**차원축소**

매우 많은 피처로 구성된 다차원 데이터 세트의 차원을 축소해 새로운 차원의 데이터 세트 생성

차원이 증가할수록 데이터 간의 거리가 멀어지고 희소한 구조를 가짐

* 상대적으로 적은 차원에서 학습된 모델보다 예측 신뢰도 떨어짐
* 피처가 많을 경우 개별 피처 간의 상관관계가 높을 가능성이 큰데 선형모델에서는 이로 인한 다중 공선성 문제로 모델의 예측 성능이 저하됨
* 다차원 피처를 차원축소해 피처 수를 줄이면 직관적으로 데이터 해석 가능(시각화, 학습에 필요한 처리 능력 감소)

차원 축소는 두가지로 나눌 수 있음

1. 피처 선택(feature selection): 특정 피처에 종속성이 강한 불필요한 피처는 아예 제거
2. 피처 추출(feature extraction)

* 기존 피처를 저차원의 중요 피처로 압축해 추출, 기존 피처와는 완전히 다름
* 피처를 함축적으로 더 잘 설명할 수 있는 또 다른 공간으로 매핑해 추출
* 잠재적인 요소 추출

PCA, SVD, NMF는 잠재적인 요소를 찾는 대표적인 차원 축소 알고리즘

* 매우 많은 픽셀로 이루어진 이미지 데이터에서 잠재된 특성을 피처로 도출해 함축적 형태의 이미지 변환과 압축 수행
* 분류 수행 시에 과적합 영향력이 작아져서 예측 성능 향상 가능
* 텍스트 구성에서 숨겨져 있는 Semantic 의미나 토픽을 잠재 요소로 간주하고 찾아낼 수 있음

**PCA (Principal Component Analysis)**

여러 변수 간에 존재하는 상관관계를 이용해 이를 대표하는 주성분 추출해 차원 축소

가장 높은 분산을 가지는 데이터의 축을 찾아 이 축으로 차원 축소 🡪 이것이 PCA의 주성분

* 가장 큰 데이터 변동성을 기반으로 첫번째 벡터 축 생성하고, 두번째 축은 이 벡터 축의 직교벡터를 축으로 하고 세번째 축은 다시 두번째 축의 직교 벡터로 축 생성 🡪 이렇게 생성된 벡터 축에 원본 데이터 투영하면 벡터 축의 개수만큼의 차원으로 차원 축소

PCA 즉 주성분 분석은 원본 데이터의 피처 개수에 비해 매우 작은 주성분으로 원본 데이터의 총 변동성을 대부분 설명 가능

선형대수의 관점!

1. 선형변환: 특정 벡터에 행렬을 곱해 새로운 벡터로 변환하는 것, 특정 벡터를 하나의 공간에서 다른 공간으로 투영하는 개념
2. 공분산 행렬: 정방 행렬 & 대칭행렬, 대각선 원소는 각 변수의 분산이고 나머지는 공분산

* 대칭 행렬은 항상 고유벡터를 직교행렬로, 고유값을 정방행렬로 대각화 가능

1. 고유벡터: 행렬을 곱하더라도 방향 변하지 않고 크기만 변하는 벡터, 정방 행렬은 최대 그 차원 수만큼의 고유벡터 가질 수 있음

PCA는 입력 데이터의 공분산 행렬을 고유값 분해하고, 고유벡터에 입력 데이터를 선형 변환하는 것, 이 고유벡터가 PCA의 주성분벡터로서 입력 데이터의 분산이 큰 방향 나타냄

공분산 행렬 = 고유벡터 직교행렬 X 고유값 정방 행렬 X 고유벡터 직교행렬의 전치행렬, ei는 i번째 고유벡터, e1은 가장 분산이 큰 벡터, e2는 e1에 수직이면서 다음으로 가장 분산이 큰 방향을 가진 고유벡터

**입력 데이터의 공분산 행렬이 고유벡터와 고유값으로 분해됨, 분해된 고유벡터를 이용해 입력 데이터를 선형 변환하는 방식이 PCA이다.**

PCA의 STEP

입력 데이터 세트의 공분산 행렬 생성 🡪 공분산 행렬의 고유벡터와 고유값 계산 🡪 고유값이 가장 큰 순으로 K개(PCA 변환 차수만큼)만큼 고유벡터 추출 🡪 고유값이 가장 큰 순으로 추출된 고유벡터를 이용해 새롭게 입력 데이터 변환

PCA는 컴퓨터 비전에 활발히 적용, 얼굴 인식의 경우 Eigen-face라 불리는 PCA 변환으로 원본 얼굴 이미지 변환해 사용

**LDA (Linear Discriminant Analysis)**

선형 판별 분석법으로 불리며 PCA와 유사

PCA와 유사하게 입력 데이터 세트를 저차원 공간에 투영해 차원 축소하는 기법이지만, 분류에서 사용하기 쉽도록 개별 클래스를 분별할 수 있는 기준을 최대한 유지하면서 차원 축소

