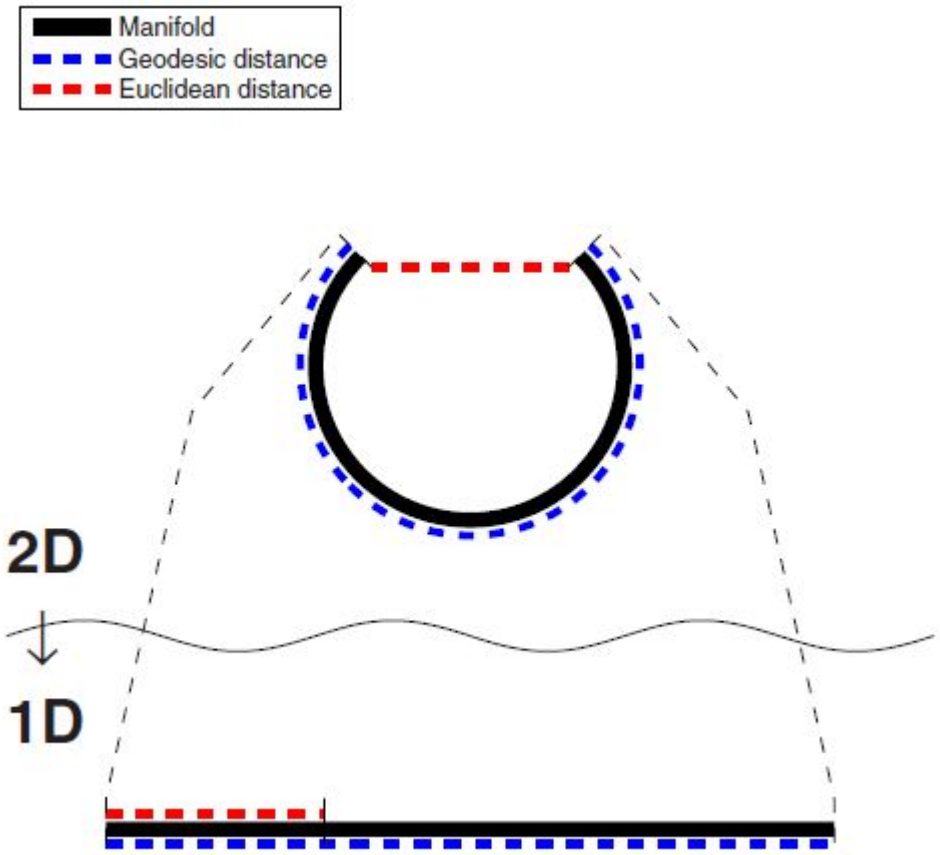
-2.공분산행렬의 고유값과 고유벡터를 계산한다.
(고유벡터가 주성분이 된다.)

-3.고유값 크기 순서대로 고유벡터를 나열한다.

-4.정렬된 고유벡터 가운데 일부를 선택한다.

-5.해당 고유벡터와 x의 내적을 구한다.

3.ISOMAP(Isomatric feature mapping)



ISOMAP은 비선형 구조를 가진 데이터의 차원 축소를 위해서 사용되며, 차원을 한단계 줄여주는 방법입니다.

ISOMAP의 경우 MDS와 매우 유사하지만 인스턴스 사이의 거리를 비선형 데이터 구조에 맞도록 Euclidean distance가 아닌 Geodesic distance를 사용합니다.

4.MDS (Multi Dimensional Scaling)

MDS는 PCA처럼 데이터 행렬을 사용하지 않고, 관찰된 데이터들 사이의 거리를 이용합니다. 따라서 Euclidean distance를 계산한 거리 행렬이 필요합니다.

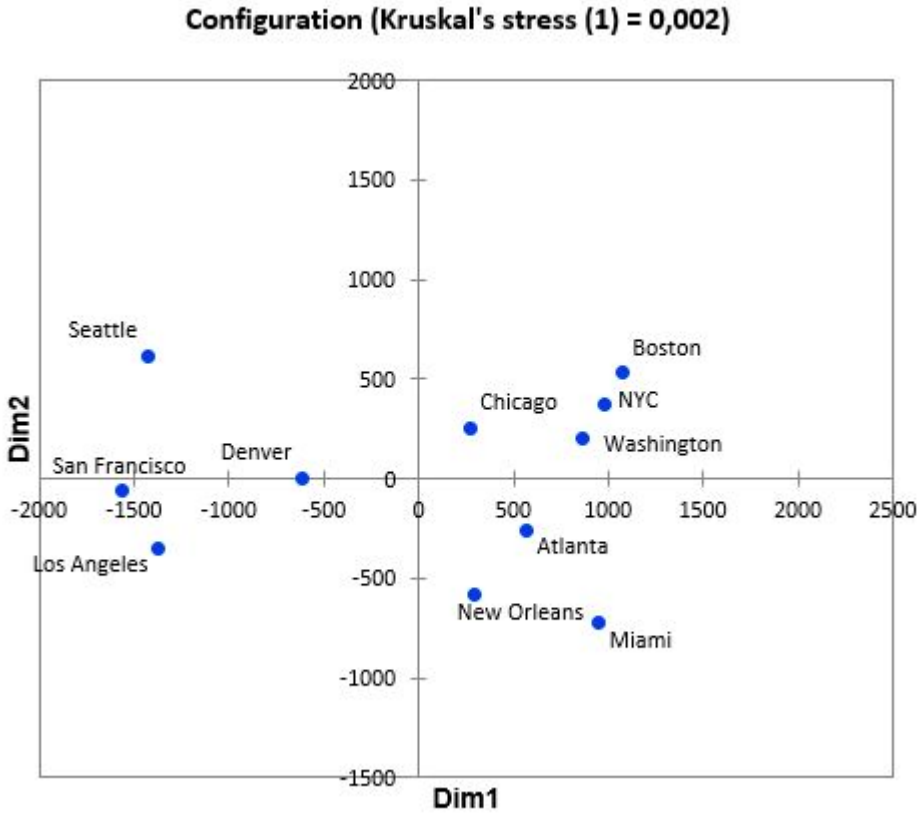
MDS또한 비선형 구조를 가진 데이터에 사용하며, 거리 행렬을 이용하여 자료들의 유사성 학습 결과를 사용해 다중의 변수들의 차원을 낮추어 나타내는 기법입니다.

MDS 사용시에는 자료가 가까이 위치할수록 유사성이 높고, 멀리 위치할수록 유사성이 낮습니다.

MDS에는 계량적 MDS(Metric MDS)와 비계량적 MDS(Nonmetric MDS) 두가지 방법이 있습니다.

4.MDS (Multi Dimensional Scaling)

아래는 미국의 도시들간 거리 행렬을 이용하여 MDS로 나타낸 결과입니다.



참조

https://towardsdatascience.com/preserving-geodesic-distance-for-non-linear-datasets-isomap-d24a1a1908b2

https://towardsdatascience.com/isomap-embedding-an-awesome-approach-to-non-linear-dimensionality-reduction-fc7efbca47a0

https://velog.io/@swan9405/MDS-Multidimensional-Scaling

https://yngie-c.github.io/machine%20learning/2020/10/06/isomap_and_lle/

https://blog.naver.com/PostView.naver?blogId=sw4r&logNo=221034788697&parentCategoryNo=&categoryNo=157&viewDate=&isShowPopularPosts=false&from=postView

https://ratsgo.github.io/machine%20learning/2017/04/24/PCA/

https://docs.sangyunlee.com/ml/analysis/undefined-1