Chapter 8 차원축소

차원의 저주

✓ **차원의 저주**: 무작위로 선택한 고차원 벡터는 매우 희소해서 과대적합의 위험이 크고 많은 양의 데이터가 있지 않으면 데이터의 패턴을 잡아내기 어려워지는 문제

- **차원 축소의 주요 목적**

> - 훈련 알고리즘의 속도를 높이기 위해
> - 데이터를 시각화하기 용이하도록
> - 메모리 공간 절약

- **주요 단점**

> - 일부 정보가 유실되어 훈련 알고리즘 성능을 낮출 수 있다.
> - 변환된 데이터를 이해하기 어려운 경우가 많다

차원 축소 접근 방법

• 투영

대부분의 훈련샘플들은 몇몇의 특성에 대해서만 강한 연관을 띄고 있어서 고차원 공간 안의 **저차원 부분공간**에 놓여있다.

✓**주의**: 스위스 롤과 같이 부분공간이 뒤틀려있거나 휘어있을 때 투영은 좋지 않다.

• 매니폴드 학습

✓ **d차원 매니폴드(manifold, 다양체)**: d < n일 때 d차원 매니폴드는 국부적으로 d차원 초평면으로 보일 수 있는 n차원 공간의 일부

✓ **매니폴드 가정**: 대부분의 실제 고차원 데이터셋은 더 낮은 저차원 매니폴드에 가깝게 놓여있다는 가정

암묵적으로 저차원의 매니폴드 공간에 표현되면 더 간단해질 것이라는 가정이 병행되지만 무조건적인 것은 아니다.

매니폴드 설명 참조

PCA(주성분 분석, Principal Component Analysis)

데이터에 가장 가까운 초평면을 정의한 다음 데이터를 이 평면에 투영하는 차원 축소기법

• 분산보존

분산이 최대한 보존되는 축을 선택하는 것이 정보가 가장 적게 손샐되므로 합리적

즉 분산이 최대한 보존되려면 **원본 데이터셋과 투영된 데이터셋 사이의 평균 제곱거리가 최소화되도록 하는 축을 선택한다.**

- ## 주성분

데이터셋의 분산이 최대한 보존되는 축을 찾는다. 이 후 이전의 모든 축들과 직교하고 분산을 최대한 보존하는 새로운 축을 찾는다. 데이터셋의 차원의 수 만큼 찾는다.

이때 i번째 축을 **데이터의 i번째 주성분**이라한다.

✓ 주성분 찾는법

: 훈련 세트 행렬이 X라 할때 **특이값 분해(SVD)**를 통해 $X = U\Sigma V^T$ 로 분해하면 모든 주성분의 단위벡터(C_1, C_2, C_3)가 V에 있다.

$$V = \begin{pmatrix} | & | & | \\ c_1 & c_2 & c_3 \\ | & | & | \end{pmatrix}$$

특이값 분해 참조

- ## d차원으로 투영하기

$$X_{d-proj} = XW_d$$

> X_{d-proj}: d차원에 투영된 훈련세트
>
> X: 훈련세트
>
> W_d: V의 첫 d개의 열로 구성된 행렬

- ## 적절한 차원의 수 선택

분산을 차원 수에 대한 함수로 그리고 분산의 빠른 성장이 멈추는 변곡점에 해당하는 차원을 택한다.

- ## 랜덤 PCA

확률적 알고리즘을 사용해 처음 d개의 주성분에 대한 근삿값을 빠르게 찾는다.

d값이 n값보다 작을 때 훨씬 빠르게 동작

- ## 점진적 PCA(IPCA, Incremental PCA)

PCA 구현의 문제: SVD 알고리즘을 실행하기 위해 전체 훈련 세트를 메모리에 올려야 함

점진적 PCA: 훈련 세트를 미니배치로 나눈 뒤 IPCA 알고리즘에 한 번에 하나씩 주입. 이런 훈련 방식은 훈련 세트가 클 때 유용하고 온라인으로 (즉, 새로운 데이터가 준비되는 대로 실시간으로) PCA를 적용할 수도 있다.

kPCA

고차원 특성 공간에서의 선형 결정 경계 = 원본공간(비교적 저차원)에서의복잡한 비선형 결정 경계

이를 pca에 적용하여 **차원축소를 위한 복잡한 비선형 투영**을 수행할 수 있다.

