이상 탐지 A to Z

4편.

비지도 학습 기반의 머신러닝 기법(1부)

데이크루 2기 Team Zoo

DACON
DATA TO VALUE

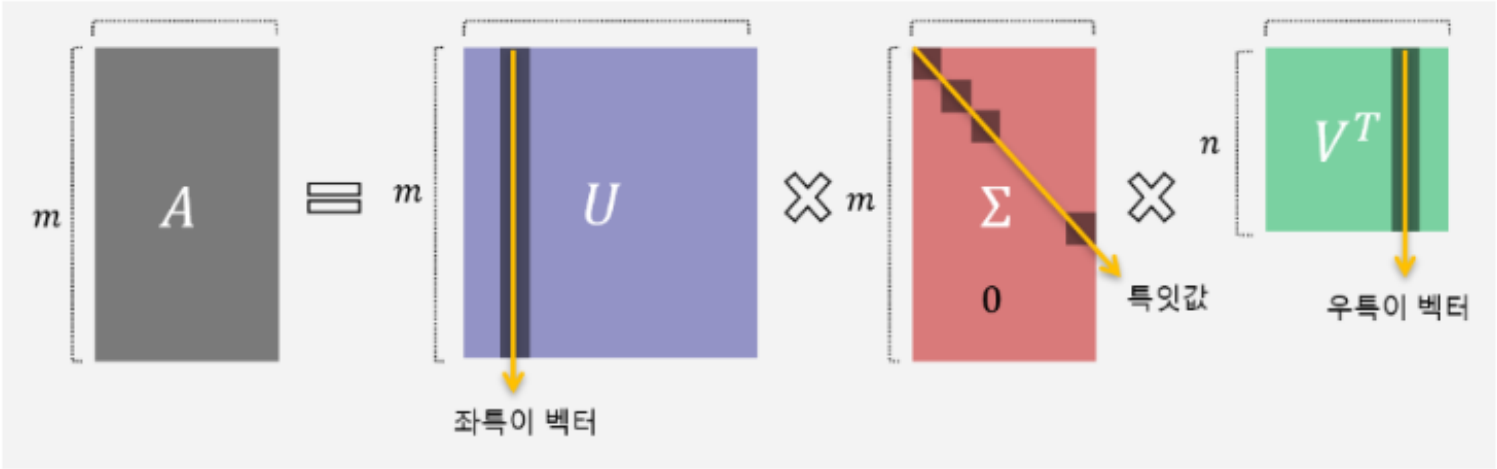
목차

1. SVD
2. Random Projection
3. LLE
4. t – SNE
5. 사전학습
6. ICA

1. SVD

- SVD란, 임의의 m x n 직사각 행렬을 다음의 세 가지 행렬로 대각화하여 분해하는 방법을 말한다.
- A = U $\sum VT$
- U = m x m 직교행렬
- $\sum = m\ x\ n$ 대각행렬
- V = n x n 직교행렬

1. SVD



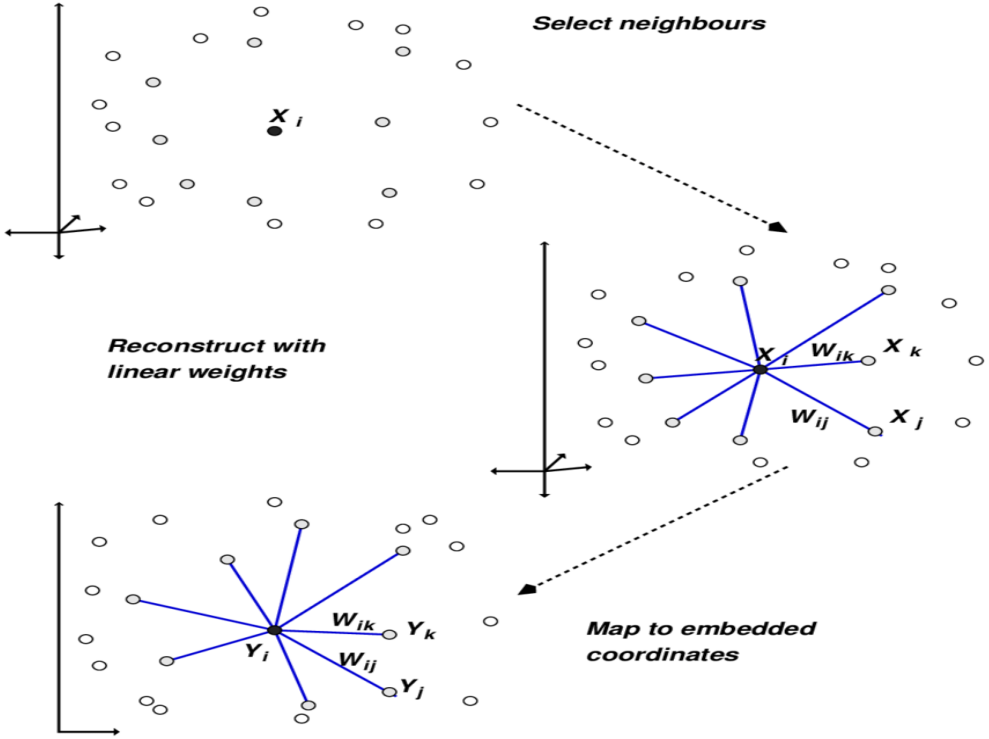
- 특이값 분해란, [m x m] 행렬을 [m x r], [r x r], [r x n] 행렬들의 곱으로 근사하게 표현하는 것. 이 때, [m xr], [r x n]은 정규직교 (Orthornomal)이며, [r x r]은 대각 (Diagomal) 행렬.
- 특이값 분해는 정방 행렬뿐만 아니라 행과 열의 크기가 다른 행렬에 대해서도 적용할 수 있습니다. 즉, 특이값 분해는 모든 직각 행렬에 대해 가능

