CHAPTER8. 차원 축소

- 차원의 저주
- 차원 축소를 위한 접근 방법
- PCA
- 커널 PCA
- LLE
- 다른 차원 축소 기법

8.1 차원의 저주

- 고차원 초입방체에 있는 대다수이 점은 경계와 매우 가까이 있습니다.
- 저차원일때보다 고차원에서 두 점을 무작위로 선택하면 거리가 더 멉니다.
- 이는 고차원의 데이터셋이 매우 희박한 상태일 수 있음을 의미합니다.
- 대부분 Train 데이터가 서로 멀리 떨어져 있다는 것을 의미합니다.
- 이럴 때 예측을 위해선 훨씬 더 많은 외삽을 해야하기 때문에 저차원보다 불안정합니다.
- 간단히 말해 훈련 세트의 차원이 클수록 과대적합 위험이 커집니다.
- 이론적으로 차원의 저주를 해결하는 방법은 Train 데이터를 키우는 것입니다.
- 하지만 데이터를 키우면 차원이 기하급수적으로 늘어납니다.

8.2 차원 축소를 위한 접근 방법

8.2.1 투영 (Projection)

- 데이터의 많은 특성들은 변화가 없는 반면 다른 특성들은 서로 강하게 연관되어 있습니다.
- 즉, 고차원 공간 안의 저차원 부분 공간에 있는 데이터를 저차원으로 투영하면 차원을 줄일 수 있습니다.

8.2.2 매니폴드 학습 (Manifold Learning)

- d차원 매니폴드는 국부적으로 d차원 초평면으로 보일 수 있는 n차원 공간의 일부입니다. (d < n)
- 스위스롤의 경우 국부적으로는 2D 평면으로 보이지만 3차원으로 말려 있습니다.
- 실제 고차원 데이터셋이 더 낮은 저차원 매니폴드에 가깝게 놓여있다는 매니폴드 가설에 근거합니다.
- 이에 따라 훈련 샘플이 놓여 있는 매니폴드를 모델링하는 알고리즘을 의미합니다.

8.3 PCA (Principal Component Analysis)

- 가장 인기있는 차원 축소 알고리즘입니다.
- 먼저 데이터에 가장 가까운 초평면을 정의한 다음, 데이터를 이 평면에 투영시킵니다.

8.3.1 분산 보존

- 분산이 최대로 보존되는 축을 선택하는 것이 정보가 가장 적게 손실되므로 좋습니다.
- 즉 원본 데이터셋과 투영된 것 사이의 평균 제곱 거리를 최소화하는 축입니다.

8.3.2 주성분 (PC, Principal Componenet)

- PCA는 Train 데이터에서 분산이 최대인 축을 찾습니다.
- i번째 축을 정의하는 단위 벡터를 i번째 주성분이라고 부릅니다.
- Train 데이터의 주성분을 찾는 방법은 특잇값 분해 (SVD, Singular Value Decomposition) 라는 표준 행렬 분해 기술을 이용합니다.
- 훈련세트 행렬 X를 세 개 행렬의 점곱인 $U \cdot \sum \cdot V^T$ 로 분해합니다.

8.3.3 d차원으로 투영하기

- 주성분을 추출했으면 처음 d개의 주성분으로 정의한 초평면에 투영해 데이터셋의 차원을 d차원으로 축소시킬 수 있습니다.

$$X_{d-proj} = X \cdot W_d$$

8.3.4 적절한 차원 수 선택하기

- 데이터 시각화를 위해 차원을 축소하는 경우에는 차원을 2개나 3개로 줄이는 것이 일반적입니다.

8.4 커널 PCA

- 커널트릭을 PCA에 적용해 복잡한 비선형 투영으로 차원을 축소하는 방법입니다.

8.5 LLE (Locally Linear Embeding)

- 이전 알고리즘처럼 투영에 의존하지 않는 매니폴드 학습입니다.
- 먼저 각 Train 데이터가 가장 가까운 이웃에 얼마나 선형적으로 연관되어 있는지 측정합니다.

- 다음 국부적인 관계가 가장 잘 보존되는 Train 데이터의 저차원 표현을 찾습니다.
- 특히 잡음이 너무 많지 않은 경우 꼬인 매니폴드를 펼치는 데 잘 작동합니다.

8.6 다른 차원 축소 기법

8.6.1 다차원 스케일링 (MDS, Multidimensional Scaling)

- 샘플 간의 거리를 보존하면서 차원을 축소합니다.

8.6.2 Isomap

- 각 샘플을 가장 가까운 이웃과 연결하는 식으로 그래프를 만듭니다.
- 그런 다음 샘플 간의 지오데식 거리를 유지하면서 차원을 축소합니다.

8.6.3 t-SNE (t-Distributed Stochastic Neighbor Embedding)

- 비슷한 샘플은 가까이, 비슷하지 않은 샘플은 멀리 떨어지도록 하면서 차원을 축소합니다.
- 주로 시각화에 많이 사용되며 특히 고차원 공간에 있는 샘플의 군집을 시각화할 때 사용합니다.

8.6.4 선형 판별 분석 (LDA, Linear Discriminant Analysis)

- 데이터가 투영되는 초평면을 정의하는 데 사용할 수 있습니다.
- 투영을 통해 가능한 클래스를 멀리 떨어지게 유지시킵니다.
- SVM 분류기 같은 다른 분류 알고리즘을 적용하기 전에 차원을 축소시키는데 좋습니다.