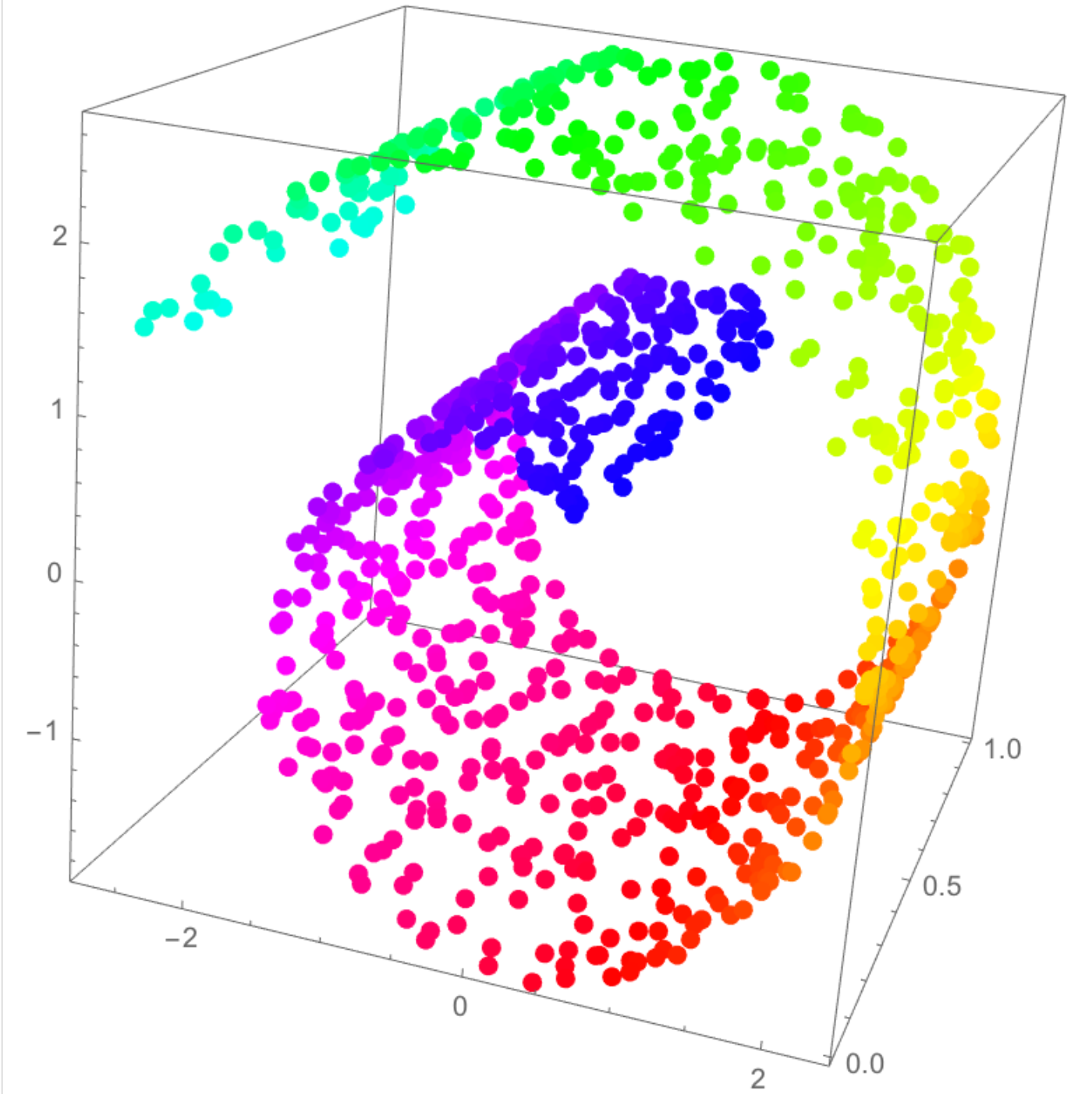
[인공지능 이론] Manifold Learning

2021. 9. 8. 23:23

#머신러닝 #딥러닝 #매니폴드 #Manifold #ManifoldLearning #PCA #LDA #t-sne #Isomap #MDS #AutoEncoder



