차원 축소**(Dimension Reduction)**

- 매우 많은 피처로 구성된 다차원 데이터 세트의 차원을 축소해 새로운 차원의 데이터 세트를 생성하는 것
- 일반적으로 차원이 증가할수록 데이터 포인트 간의 거리가 기하급수적으로 멀어지게 되고 희소(sparse)한 구조를 가지게 됨
- 수백 개 이상의 피처로 구성된 데이터 세트의 경우 예측 신뢰도가 떨어진
- 피처가 많을 경우 개별 피처 간에 상관관계가 높을 가능성이 큼 -> 선형 회귀와 같은 선형 모델에서는 입력 변수 간의 상관관계가 높을 경우 이로 인한 다중 공선성 문제로 모델의 예측 성능 저하됨
- 차원 축소 -> 더 직관적으로 데이터를 해석할 수 있음
- 차원 축소를 통해 좀 더 데이터를 잘 설명할 수 있는 잠재적인 요소를 추출
- 피처 선택(feature selection), 피처 추출(feature extraction)

피처 선택

- 특정 선택 특정 피처에 종속성이 강한 불필요한 피처는 아예 제거, 데이터의 특징을 잘 나타내는 주요 피러만 선택

피처**(**특성**)** 추출

- 기존 피처를 저차원의 중요 피처로 압축하여 추출
- 기존의 피처와는 완전히 다른 값
- 또 다른 공간으로 매핑해 추출
- 더 함축적인 요약 특성으로 추출, 즉 잠재적인 요소를 추출

PCA(Principal Component Analysis)

- 차원 축소 기법
- 여러 변수 간에 존재하는 상관관계를 이용해 이를 대표하는 주성분을 추출해 차원을 축소
- PCA의 주성분 : 가장 높은 분산을 가지는 데이터의 축, 이 축으로 차원 축소(=분산이 데이터의 특성을 가장 잘 나타내는 것으로 간주)
- 주성분 분석 : 원본 데이터의 피처 개수에 비해 매우 작은 주성분으로 원본 데이터의 총 변동성을 대부분 설명할 수 있는 분석법

과정

- 가장 큰 데이터 변동성을 기반으로 첫번째 벡터 축 생성
- 두번째 축은 이 벡터 축에 직각이 되는 벡터(직교 벡터)를 축으로
- 세번째 축은 두번째 축과 직각이 되는 벡터를 설정하는 방식으로 축 생성
- 생성된 벡터 축에 원본 데이터를 투영하면 벡터 축의 개수만큼의 차원으로 원본 데이터가 차원 축소됨

=> 입력 데이터의 공분산 행렬이 고유 벡터와 고유값으로 분해될 수 있으며 이렇게 분해된 고유벡터를 이용해 입력 데이터를 선형 변환하는 방식

1. 입력 데이터 세트의 공분산 행렬 생성
2. 공분산 행렬의 고유 벡터와 고유값을 계산
3. 고유값이 가장 큰 순으로 K개(PCA 변환 차수만큼)만큼 고유벡터를 추출
4. 고유값이 가장 큰 순으로 추출된 고유 벡터를 이용해 새롭게 입력 데이터를 변환

선형대수

선형 변환

- 특정 벡터에 행렬 A를 곱해 새로운 벡터로 변환하는 것

공분산

- 두 변수 간의 변동 의미

공분산 행렬

- 여러 변수와 관련된 공분산을 포함하는 정방형 행렬
- 개별 분산값을 대각 원소로 하는 대칭행렬
- 정방행렬이며 대칭행렬임
  - 정방행렬 : 열과 행이 같은 행렬
  - 대칭행렬 : 정방행렬 중에서 대각 원소를 중심으로 원소 값이 대칭되는 행렬

고유 벡터

- 행렬 A를 곱하더라도 방향이 변하지 않고 그 크기만 변하는 벡터

LDA(Linear Discriminant Analysis)

- 선형 판별 분석법
- PCA와 유사하게 입력 데이터 세트를 저차원 공간에 투영해 차원 축소
- 지도학습의 분류에서 사용하기 쉽도록 개별 클래스를 분별할 수 있는 기준을 최대한 유지하며 차원 축소
- 입력 데이터의 결정 값 클래스를 최대한으로 분리할 수 있는 축을 찾음
- 특정 공간상에서 클래스 분리를 최대화하는 축을 찾기 위해 클래스 간 분산과 클래스 내부 분산의 비율을 최대화하는 방식으로 차원 축소
- 클래스 간 분산은 최대한 크게, 클래스 내부의 분산은 최대한 작게

과정

1. 클래스 내부와 클래스 간 분산 행렬을 구함, 이 두개의 행렬은 입력 데이터의 결정 값 클래스별로 개별 피처의 평균 벡터를 기반으로 구함
2. 클래스 내부 분산 행렬을 Sw, 클래스 간 분산 행렬을 SB라고 하면 다음 식으로 두 행렬을 고유 벡터로 분해할 수 있음
3. 고유값이 가장 큰 순으로 K개(LDA 변환 차수만큼) 추출

4. 고유값이 가장 큰 순으로 추출된 고유 벡터를 이용해 새롭게 입력 데이터를 변환

SDA(Singular Value Decomposition)

- PCA와 유사한 행렬 분해 기법 사용
- 정방행렬뿐만 아니라 행과 열의 크기가 다른 행렬에도 적용 가능
- m * n 크기의 행렬 A를 분해하는 것을 의미
- 특이값 분해로 불림
- 행렬 U와 V에 속한 벡터는 특이 벡터로 모든 특이 벡터는 서로 직교하는 성질을 가짐

NMF(Non-Negative Matrix Factorization)

- Truncated SVD와 같이 낮은 랭크를 통한 행렬 근사 방식의 변형
- 원본 행렬 내의 모든 원소 값이 모두 양수라는 게 보장되면 다음과 같이 좀 더 간단하게 두 개의 기반 양수 행렬로 분해될 수 있는 기법을 지칭