6장. 차원 축소

EURON 6기 초급세션 김수연

6.1 차원 축소 개요

> 💡 **차원 축소**란 매우 많은 피처로 구성된 다차원 데이터 세트의 차원을 축소해 새로운 차원의 데이터 세트를 생성하는 것 → 직관적으로 데이터를 해석할 수 있다!

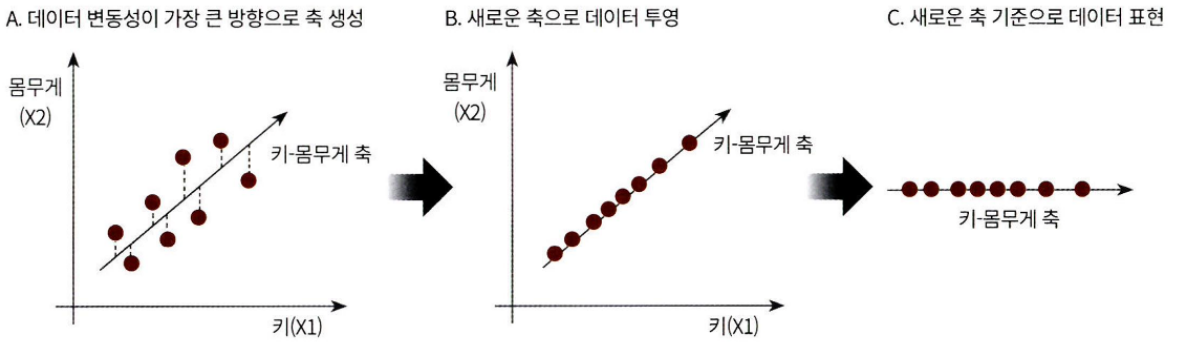
- 피처 선택, 즉 특성 선택은 말 그대로 특정 피처에 종속성이 강한 불필요한 피처는 아예 제거하고, 데이터의 특징을 잘 나타내는 주요 피처만 선택하는 것
- 피처 추출은 기존 피처를 저차원의 중요 피처로 압축해서 추출하는 것. 이렇게 새롭게 추출된 중요 특성은 기존의 피처가 압축된 것이므로 기존의 피처와는 완전히 다른 값이 된다!

6.2 PCA (Principal Component Analysis)

> 💡 **PCA**란 가장 대표적인 차원 축소 기법으로, 여러 변수 간에 존재하는 상관관계를 이용해 이를 대표하는 주성분을 추출해 차원을 축소하는 기법

기존 데이터의 정보 유실이 최소화된다!

이를 위해 PCA는 가장 높은 분산을 가지는 데이터의 축을 찾아 이 축으로 차원을 축소하는데, 이것이 PCA의 주성분이 된다.



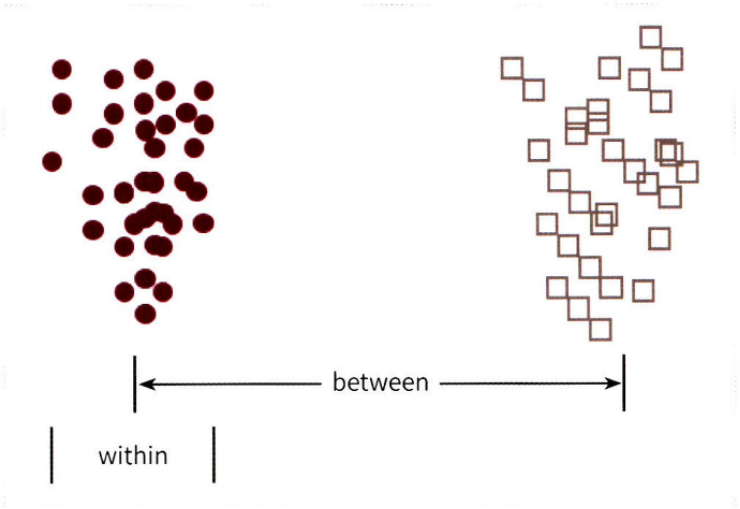
PCA는 제일 먼저 가장 큰 데이터 변동성을 기반으로 첫 번째 벡터 축을 생성하고, 두 번째 축은 이 벡터 축에 직각이 되는 벡터를 축으로 한다. 세 번째 축은 다시 두번째 축과 직각이 되는 벡터를 설정하는 방식으로 축을 생성한다. 이렇게 생성된 벡터 축에 원본 데이터를 투영하면 벡터 축의 개수만큼의 차원으로 원본 데이터가 차원 축소된다.