1. 개요

Manifold Learning이란 고차원데이터가 있을 때 고차원 데이터를 데이터 공간에 뿌리면 샘플들을 잘 아우르는 sub psace가 있을 것이라 가정에서 학습을 진행하는 방법. 이렇게 찾은 manifold는 데이터의 차원을 축소시킬 수 있다.

1.1 개념

- Manifold는 고차원의 데이터를 저차원으로 옮길 때 데이터를 잘 설명하는 집합의 모형
- Manifold Learning이란 비선형 차원 축소에 관한 접근
- 높은 차원에서 낮은 차원으로 변환하는 것을 임베딩이라 하며 그것에 대한 학습 과정을 Manifold Learning 이라 함
- 고차원 공간 중에 존재하는 실질적으로 보다 저차원으로 표시 가능한 도형

1.2 배경

- ML 알고리즘은 정보를 고차원 공간의 벡터 형태로 저장하고, 이 고차원 공간은 그대로 이해하기 어렵다. 그렇기에 이런 문제를 해결하기 위해 고차원 공간을 2차원으로 압축하여 시각화하는 임베딩 방법들이 사용될 수 있음

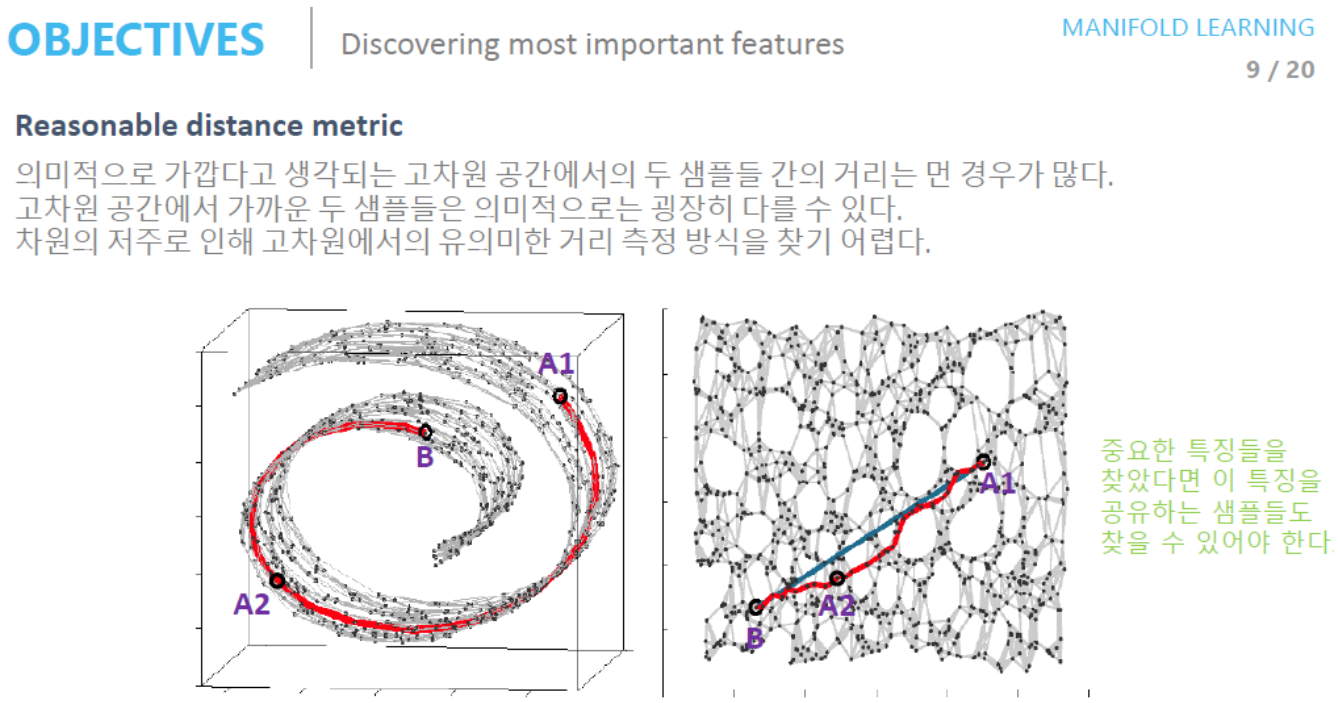
1.3 전제

- 고차원의 데이터의 밀도는 낮지만, 이들의 집합을 포함하는 저차원의 매니폴드가 있다.
- 이 저차원의 매니폴드를 벗어나는 순간 급격히 밀도는 낮아진다.

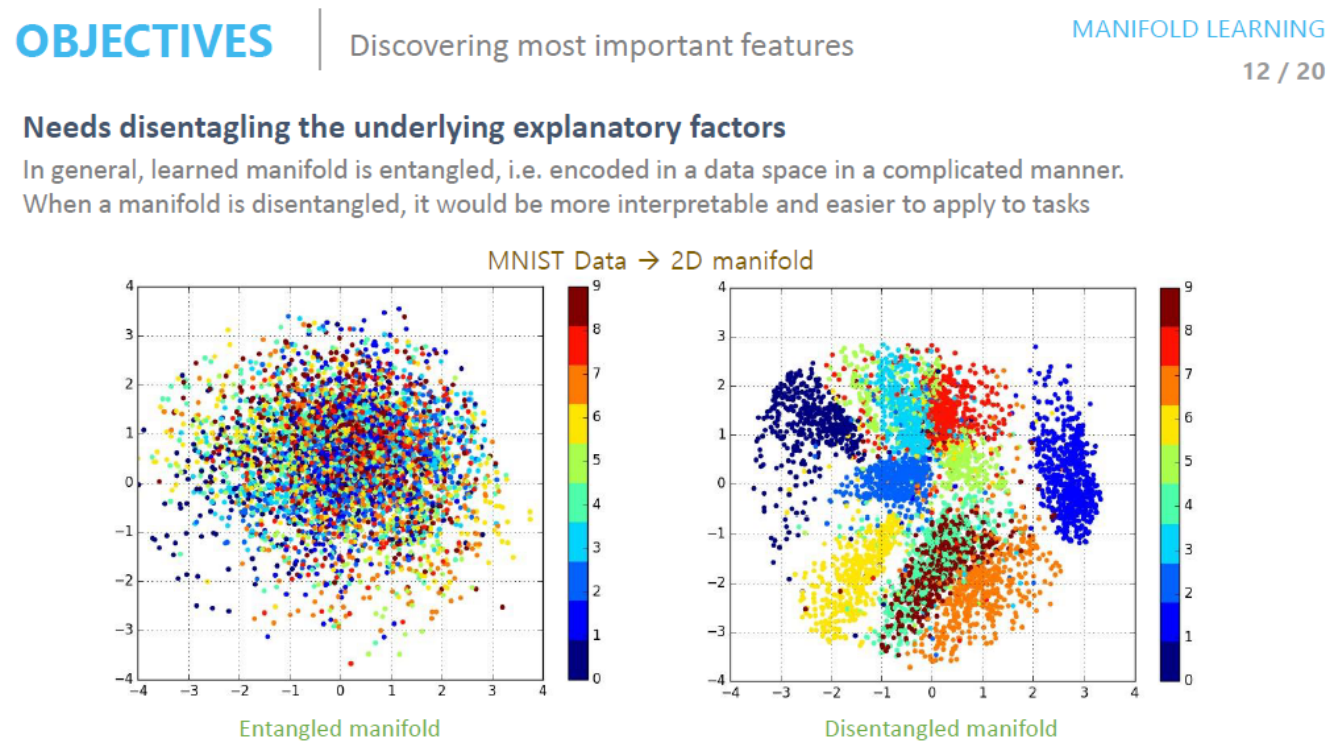
→ 고차원의 데이터를 잘 표현하는 manifold를 통해 샘플 데이터의 특징을 파악할 수 있는 것

