Week 10 예습과제

파이썬 머신러닝 완벽가이드 6장 필사 & 개념정리

(6.1 ~ 6.5)

01 차원축소

- 차원 축소: 매우 많은 피처로 구성된 다차원 데이터 세트의 차원을 축소해 새로운 차원의 데이터 세트를 생성하는 것
    - 일반적으로 차원이 증가 → 데이터 포인트 간의 거리가 기하급수적으로 멀어짐 → 희소(sparse)한 구조를 가지게 됨
- 차원 축소는 피처 선택(feature selection)과 피처 추출(feature extraction)으로 나뉨
    - 피처 선택: 특정 피처에 종속성이 강한 불필요한 피처 제거 및 주요 피처만 선택
        - 기존 피처를 저차원의 중요 피처로 압축해서 추출 → 기존의 피처가 압축된 것이므로 기존의 피처와는 완전히 다른 값
    - 피처 추출: 기존 피처를 단순 압축이 아닌, 피처를 함축적으로 더 잘 설명할 수 있는 또 다른 공간 매핑 및 추출
        - 피처 추출은 기존 피처가 전혀 인지하기 어려웠던 잠재적인 요소를 추출
- PCA, SVD, NMF은 이처럼 잠재적인 요소를 찾는 대표적인 차원 축소 알고리즘
    - 매우 많은 픽셀로 이뤄진 이미지 데이터에서 잠재된 특성을 피처로 도출
    - 텍스트 문서의 숨겨진 의미 추출 - 단어들의 구성에서 숨겨져 있는 시맨틱(Semantic)의미나 토픽(Topic)을 잠재 요소 간주하고 찾아냄

02 PCA(Principal Component Analysis)

- PCA: 가장 대표적인 차원 축소 기법으로 여러 변수 간에 존재하는 상관관계를 이용해 이를 대표하는 주성분(Principal Component)을 추출해 차원을 축소하는 기법
    - PCA 차원 축소로 기본 데이터 정보 유실이 최소화된다. PCA는 가장 높은 분산을 가지는 데이터의 축을 찾아 이 축으로 차원 축소 → 이것이 PCA의 주성분

    - (분산이 데이터의 특성 가장 잘 나타낸다함)
    - 축소 순서:

    1. 가장 큰 데이터 변동성을 기반으로 첫 번쨰 벡터 축 생성
    2. 두 번째 축은 이 벡터 축에 직각이 되는 벡터 (직교 벡터)를 축으로 함
    3. 세 번째 축을 다시 두 번째 축과 직각이 되는 벡터를 설정하는 방식으로 축 생성
    4. 생성된 벡터 출에 원본 데이터 투영 → 벡터 축의 개수만큼의 차원으로 원본 데이터 차원 축소

    - PCA 순서:

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

- PCA의 또 다른 적용 영역: 컴퓨터 비전(Computer Vision) 분야
    - 특히 얼굴 인식의 경우 Eigen-face라고 불리는 PCA 변환으로 원본 얼굴 이미지를 변환해 사용하는 경우 많다

03 LDA(Linear Discriminant Analysis)

- LDA: 선형 판별 분석법; PCA와 유사하게 입력 데이터 세트를 저차원 공간에 투영해 차원 축소하는 기법
    - 중요한 차이: LDA는 지도학습의 분류(Classification)에서 사용하기 쉽도록 개별 클래스를 분별할 수 있는 기준을 최대한 유지하면서 차원 축소
    - PCA는 입력 데이터의 변동성의 가장 큰 축 찾음
    - LDA는 입력 데이터의 결정 값 클래스를 최대한으로 분리할 수 있는 축 찾음
    - LDA는 특정 공간상에서 클래스 분리를 최대화하는 축을 찾기 위해 클래스 간 분산(between-class scatter)과 클래스 내부 분산(within-class scatter)의 비율을 최대화하는 방식으로 차원 축소
    - 즉, 클래스 간 분산은 최대한 크게 가져가고, 클래스 내부의 분산 작게 가져가는 방식

- LDA 구하는 스텝

1. 클래스 내부와 클래스 간 분산 행렬 구하기. 이 두 개의 행렬은 입력 데이터의 결정 값 클래스별로 개별 피처의 평균 벡터(mean vector)를 기반으로 구함
2. 클래스 내부 분산 행렬을 SW, 클래스 간 분산 행렬을 SB라고 한다
3. 고유 값이 가장 큰 순으로 K개(LDA변환 차수만큼) 추출
4. 고유값이 가장 큰 순으로 추출된 고유벡터를 이용해 새롭게 입력 데이터를 변환

04 SVD(Singular Value Decomposition)

- SVD: 정방행렬(PCA) 뿐만 아니라 행과 열의 크기가 다른 행렬에도 적용 가능한 행렬 분해 기법
- 일반적으로 SVD는 mxn 크기의 행렬 A를 다음과 같이 분해하는 것을 의미
- SVD는 특이값 분해로 불리며, 행렬 U와 V에 속한 벡터는 특이벡터(singular vector)이며, 모든 특이 벡터는 서로 직교하는 성질 가진다.
- SVD는 PCA와 유사하게 컴퓨터 비전 영역에서 이미지 압축을 통한 패턴 인식과 신호 처리 분야에 사용됩니다. 또한 텍스트의 토픽 모델링 기법인 LSA (Latent Semantic Analysis)의 기반 알고리즘이다

05 NMF(Non-Negative Matrix Factorization)

- NMF: Truncated SVD와 같이 낮은 랭클르 통한 행렬 근사(Low-Rank Approximation)방식의 변형
- IF NMF는 원본 행렬 내의 모든 원소 값이 모두 양수 → 더 간단한 두 개의 기반 양수 행렬로 분해
    - 분해된 행렬은 잠재 요소를 특성으로 가짐
    - NMF는 SVD와 유사하게 차원 축소를 통해 잠재 요소 도출로 이미지 변환 및 압축, 텍스트의 토픽 도출 등의 영역에서 사용
- NMF는 또 문서 유사도 및 클러스터링에 잘 사용됨
    - + 영화 추천과 같은 ㅊ춴 영역, 사용자의 상품 평가 데이터 세트인 사용자-평가 순위에서 평가하지 않은 상품에 대한 잠재적인 요소 추출 → 평가 순위 예측 → 높은 순위로 예측된 상품 추천
    - 이를 잠재요소 (Latent Factoring) 기반의 추천 방식이라고 함