PCA, 즉 주성분 분석은 이처럼 원본 데이터의 피처 개수에 비해 매우 작은 주성분으로 원본 데이터의 총 변동성을 대부분 설명할 수 있는 분석법이다.

6.3 LDA (Linear Discriminant Analysis)

💡 LDA란 선형 판별 분석법으로 불린다. PCA와 유사하게 입력 데이터 세트를 저차원 공간에 투영해 차원을 축소하는 기법이지만, 지도학습의 분류에서 사용하기 쉽도록 개별 클래스를 분별할 수 있는 기준을 최대한 유지하면서 자원을 축소한다는 점에서 다르다.

PCA는 입력 데이터 변동성의 가장 큰 축을 찾았지만, LDA는 입력 데이터의 결정 값 늘래스를 최대한으로 분리할 수 있는 축을 찾는다.



LDA는 특정 공간상에서 클래스 분리를 최대화하는 축을 찾기 위해 클래스 간 분산과 클래스 내부 분산의 비율을 최대화하는 방식으로 차원을 축소한다. 즉, 클래스 간 분산은 최대한 크게 가져가고, 클래스 내부의 분산은 최대한 작게 가져가는 방식이다.

다음 스텝을 따른다.

1. 클래스 내부와 클래스 간 분산 행렬을 구합니다. 이 두 개의 행렬은 입력 데이터의 결정 값 클래스별로 개별 피처의 평균 벡터(mean vector)를 기반으로 구합니다.
2. 클래스 내부 분산 행렬을 S_W, 클래스 간 분산 행렬을 S_B라고 하면 다음 식으로 두 행렬을 고유벡터로 분해할 수 있습니다.

$$S_W^T S_B = \begin{bmatrix} e_1 & \cdots & e_n \end{bmatrix} \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \cdots & \cdots & \cdots \\ 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} e_1^T \\ \cdots \\ e_n^T \end{bmatrix}$$

3. 고유값이 가장 큰 순으로 K개(LDA변환 차수만큼) 추출합니다.
4. 고유값이 가장 큰 순으로 추출된 고유벡터를 이용해 새롭게 입력 데이터를 변환합니다.

6.4 SVD (Singular Value Decomposition)

💡 SVD 역시 PCA와 유사한 행렬 분해 기법을 이용한다. 정방행렬만을 고유벡터로 분해하는 PCA와 달리, SVD는 정방행렬뿐만 아니라 행과 열의 크기가 다른 행렬에도 적용할 수 있다.

일반적으로 SVD는 mXn 크기의 행렬 A를 다음과 같이 분해하는 것을 의미한다.

$$A = U \sum V^T$$

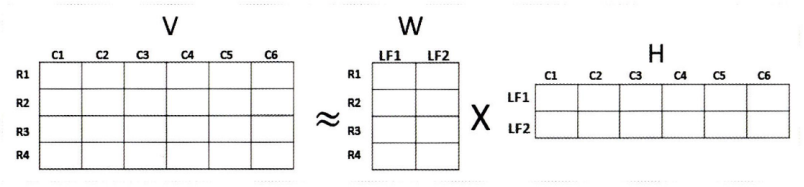
SVD는 특이값 분해로 불리며, 행렬 U와 V에 속한 벡터는 특이벡터(singular vector)이며, 모든 특이벡터는 서로 직교하는 성질을 가집니다. Σ는 대각행렬이며, 행렬의 대각에 위치한 값만 0이 아니고 나머지 위치의 값은 모두 0입니다. Σ이 위치한 0이 아닌 값이 바로 행렬 A의 특이값입니다. SVD는 A의 차원이 m × n일 때 U의 차원이 m × m, Σ의 차원이 m × n, V^T의 차원이 n × n으로 분해합니다.

수학 기호가 많아서 캡쳐로 대체

6.5 NMF (Non-Negative Matrix Factorization)

💡 **NMF**는 Truncated SVD와 같이 낮은 랭크를 통한 행렬 근사 방식의 변형이다.

NMF는 원본 행렬 내의 모든 원소 값이 모두 양수라는 게 보장되면 다음과 같이 좀 더 간단하게 두 개의 기반 양수 행렬로 분해될 수 있는 기법을 지칭한다.



행렬 분해를 하게 되면 W 행렬과 H 행렬은 일반적으로 길고 가는 행렬 W와 작고 넓은 행렬 H로 분해된다. 이렇게 분해된 행렬은 잠재 요소를 특징으로 가지게 된다. 분해 행렬 W는 원본 행에 대해서 이 잠재 요소의 값이 얼마나 되는지에 대응하며, 분해 행렬 H는 이 잠재 요소가 원본 열로 어떻게 구성되었는지를 나타낸다.