

Euron 10주차 출석 과제 (ML 조한희)

차원 축소는 고차원 데이터의 복잡한 구조를 이해하기 쉽게 표현하기 위한 핵심적인 데이터 전처리 및 시각화 기법이다. 데이터의 특성이 수십 개 이상일 경우, 모든 차원을 그대로 분석하는 것은 비효율적이며, 시각화조차 불가능하다. 이때 차원 축소는 원본 데이터의 정보 손실을 최소화하면서 더 낮은 차원으로 표현하여, 데이터의 본질적인 구조나 패턴을 파악할 수 있게 해준다. 대표적인 방법으로는 PCA(주성분 분석), MDS(다차원 척도법), t-SNE가 있다.

1. PCA (Principal Component Analysis, 주성분 분석)

PCA는 데이터의 분산(variance)을 가장 잘 보존하는 새로운 축(주성분)을 찾아 데이터를 투영하는 방식이다. 즉, 여러 특성이 서로 상관되어 있을 때 이를 직교하는 새로운 축으로 변환하여, 정보의 손실을 최소화하면서 차원을 줄이는 통계적 방법이다. 각 주성분(Principal Component)은 데이터가 퍼져 있는 방향을 나타내며, 첫 번째 주성분은 가장 큰 분산을, 두 번째 주성분은 그 다음으로 큰 분산을 설명한다. Kaggle의 유방암 진단 데이터 분석 예제에서는 30개의 피처를 표준화한 후 PCA를 적용해 2개의 주성분으로 축소했다. 결과적으로 악성(M)과 양성(B) 두 그룹이 2차원 상에서 명확히 분리되어 시각적으로 구별 가능했다. 추가로 .components_ 속성을 통해 각 주성분이 원래의 30개 변수와 어떤 상관관계를 갖는지도 확인할 수 있었는데, 첫 번째 주성분은 대부분의 변수와 양의 관계를, 두 번째 주성분은 상반된 방향의 패턴을 나타냈다. PCA의 장점은 계산이 빠르고 데이터의 주요 구조를 효율적으로 요약할 수 있다는 점이지만, 비선형적 구조나 복잡한 군집 패턴을 반영하기 어렵다는 한계가 있다.

2. MDS (Multi-Dimensional Scaling, 다차원 척도법)

MDS는 데이터 간의 '거리(distance)'를 보존하는 것을 목표로 하는 차원 축소 기법이다. PCA가 분산을 기준으로 작동한다면, MDS는 각 샘플 간의 유사성 또는 차이를 거리로 계산하여 저차원 공간에 재배치한다. 즉, 고차원 공간에서 서로 가까웠던 점들은 2차원에서도 가깝게 유지되도록 배치한다. 유방암 데이터에 MDS를 적용한 결과, 비록 PCA와는 기준이 다르지만 여전히 악성과 양성 그룹이 분리되어 나타났다. 이는 MDS가 원본 공간에서의 상대적 거리 구조를 잘 유지했음을 보여준다. 다만, 계산 비용이 PCA보다 높고, 축(좌표축)에 대한 명확한 해석이 어렵다는 단점이 있다. 그럼에도 샘플 간 관계나 분포 형태를 보존하는 데 효과적이라는 점에서 탐색적 분석에 유용하다.

3. t-SNE (t-distributed Stochastic Neighbor Embedding)

t-SNE는 최근 고차원 데이터 시각화에서 가장 널리 사용되는 비선형 차원 축소 기법이다. 핵심 원리는 '이웃(Neighbor) 관계 보존', 즉 고차원에서 서로 가까운 점들이 저차원에서도 가까이 위치하도록 만드는 것이다. 이는 MDS가 전체적인 거리 구조를 보존하는 것과 달리, 국소(local) 구조를 강조한다. Kaggle 예제에서 t-SNE를 사용한 결과, 두 그룹(악성/양성)이 가장 명확하게 두 둉어리로 분리되었다. 시각적으로도 군집 간의 경계가 뚜렷하게 나타났으며, 이는 t-SNE가 복잡한 비선형 패턴을 효과적으로 드러냈음을 의미한다. 하지만 t-SNE는 계산량이 많고, 축의 의미나 군집 간 거리 해석이 어렵다는 점에서 주의가 필요하다. 따라서 데이터의 군집 구조를 탐색하거나 시각적으로 패턴을 파악할 때 가장 적합하다.