2. Random Projection

- Random Projection (꼐)을 이용하면 벡터 간의 거리를 보존하며 차원을 저차원으로 바꿀 수 있음.
- Random Projection은 Johnson-Linderstrauss Lemma를 이용

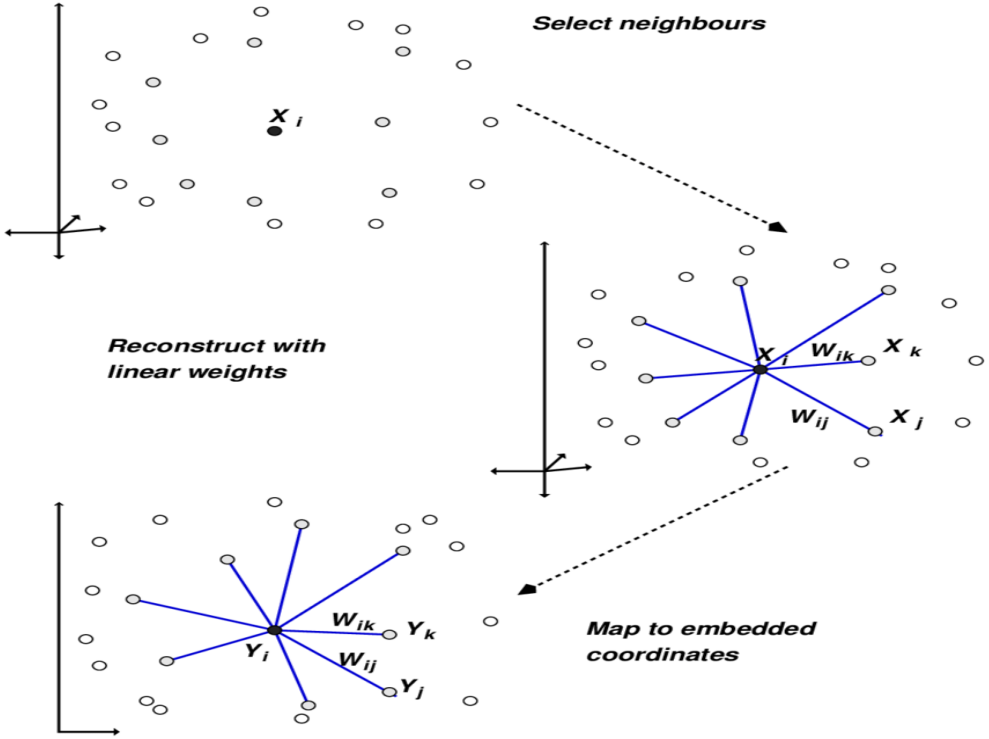
3. LLE

- LLE (Locally Linear Embedding)
- 고차원에서 최인접 이웃의 정보에 집중하는 방법



3. LLE

- LLE (Locally Linear Embedding)
- 고차원에서 최인접 이웃의 정보에 집중하는 방법



3. LLE

- Step 1: 고차원에서 x_i 와 가장 가까운 k개의 x_j들을 선택
- =〉 k는 하이퍼 파라미터로 사람이 직접 적절한 개수를 정함.

