10주차_6장. 차원축소 (6.1~6.5)

상태	진행 중

06. 차원 축소

06-1 차원 축소 개요

- 多피처 다차원 데이터 세트 차원 축소 → 새로운 차원의 데이터 세트 생성
  - 차원 ↑ - 데이터 간 거리 ↑ - '희소' 구조
  - 多피처 모델 신뢰도 < 少피처 모델 신뢰도
  - 피처 多 - 개별 피처 간 상관관계 ↑
- 직관적 데이터 해석 可, 학습 필요 처리량 ↓
- 일반적 차원 축소 : 피처 선택 / 피처 추출

  ▼ 피처 선택 feature selection
  - 특정 피처에 종속성이 강한, 불필요한 피처는 아예 제거
  - 데이터 잘 나타내는 주요 피처만 선택

  ▼ 피처 추출
  - 기존 피처 → 저차원 중요 피처 압축 추출 (완전히 다른 값 됨)
- 단순한 데이터 압축이 아니라, 데이터를 더 잘 설명할 수 있는 잠재적 요소를 추출하는 것

06-2 PCA(Principal Component Analysis)

: 변수 간 상관관계 → 이를 대표하는 주성분 추출 → 차원 축소

PCA 개요

- 기존 데이터 정보 유실 최소화
  - ← 가장 높은 분산 가지는 데이터의 축으로 차원 축소 = **PCA의 주성분**

- 첫 번째 축 : 가장 큰 데이터 변동성(Variance) 기반
- 두 번째 축 : 첫 번째 축의 직교 벡터
- 세 번째 축 : 두 번째 축의 직교 벡터

→ 이렇게 생성된 벡터 축에 원본 데이터 투영하면, 벡터 축 개수만큼의 차원으로 데이터가 차원축소됨

선형대수 관점 해석

1. 입력 데이터의 공분산(두 변수 간 변동) 행렬(Covariance Matrix) 고유값 분해
2. 구한 고유벡터에 입력 데이터 선형 변환
   - 고유벡터 = PCA 주성분 벡터 = 입력 데이터 분산이 큰 방향
   - 고윳값 = 고유벡터의 크기 = 입력 데이터의 분산

PCA 단계

1. 입력 데이터 세트의 공분산 행렬 생성
2. 공분산 행렬의 고유벡터/고유값 계산
3. 고유값 가장 큰 순으로 K개만큼=PCA변환 차수만큼 고유벡터 추출
4. 고유값 가장 큰 순으로 추출된 고유벡터를 이용해 입력 데이터 새롭게 변환

06-3 LDA (Linear Discriminant Analysis)

LDA 개요

- 선형 판별 분석법. PCA와 동일하게 입력 데이터를 저차원에 투영해 차원을 축소하지만, 지도학습의 classification에서 사용하기 쉽도록 개별 클래스를 분별할 수 있는 기준을 최대한 유지하며 차원 축소
- 입력 데이터의 결정 값 클래스를 최대한으로 분리할 수 있는 축 찾음
  - 클래스 간 분산(between-class scatter)과 클래스 내부 분산(within-class scatter)의 비율을 최대화
  - 클래스 간 분산 최대화, 클래스 내부 분산 최소화

LDA 단계

1. 클래스 내부와 클래스 간 분산 행렬 구함. 이 두 개의 행렬을 결정 값 클래스별로 개별 피처의 평균 벡터를 기반으로 구함

2. 클래스 내부 분산 행렬과 클래스 간 분산 행렬을 기반으로 두 행렬을 고유벡터로 분해
3. 고유값이 가장 큰 순으로 K개(LDA 변환 차수만큼) 추출
4. 고유값이 가장 큰 순으로 추출된 고유벡터 이용해 새롭게 입력 데이터 변환

06-4 SVD (Singular Value Decomposition)

SVD 개요

: PCA와 유사한 행렬 분해 기법 이용하지만, 정방행렬만을 고유벡터로 분해하는 PCA와 달리 SVD는 행과 열이 다른 행렬에도 적용 가능

- = 특이값 분해, 각 행렬에 속한 벡터 = 특이 벡터
- 모든 특이 벡터는 서로 직교

사이킷런 TruncatedSVD 클래스를 이용한 변환

- fit(), transform() 호출해 원본 데이터의 주요 컴포넌트로 차원 축소해 변환.

06-5 NMF (Non-Negative Matrix Factorization)

: 랭크를 통한 행렬 근사(Low-Rank Approximation) 방식의 변형.

- 원본 행렬 내 모든 원소 값이 모두 양수라는 보장되면, 더 간단한 두 개의 양수 행렬로 분해될 수 있는 기법 지칭 = 행렬 분해(Matrix Factorization)
- 분해된 행렬은 잠재 요소를 특성으로 가짐
  - W : 원본 행에 대해 이 잠재 요소의 값이 얼마나 되는지 대응
  - H : 이 잠재 요소가 원본 열(원본 속성)로 어떻게 구성됐는지 나타냄
- SVD와 유사하게 이미지 변환, 텍스 추출 등에 사용