EURON 9주차 과제

01. 차원 축소(Dimension Reduction) 개요

차원 축소란?

- 매우 많은 피처로 구성된 다차원 데이터 세트의 차원을 축소해 새로운 차원의 데이터 세트를 생성하는 것
- 일반적으로 차원이 증가할수록 데이터 포인트 간의 거리가 기하급수적으로 멀어지게 되어 희소(sparse)한 구조를 가지게 됨
- 피처가 많을수록 예측 신뢰도가 떨어지고, 선형 모델에서는 다중 공선성 문제로 예측 성능이 저하됨

대표적인 차원 축소 알고리즘

- PCA
- LDA
- SVD
- NMF

피처 선택(feature selection)

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

피처 추출(feature extraction)

- 기존 피처를 저차원의 중요 피처로 압축해 추출하는 것
- 새롭게 추출된 중요 특성은 기존의 피처와 완전히 다른 값
- 단순 압축X, 피처를 함축적으로 더 잘 설명할 수 있는 또 다른 공간으로 매핑해 추출하는 것

차원 축소를 통해 좀 더 데이터를 잘 설명할 수 있는 잠재적인 요소를 추출하는 것이 핵심

- PCA, SVD, NMF 가 이에 해당

차원 축소 알고리즘이 사용되는 영역

▼ 이미지 데이터

- 많은 픽셀로 이루어진 이미지 데이터에서 잠재된 특성을 피처로 도출해 함축적 형태의 이미지 변환과 압축 수행
- 이렇게 변환된 이미지는 원본 이미지보다 훨씬 적은 차원이므로 분류 수행 시에 과적합 영향력이 작아져 예측 성능을 끌어올릴 수 있음

▼ 텍스트 문서

- 문서는 많은 단어로 구성, 의미나 의도를 가지고 문서를 작성하면서 단어 사용
- 차원 축소 알고리즘은 문서 내 단어들의 구성에서 숨겨져 있는 시맨틱 의미나 토픽을 잠재 요소로 간주하고 찾아냄
- SVD와 NMF는 이러한 시맨틱 토픽 모델링을 위한 기반 알고리즘으로 사용

02. PCA(Principal Component Analysis)

PCA 개요

- 가장 대표적인 차원 축소 기법
- 여러 변수 간에 존재하는 상관관계를 이용해 이를 대표하는 주성분을 추출해 차원을 축소
- 기존 데이터의 정보 유실이 최소화
- 가장 높은 분산을 가지는 데이터의 축을 찾아 이 축으로 차원을 최소화 → PCA의 주성분
- 가장 큰 데이터 변동성을 기반으로 첫 번째 데이터 축 생성 → 두 번째 축은 직교 벡터를 축으로 함 → 세 번째 축은 다시 두 번째 축과 직각이 되는 벡터를 설정하는 방식으로 축 생성
- 벡터 축의 개수만큼의 차원으로 원본 데이터가 차원 축소됨
- 원본 데이터의 피처 개수에 비해 매우 작은 주성분으로 원본 데이터의 총 변동성을 대부분 설명할 수 있음

선형 대수 관점

- 입력 데이터의 공분산 행령을 고유값 분해, 이렇게 구한 고유벡터에 입력 데이터 선형 변환
- 고유벡터가 PCA의 주성분 벡터로서 입력 데이터의 분산이 큰 방향을 나타냄
- 고윳값은 고유벡터의 크기를 나타내며, 동시에 입력 데이터의 분산을 나타냄
- 입력 데이터의 공분산 행렬이 고유벡터와 고유값으로 분해될 수 있으며, 이렇게 분해된 고유벡터를 이용해 입력 데이터를 선형 변환하는 방식이 PCA
- 다음과 같은 스텝으로 수행됨
    1. 입력 데이터 세트의 공분산 행렬을 생성
    2. 공분산 행렬의 고유벡터와 고유값 계산
    3. 고유값이 가장 큰 순으로 K개(PCA 변환 차수)만큼 고유벡터 추출
    4. 고유값이 가장 큰 순으로 추출된 고유벡터를 이용해 새롭게 입력 데이터 변환