1.4 특징

- manifold는 고차원 데이터를 잘 표현하고 이는 데이터의 중요한 특징을 발견하는 것
- 고차원 데이터의 manifold 좌표들을 조정해보면 manifold의 변화에 따라 학습 데이터도 유의미하게 조금씩 변하는 것을 확인가능
- 데이터(샘플)을 잘 아우르는 manifold를 찾게 되면 feature를 잘 찾았기 때문에 manifold의 좌표르 조금씩 변경해가면서 데이터를 유의미하게 조금씩 변화시킬 수 있다. → 데이터의 dominant한 feature를 잘 찾았기 때문임
- 역으로 manifold를 잘 찾았다면 dominant feature가 유사한 sample들을 찾을 수 있음



일반적으로 학습된 manifold는 얽혀 있고, manifold가 풀리면 해석이 쉽고 작업에 쉽게 적용가능

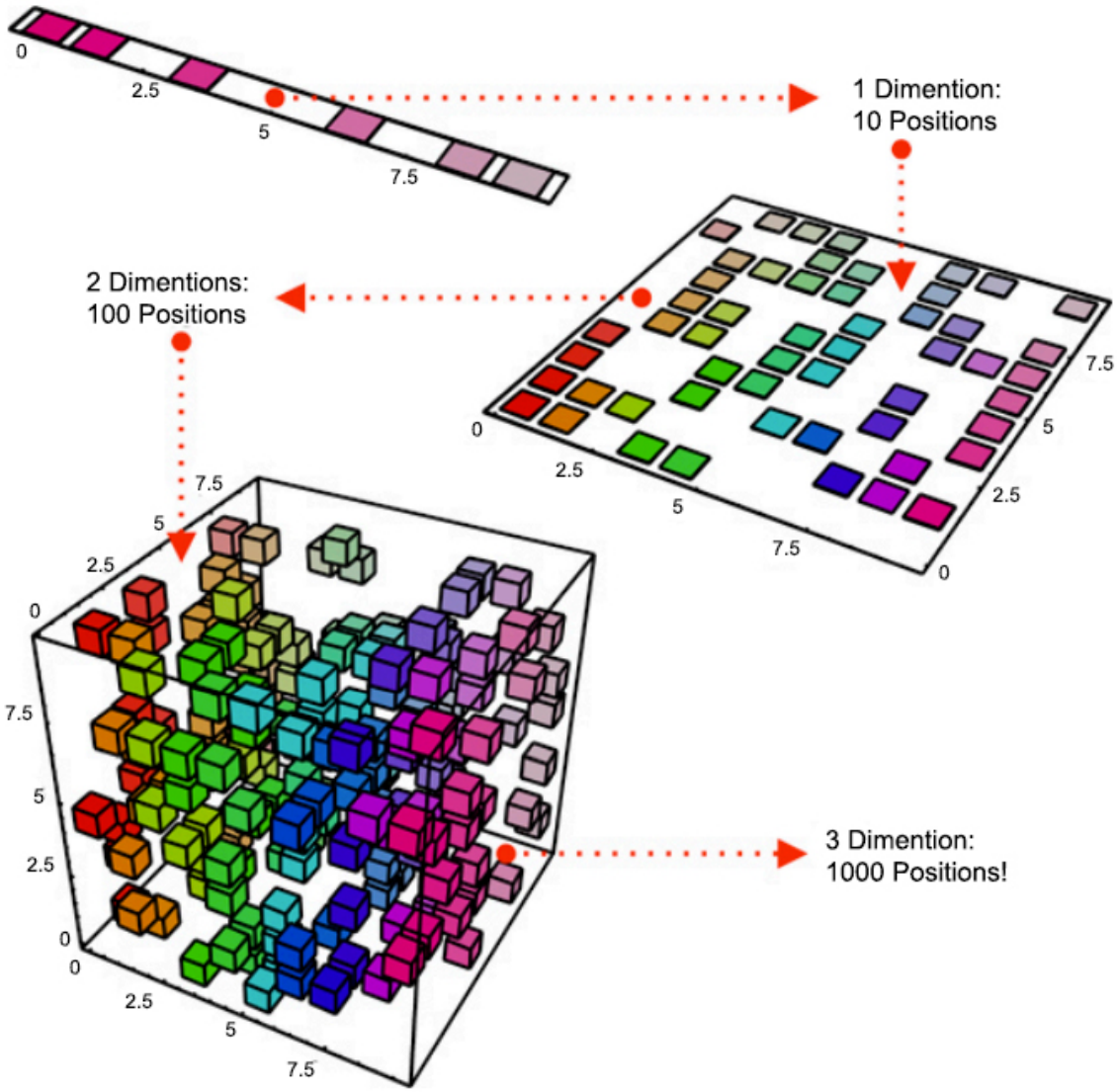


1.5 활용