- ## 커널 선택과 하이퍼파라미터 튜닝
    - 그리드 탐색을 사용하여 성능이 가장 좋은 커널과 하이퍼파라미터를 선택한다.
    - 가장 낮은 재구성 오차를 만드는 커널과 하이퍼파라미터를 선택한다(PCA의 원래 목표는 원래 데이터를 가장 잘 설명해주는 주성분을 찾는 것, **복원될 데이터의 재구성 오차가 가장 낮도록 하는 기처를 찾는 것**)

✓ kPCA는 선형 PCA 만큼 재구성이 쉽지 않다

✓ **재구성 원상**: 축소된 공간의 샘플에 대해 선형 PCA를 역전시기면 재구성된 데이터 포인트는 원본공간이 아니라 다차원의 특성공간에 놓이게 되므로 오차를 계산하기 어렵다. 따라서 투영된 샘플을 훈련세트로, 원본세트를 타깃(레이블)으로 하는 지도 학습 회귀 모델을 훈련시켜 재구성된 포인트에 가깝게 매핑된 원본공간의 포인트를 찾을 수 있다.

즉 재구성 원상과 원본 샘플과의 제곱거리를 측정하여 오차를 최소화하는 커널과 하이퍼파라미터를 선택하게 한다.

LLE(지역 선형 임베딩)

- **강력한 비선형 차원 축소기술**
- **매니폴드 학습의 일종**
- 대량의 데이터셋에 적용하기 어렵다.

각 훈련 샘플마다 **가장 가까운 이웃에 얼마나 선형적으로 연관되어 있는지 측정**하여 이 국부적 선형 관계가 가장 잘 보존되는 훈련세트의 저차원 표현을 찾는 알고리즘

다른 차원 축소 기법

• 랜덤 투영

랜덤한 선형 투영을 사용하여 저차원 공간으로 투영하는 기법

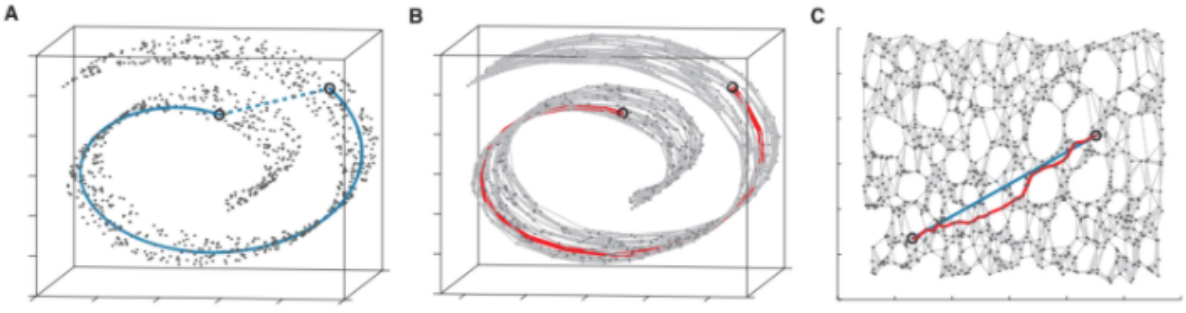
• 다차원 스케일링

샘플 간의 거리를 보존하며 차원을 축소하는 기법.

주로 거리행렬으로 표현된 데이터셋을 시각화하는 데 이용

• Isomap

훈련 샘플 간의 지오데식 거리(geodesic distance,그래프에서 두 노드사이의 최단경로를 이루는 노드의 수)를 유지하도록 차원을 축소하는 기법



• t - SNE

t분포를 사용한 SNE.

✓**SNE**: 고차원 공간의 훈련 샘플들 간의 거리를 임의의 훈련 샘플의 이웃에 대한 조건부 확률로 나타내 이 확률 정보를 저차원으로 차원축소를 시켰을 때도 보존하도록 하는 매핑기법

주로 시각화에 많이 사용되며 특히 고차원 공간에 있는 샘플의 군집을 시각화할 때 사용됨

- **LDA(선형판별분석)**

  데이터 분포를 학습해 클래스 사이를 가장 잘 구분하는 축을 만들어 데이터를 분류하는 알고리즘

  투영을 통해 가능한 한 클래스를 서로 멀리 떨어지게 만들기 때문에 분류를 위한 차원 축소에 유용하다.