03. LDA(Linear Discriminant Analysis)

LDA 개요

- 선형 판별 분석법, PCA와 매우 유사
- 입력 데이터 세트를 저차원 공간에 투영해 차원 축소
- 지도학습의 분륭서 사용하기 쉽도록 개별 클래스를 분별할 수 있는 기준을 최대한 유지하면서 차원을 축소
- 입력 데이터의 결정 값 클래스를 최대한으로 분리할 수 있는 축을 찾음
- 클래스 간 분산과 클래스 내부 분산의 비율을 최대화하는 방식으로 차원 축소

  → 클래스 간의 분산은 최대한 크게 가져가고, 클래스 내부의 분산을 최대한 작게 가져감

LDA를 구하는 스텝은 다음과 같음

1. 클래스 내부와 클래스 간 분산 행렬을 구합니다. 이 두 개의 행렬은 입력 데이터의 결정 값 클래스별로 개별 피처의 평균 벡터를 기반으로 구합니다.
2. 클래스 내부 분산 행렬을 S_W, 클래스 간 분산 행렬을 S_B라고 하면 다음 식으로 두 행렬을 고유벡터로 분해할 수 있음 (식 416pg 참고)

3. 고유값이 가장 큰 순으로 K개(LDA변환 차수만큼) 추출합니다.
4. 고유값이 가장 큰 순으로 추출된 고유벡터를 이용해 새롭게 입력 데이터를 변환합니다.

04. SVD(Singular Value Decomposition)

SVD 개요

- PCA와 유사한 행렬 분해 기법 이용
- 정방행렬뿐만 아니라 행과 열의 크기가 다른 행렬에도 적용할 수 ㅣㅇㅆ음
- 일반적으로 m*n 크기의 행렬 A를 다음과 같이 분해하는 것을 의미(식 418pg 참고)
- 특이값 분해

05. NMF(Non-Negative Matrix Factorization)

NMF 개요

- Truncated SVD와 같이 낮은 랭크를 통한 행렬 근사 방식의 변형
- 원본 행렬 내의 모든 원소 값이 모두 양수라는게 보장되면 좀 더 간단하게 두 개의 기반 양수 행렬로 분해될 수 있는 기법을 지칭
- 분해된 행렬은 잠재 요소를 특성으로 가지게 됨
- 차원 축소를 통한 잠재 요소 도출로 이미지 변환 및 압축, 텍스트의 토픽 도출 등의 영역에서 사용

06. 정리

- 많은 피처로 이뤄진 데이터 세트를 PCA같은 차원 축소를 통해 더욱 직관적으로 이해할 수 있음
- 차원 축소는 단순히 피처의 개수를 줄이는 개념보다 이를 통해 데이터를 잘 설명할 수 있는 잠재적인 요소를 추출하는 데 큰 의미가 있음
- 많은 차원을 가지는 이미지나 텍스트에 차원 축소 알고리즘이 활발하게 사용

PCA

- 입력 데이터의 변동성이 가장 큰 축을 구하고, 다시 이 축에 직각인 축을 반복적으로 축소하려는 차원 개수만큼 구한 뒤 입력 데이터를 이 축들에 투영해 차원을 구하는 방식
- 입력 데이터의 공분산 행렬을 기반으로 고유 벡터를 생성하고 이렇게 구한 고유 벡터에 입력 데이터를 선형 변환하는 방식

LDA

- PCA와 매우 유사한 방식
- 입력 데이터의 결정 값 클래스를 최대한으로 분리할 수 있는 축을 찾는 방식으로 차원 축소

SVD & NMF

- 매우 많은 피처 데이터를 가진 고차원 행렬을 두 개의 저차원 행렬로 분리하는 행렬 분해 기법
- 토픽 모델링이나 추천 시스템에서 활발하게 사용