manifold learning은 크게 4가지에 사용될 수 있다.

- Data Compression

- Data Visualization
- Curse of dimensionality(차원의 저주)
    - 데이터의 차원이 증가할수록 해당 공간의 부피는 기하급수적으로 증가
    - 동일한 개수의 데이터 밀도는 차원이 증가할수록 급속도로 희박



    - 따라서 차원이 증가할수록 데이터 분포 분석 또는 모델 추정에 필요한 샘플 데이터의 개수가 기하급수적으로 증가하게 됨
    - 차원이 늘어갈수록 사용하는 공간대비 아는 정보에 대한 밀도가 희박해짐. 따라서 차원이 증가할수록 데이터 분석에 필요한 데이터수가 기하급수적으로 늘어나게 됨

2. 종류

차원 축소 분류 체계

2.1 선형

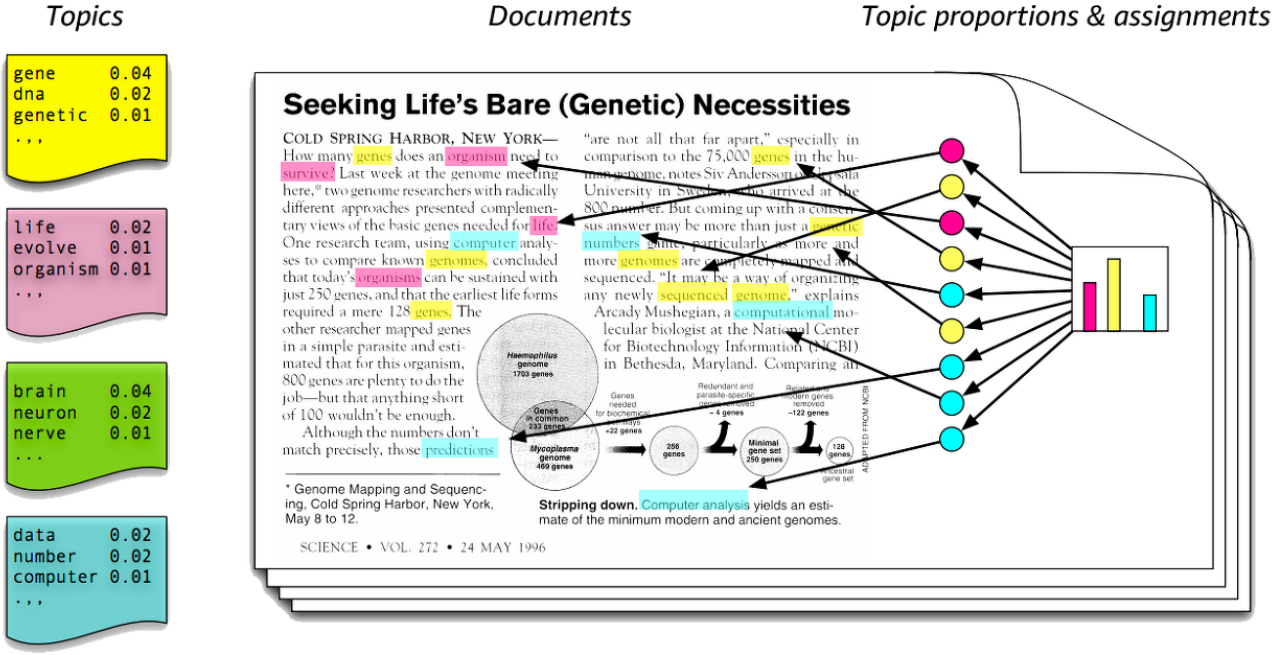
2.1.1 PCA (Principal Component Analysis)