* PCA는 입력 데이터의 변동성의 가장 큰 축 찾지만 LDA는 **입력데이터의 결정 값 클래스를 최대한으로 분리할 수 있는 축** 찾음
* 클래스 간 분산은 최대한 크게, 클래스 내부 분산은 최대한 작게

LDA step: 클래스 간 분산과 클래스 내부 분산 행렬 구하기(입력 데이터의 결정 값 클래스 별로 개별 피처의 평균 벡터를 기반으로) 🡪 내부 분산 행렬을 SW, 클래스 간 분산 행렬을 SB 🡪 SWT X SB를 고유벡터로 분해 가능 🡪 고유값이 가장 큰 순으로 k개(LDA 변환 차수만큼) 추출 🡪 고유값이 가장 큰 순으로 추출된 고유벡터를 이용해 새롭게 입력 데이터 변환

**SVA (Singular Value Decomposition): 특이값 분해**

PCA는 정방행렬만을 고유벡터로 분해할 수 있지만, SVD는 정방행렬 뿐만 아니라 행과 열의 크기가 다른 행렬에도 적용 가능

A = U X ∑ X VT

행렬U, V에 속한 벡터는 특이 벡터이며 모든 특이 벡터는 서로 직교

∑는 대각 행렬이며 대각에 위치한 값만 0이 아니고 행렬 A의 특이값, 나머지 위치의 값은 0

A의 차원이 m\*n 일 때 U의 차원이 m\*m, ∑ 차원이 m\*n, VT의 차원이 n\*n로 분해됨

Truncated SVD는 대각행렬의 대각 원소 중 상위 몇 개만 추출해서 여기에 대응하는 U, V 원소도 함께 제거해 더욱 차원을 줄인 형태로 분해

* 이렇게 분해하면 인위적으로 더 작은 차원으로 분해하기 때문에 원본 행렬을 정확하게 복원할 수 없지만 근사는 가능
* 원본 차원의 차수에 가깝게 잘라낼수록 원본 행렬에 더 가깝게 복원 가능

PCA와 SVD는 동일한 변환 수행 🡪 PCA는 밀집행렬에 대해서만 변환 가능하지만 SVD는 희소 행렬에 대한 변환도 가능 🡪 SVD는 PCA와 유사하게 컴퓨터 비전 영역에서 이미지 압축 통한 패턴 인식과 신호 처리 분야에 사용됨, 텍스트의 토픽 모델링 기법인 LSA의 기반 알고리즘

**NMF (Non-Negative Matrix Factorization)**

NMF는 Truncated SVD와 같이 낮은 랭크를 통한 행렬 근사 방식의 변형

원본 행렬 내의 모든 원소 값이 모두 양수라는 게 보장되면 두개의 기반 영수 행렬로 분해 가능 🡪 SVD 같은 행렬 분해 기법 통칭

V(4\*6) = W(4\*2) X H(2\*6)

분해된 행렬은 잠재요소를 특성으로 가지게 됨

* 분해행렬 W는 원본 행에 대해서 이 잠재요소의 값이 얼마나 되는지에 대응
* 분해행렬 H는 이 잠재요소가 원본 열(즉, 원본 속성)로 어떻게 구성됐는지 나타내는 행렬

NMF는 SVD와 유사하게 차원 축소를 통한 잠재요소 도출로 이미지 변환 및 압축, 텍스트 토픽 도출 등의 영역에서 사용됨

영화 추천 같은 추천 영역에 활발히 적용 🡪 사용자의 상품 평가 데이터 세트인 사용자-평가 순위 데이터 세트를 행렬 분해 기법을 통해 분해하면서 평가하지 않은 상품에 대한 잠재적인 요소를 추출해 이를 통해 평가 순위 예측하고 높은 순위로 예측된 상품 추천

**정리**

차원축소는 단순히 피처의 개수를 줄이는 개념보다는 이를 통해 데이터 잘 설명할 수 있는 잠재적인 요소 추출하는 데 큰 의미가 있음

PCA는 입력데이터의 변동성이 가장 큰 축 구하고, 다시 이 축에 직각인 축을 반복적으로 축소하려는 차원 개수만큼 구한 뒤 입력 데이터를 이 축들에 투영해 차원 축소 🡪 입력 데이터의 공분산 행렬을 기반으로 고유 벡터 생성하고 이 고유 벡터에 입력 데이터를 선형 변환

LDA는 PCA와 유사, 입력 데이터의 결정 값 클래스를 최대한으로 분리할 수 있는 축 찾는 방식

SVD와 NMF는 매우 많은 피처 데이터를 가진 고차원 행렬을 두 개의 저차원 행렬로 분리하는 행렬 분해 기법, 원본 행렬에서 잠재된 요소 추출하기 때문에 토픽 모델링이나 추천 시스템에 활용