- Step 2: x_i 와 가장 가까운 x_j를 가장 잘 재구성하는 가중치 w_{ij}를 구하고, $\sum_{j=1}^{k} w_{ij} x_j \approx x_i$ 를 만족하는 w_{ij}를 학습

- Step 3: 저차원 y_i에서 y_j에 대해 재구성하고, Step 2에서 구한 w_{ij}를 이용하여 Y를 찾는 최소화 문제

=〉 $\min Y = \sum_{i=1}^{m} ||y_i - \sum_{j=1, j\neq i}^{k} w_{ij} y_j||^2$

4. t – SNE

- **t – SNE (t - Distributed Stochastic Neighbor Embedding)**
- 고차원에서 가까운 점들은 저차원에서도 가깝게 이동하고, 고차원에서 먼 점들은 저차원에서도 먼 곳으로 이동시킴.
- LLE 방법과는 다르게 좀 더 멀리있는 점들의 위치로 고려하는 방법
- 고차원의 유사도: $p_{ij=}\ \frac{p_{j|i+}\, p_{i|j}}{2n}$ (정규분포)
- 저차원의 유사도: q_{ij} (t 분포)

$$p_{j|i} = \frac{\exp(\frac{-|x_i - x_j|^2}{2{\sigma_i}^2})}{\sum_{k \neq i} \exp(\frac{-|x_i - x_k|^2}{2{\sigma_i}^2})}$$

$$q_{ij} = \frac{(1 + |y_i - y_j|^2)^{-1}}{\sum_{k \neq l}(1 + |y_k - y_j|^2)^{-1}}$$

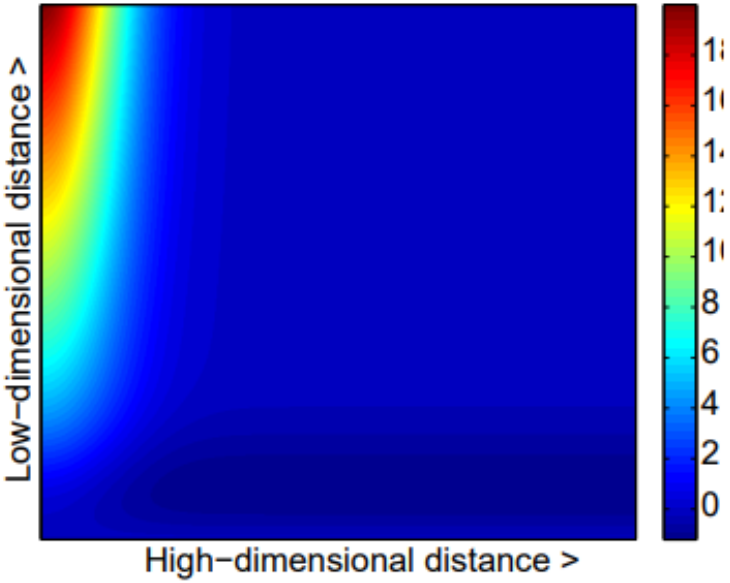
4. t – SNE

- **t 분포를 사용하는 이유?**
- t- SNE 전에 SNE라는 방법이 있었는데, 저차원의 공간에서도 정규분포를 가정하여 사용
- 하지만, 거리가 가까운 점들은 무한대 값에 가까워지므로 t – 분포를 사용
- 또한, 정규분포에서는 거리가 먼 데이터 포인트들과 약간 먼 데이터 포인트들을 잘 잡지 못하는 일이 발생 (crowding Problem) =〉 확률이 비슷하게 정의되기 때문
- t 분포의 특성 중 끝단의 꼬리가 두꺼운 부분을 이용해서 먼 데이터 포인트들과 약간 먼 데이터들의 포인트를 거리적으로 배치 가능

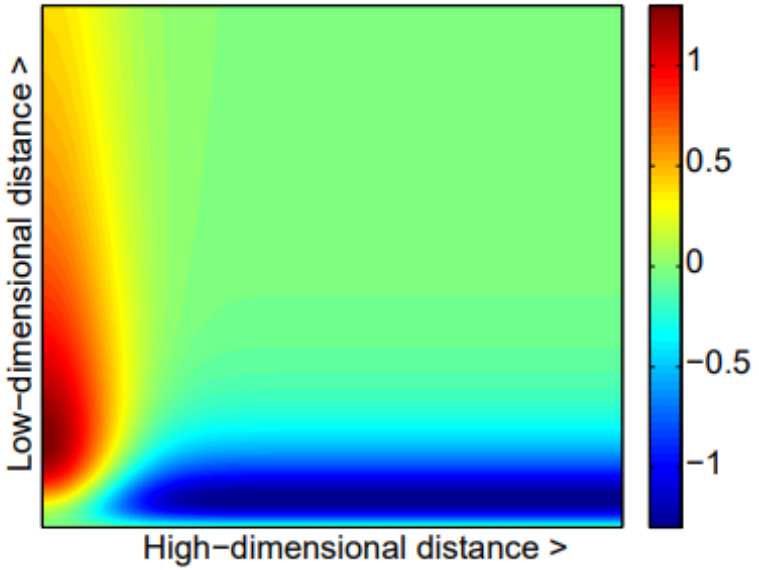
- 임베딩 공간의 점들 간 유사도 분포를 t 분포로 더 안정적인 학습 결과를 얻어냄.