- 간단히 말해 원 데이터를 공간에 뿌려 하이퍼플레인(hyperplane)을 찾는 방법이다. (1993)

2.1.2 LDA (Latent Dirichlet Allocation)

잠재디리클레할당이라 한다.

- (활용) 토픽 모델링에 사용하는 알고리즘
- (개념) 주어진 문서에 대해 각 문서에 어떤 주제들이 존재하는지에 대한 확률모형
- (특징) LDA는 토픽별 단어의 분포, 문서별 토픽의 분포를 모두 추정해냄



2.2 비선형

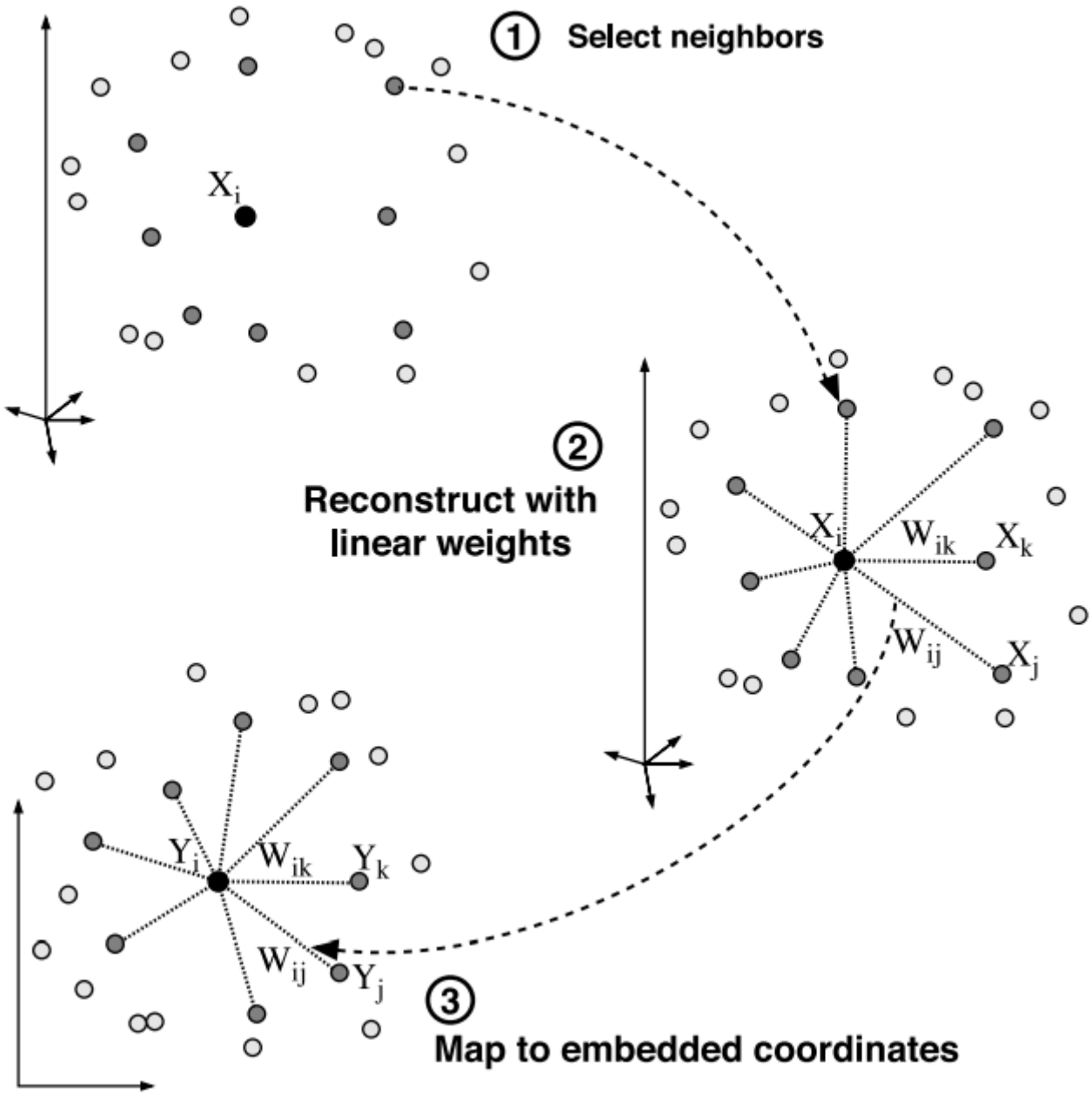
2.2.1 t-distributed Stochastic Neighbor Embedding(t-SNE)

