[Week7] 파머완 6장

CHAPTER 06 차원 축소

01 차원 축소(Dimension Reduction) 개요

차원 축소

- 매우 많은 피처로 구성된 다차원 데이터 세트의 차원을 축소해 새로운 차원의 데이터 세트를 생성하는 것
- 차원이 증가할수록 데이터 포인트 간의 거리가 기하급수적으로 멀어지게 되고 희소(sparse)한 구조를 갖게 됨
- 더 직관적으로 데이터 해석 가능
- 피처 선택(feature selection) : 특정 피처에 종속성이 강한 불필요한 피처는 제거 & 데이터의 특징을 잘 나타내는 주요 피처만 선택
- 피처 추출(feature extraction)
  - 기존 피처를 저차원의 중요 피처로 압축해서 추출 → 기존의 피처와 완전히 다른 값
  - 단순 압축이 아닌 피처를 함축적으로 더 잘 설명할 수 있는 또 다른 공간으로 매핑해 추출하는 것
- 이미지 변환과 압축, 텍스트 의미 추출 등에 사용

02 PCA(Principal Component Analysis)

PCA 개요

- 여러 변수 간에 존재하는 상관관계를 이용해 이를 대표하는 주성분을 추출하여 차원 축소
- 기존 데이터의 정보 유실 최소화 → 가장 높은 분산을 가지는 데이터 축을 찾아 이 축으로 차원 축소 ⇒ 주성분 = 분산
- 가장 큰 데이터 변동성 기반으로 첫번째 벡터 축 생성 → 이 벡터 축에 직각이 되는 직교 벡터를 두번째 벡터 축으로 생성 → 다시 두번째 축과 직각이 되는 벡터를 설정하는 방식으로 세번째 축 생성

- 원본 데이터의 피처 개수에 비해 매우 작은 주성분으로 원본 데이터 총 변동성을 대부분 설명 가능
- 입력 데이터의 공분산 행렬을 고유값 분해 → 고유벡터에 입력데이터를 선형 변환
  - 고유벡터 : PCA의 주성분 벡터, 입력 데이터의 분산이 큰 방향을 나타냄, 행렬을 곱해도 방향은 변하지 않고 크기만 변함
    - 고유벡터는 여러 개 존재
    - 정방행렬은 최대 그 차원 수만큼의 고유벡터를 가질 수 있음
  - 고윳값(eigenvalue) : 고유벡터의 크기. 입력 데이터의 분산
  - 선형 변환 : 특정 벡터에 행렬 A를 곱해 새로운 벡터로 변환하는 것
  - 공분산 행렬 : 여러 변수와 관련된 공분산을 포함하는 정방형 행렬, 대칭행렬
    - 공분산행렬$C = P \sum P^T$
    - P = n X n 직교행렬
    - $\sum$ = n X n 정방행렬
- 입력 데이터의 공분산 행렬이 고유벡터와 고유값으로 분해될 수 있고 이렇게 분해된 고유벡터를 이용해 입력 데이터를 선형 변환하는 방식이 PCA!

PCA 스탭

1. 입력 데이터 세트의 공분산 행렬 생성
2. 공분산 행렬의 고유벡터와 고유값 생성
3. 고유값이 가장 큰 순으로 K개(PCA 변환 차수)만큼 고유벡터를 추출
4. 고유값이 가장 큰 순으로 추출된 고유벡터를 이용해 새롭게 입력 데이터 변환

사이킷런 PCA 클래스

- n_components : PCA로 변환할 차원의 수
- explained_variance_ratio_ : 전체 변동성에서 개별 PCA 컴포넌트별로 차지하는 변동성 비율 정보 제공

03 LDA(Linear Discriminant Analysis)

LDA 개요

- 선형 판별 분석법
- 입력 데이터 세트를 저차원 공간에 투영해 차원 축소 → PCA와 매우 유사
- 지도학습의 분류에서 사용하기 쉽도록 개별 클래스를 분별할 수 있는 기준을 최대한 유지하면서 차원 축소 → 입력 데이터의 결정 값 클래스를 최대한으로 분리할 수 있는 축 탐색
- 클래스 간 분산과 클래스 내부 분산의 비율을 최대화하는 방식으로 차원 축소
- 클래스 간 분산은 최대한 크게, 클래스 내부 분산은 최대한 작게 → 내부 분산 행렬 생성 후 투영

LDA 스탭

1. 클래스 내부와 클래스 간 분산 행렬 계산 → 입력 데이터의 결정값 클래스별로 개별 피처의 평균 벡터를 기반으로 구함
2. 클래스 내부 분산 행렬 S_W, 클래스 간 분산 행렬 S_B를 고유 벡터로 분해
3. 고유값이 가장 큰 순으로 K개 추출
4. 고유값이 가장 큰 순으로 추출된 고유벡터를 이용해 새롭게 입력 데이터 변환

사이킷런 LDA 클래스

- LinearDiscriminantAnalysis 클래스

04 SVD(Singular Value Decomposition)

SVD 개요

- 정방행렬뿐만 아니라 행과 열의 크기가 다른 행렬에도 적용 가능
- $A = U \sum V^T$ (A = m x n 행렬)
- 특이값 분해
- 행렬 U와 V에 속한 벡터는 특이 벡터 : 서로 직교하는 성질
- $\sum$: 대각행렬
- 특이값 : 대각행렬에서 0이 아닌 값

SVD 모듈

- numpy.linalg.svd
- 사이킷런 TruncatedSVD 클래스

05 NMF(Non-Negative Matrix Factorization)

NMF 개요

- 낮은 랭크를 통한 행렬 근사 방식의 변형
- 사이킷런 NMF 클래스