4. t – SNE

- t 분포를 사용하는 이유?



(a) Gradient of SNE.



(c) Gradient of t-SNE.

4. t – SNE

- **t –SNE가 학습되는 과정**
- Gradient Descent를 이용
- $\frac{\partial C}{\partial y_i} = \sum_j (p_{ij} - q_{ij})(y_i - y_j) \frac{1}{1+|y_i - y_j|^2}$ (Kullback–Leibler divergence 미분)
- Kullback–Leibler divergence는 두 확률분포의 차이를 계산하기 위해 사용되는 함수 => p_{ij} 와 q_{ij}의 분포가 같아지도록 업데이트
- t 분포의 특성을 이용해서 y_i 가 고차원에서 멀리있던 점들이 저차원에서 가까웠던 점들은 멀리 배치시킬 수 있도록 하고, 고차원에서 가까이 있던 점이 저차원에서 멀리 떨어져 있으면 가까이 배치 시킬 수 있도록 이동

4. t – SNE

- **t –SNE가 학습되는 과정**
- Gradient Descent를 이용
- $\frac{\partial C}{\partial y_i} = \sum_j (p_{ij} - q_{ij})(y_i - y_j) \frac{1}{1+|y_i - y_j|^2}$ (Kullback–Leibler divergence 미분)
- Kullback–Leibler divergence는 두 확률분포의 차이를 계산하기 위해 사용되는 함수 => p_{ij} 와 q_{ij}의 분포가 같아지도록 업데이트
- t 분포의 특성을 이용해서 y_i 가 고차원에서 멀리있던 점들이 저차원에서 가까웠던 점들은 멀리 배치시킬 수 있도록 하고, 고차원에서 가까이 있던 점이 저차원에서 멀리 떨어져 있으면 가까이 배치 시킬 수 있도록 이동

5. 사전 학습

- 기하학적 구조나 거리 척도에 의존하지 않는 방법 중 하나
- 주어진 간단한 단서를 이용하여 필요한 정보를 찾아내는 방법

- **스파스 사전 학습(sparse dictionary learning): sparse coding이라는 기법을 활용하는 사전 학습, 매우 적은 단서를 이용하여 필요한 정보를 찾아내는 방법**

- sparse: 드문, 희박한. 벡터나 행렬의 많은 원소가 0인 경우
- dictionary (사전): 결과 행렬
- learning(학습)
- atom(원자): 사전 안에 있는 하나의 열 벡터 (0과 1로 구성)

- 즉, label 값을 모르는 어떤 데이터를 사전 내에 있는 atom들의 선형 조합으로 나타낼 때, 선형 계수들이 최대한 0이 되도록 하는 알고리즘

5. 미니 배치 버전의 사전학습

- 성분의 수 설정
- 배치 크기, 반복 횟수 설정

6. ICA

- ICA (Independent Component Analysis): 독립 성분 분석으로 다변량의 신호를 통계적으로 독립적인 성분으로 분리하는 것
- Blind Signal Seperation가 하는 일: 섞인 두 소리를 분리해 내는 일
- **GOAL: 랜덤 변수인 s 들이 서로 독립적이라는 가정을 최대한 만족하는 W를 찾는 것**

- x : 녹음된 신호
- s : 음성 신호
- A: mixing matrix
- W: unmixing matrix
- CLT: 서로 독립적인 랜덤 변수(s)들의 분포의 선형 조합 (x) 은 가우스 분포를 따름.
- ICA: x들을 어떻게 조합하면 s를 얻을까?

6. ICA

- Bell – Sejnowski algorithm

– W를 업데이트 할 때마다 계산해줘야하는 역행렬 term이 있어 계산 속도가 느리다는 단점이 있음

- natural gradient algorithm

– Bell – Sejnowski algorithm의 단점을 보완하기 위해 만들어진 알고리즘

– 신호 처리 작업에 사용

– PCA와 달리, 가장 독립적인 축을 찾는 방법

– 독립성이 최대가 되는 벡터를 찾읍ㅁ

– 독립성은 ICA 알고리즘에 의해 계산