- 가장 많이 사용되는 벡터 시각화 임베딩 방법 중 하나

2.2.2 kernel PCA

2.2.3 Locally Linear Embedding(LLE) - 로컬 선형 임베딩

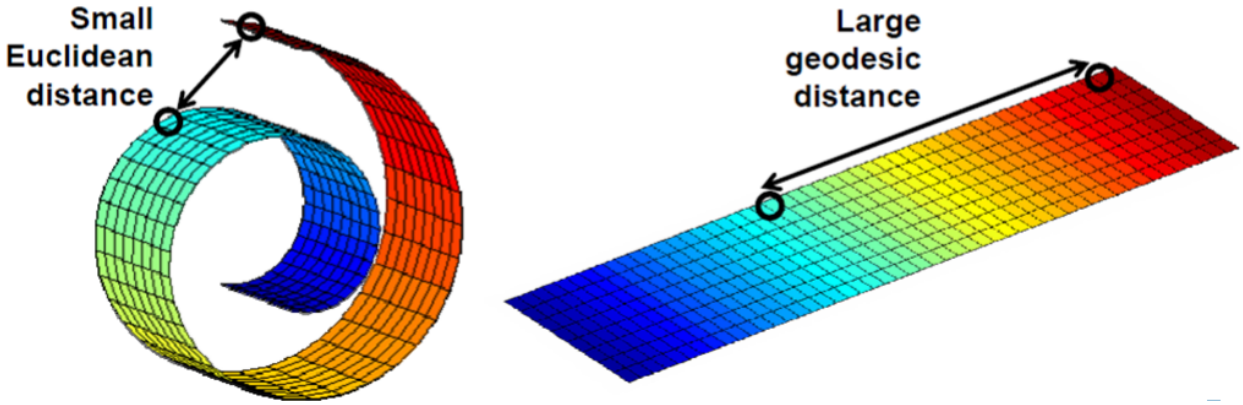
2000년 Science 저널에 nearest neighbors 정보를 이용하는 임베딩 방법 두 가지 중 한 가지이다.



- (개념) 고차원 공간에서 인접해 있는 데이터들 사이의 선형적 구조를 보존하면서 저차원으로 임베딩하는 방법론
- (단계) 3단계로 구성된다
  - 가장 가까운 이웃 검색
  - 가중치 메트릭스 구성
  - 부분 고유치 분해
- LLE는 ISOMAP과 목적은 동일하지만 지역성(Locality)을 어떻게 수학적으로 반영하느냐에 따라 지역적 선형 임베딩과 ISOMAP의 차이가 발생한다.
- 쉽게 말하자면 LLE는 각 훈련 샘플이 K-Neighbor들과 얼마나 선형적으로 연결되어있는가를 추론하는 과정이다.
- LLE는 Locally Linear Embedding)의 약자로 PCA로는 차원을 낮출 수 없는 비선형적 차원 축소를 가능하게 하는 방법이다.

2.2.4 ISOMAP

- 다차원 스케일링(MDS) 또는 주성분 분석(PCA)의 확장이자 두 방법론을 결합한 방법론이다. (1952)



- ISOMAP은 CPA와 MDS의 특징을 결합하여 모든 점 사이의 측지선 거리를 유지하는 더 낮은 차원의 임베딩을 추구한다. 여기서 측지거리는 두 측점 사이의 타원체면을 따라 이루어진 거리를 말한다.
- 위 그림에 따르면 두 점은 유클리디안 거리로는 가깝지만 실제 측지거리를 구할 경우 색깔이 나타내는 의미만큼 멀리 떨어져 위치함을 알 수 있다. 즉, Isomap 알고리즘은 두 데이터간의 실제 특징을 반영하는 거리 정보를 사용하는 효과적인 차원 축소를 추구한다.

2.2.5 Multi-Dimensional Scaling(MDS)

2.2.6 Auto-Encoder

- (목적) Encoder에서는 Manifold 가정을 통해 Input data의 고차원 데이터는 희박한 밀도를 가지고 있으므로 저차원의 데이터로 만들어서 원래의 데이터를 잘 설명하는 Manifold를 찾는 것
- Decoder는 왜 존재하는가? → Decoder로 latent variable을 원래 data로 만들어주게 되면 label로 input data를 사용할 수 있으므로 지도 학습이 가능해지기 때문이다.
- (활용) Auto Encoder가 생성 모델로도 쓰이고 Denoising으로 사용된다.

Reference

- https://deepinsight.tistory.com/124
- https://simpling.tistory.com/16
- https://science.sciencemag.org/content/sci/290/5500/2323.full.pdf
- https://lovit.github.io/nlp/representation/2018/09/28/mds_isomap_lle/

- https://woosikyang.github.io/first-post.html

공감 구독하기

'AI Reasearch > 머신러닝-딥러닝' 카테고리의 다른 글

[머신러닝/딥러닝] 우분투에 아나콘다 설치 (Anaconda Installation on Ubuntu) (0)	2021.11.04
[머신러닝/딥러닝] Ubuntu 20.04에 딥러닝 환경 설치 (0)	2021.11.02
[인공지능 이론] Manifold Learning (1)	2021.09.08
합성곱 신경망 (Convolutional Neural Network, CNN) (0)	2021.09.08
[인공지능 이론] 인공지능의 역사 (0)	2021.09.08
[Pandas] 데이터 프레임에 컬럼 이름 추가하기 (0)	2020.05.12

NAME

PASSWORD

HOMEPAGE

http://

SECRET WRITE

2021.12.01 14:55

다른 LDA를 넣어놨네요.

본문의 LDA (Latent Dirichlet Allocation)는 토픽 모델링으로 차원축소에도 사용가능하지만,

원래 의도는 Linear Discriminant Analysis (LDA)를 설명하려고 했는 것 같네요.

Delete Reply

PREV 1 2 3 4 5 6 7 ··· 25 NEXT

+ Recent posts

[독서노트#17] 타이탄의 도구…

[독서노트#16] 울트라 러닝 (…

[독서노트#15] 7막 7장 (홍정욱)

[독서노트#14] 라플라스의 악…