[7주차 예습과제] - 6단원

6.1

- PCA, LDA, SVD, NMF에 대해서 배울 예정.
- 차원 축소는 매우 많은 피처로 구성된 다차원 세트의 차원을 축소해 새로운 차원의 데이터 세트를 생성하는 것임.
- 일반적으로 차원이 증가할수록 데이터 포인트 간의거리가 기하급수적으로 멀어지게 되고 희소한 구조를 가지게 됨.
  -> 수백 개 이상의 피처로 구성되니 데이터 세트의 경우 상대적으로 적은 차원에서 학습된 모델보다 예측 신뢰도가 떨어짐.
  -> 피처가 많을 경우개별 피처간에 상관관계가 높을 가능 성이 큼. 선형 회귀와 같은 선형 회귀 모델에서는 입력 변수 간의 상관관계가 높을 경우 이로 인한 다중 공선성 문제로 모델의 예측 성능이 저하됨.

=> 매우 많은 다차원의 피처를 차원 축소해 피처 수를 줄이면 더 직관적으로 데이터를 해석하 수 있음.

일반적으로 차원 축소는 피처 선택과 피처 추출로 나눌 수 있음.

피처 선택 : 즉 특성 선택은 말 그대로 특정 피처에 종속성이 강한 불필요한 피처는 아예제거하고 데이터의 특징을 잘 나타내는 주요 피처만 선택하는 것임.

피처추출 : 기존 피처를 저차원의 중요 피처로 압축해서 추출하는 것임. 이렇게 새롭게 추출된 중요 특성은 기존의 피처가 압축된 것이므로 기존의 피처와는 완전히 다른 값이 됨.
압축을 할 때는, 단순 압축이 아닌 피처를 함축적으로 더 잘 설명할 수 있는 또 다른 공간으로 매핑해 추출함. 함축적인 특성 추출은 기존 피처가 전혀 인지하기 어려웠던 잠재적인 요소를 추출하는 것을 의미함.
ex. 이미지 데이터, 텍스트 문서

6.2

PCA :
- 가장 대표적인 차원 축소 기법.
- 여러 변수 간에 존재하는 상관관계를 이용해 이를 대표하는 주성분을 추출해 차원을 축소하는 기법.
- PCA로 차원을 축소할 때는 기존 데이터의 정보 유실이 최소화 됨.
- PCA는 가장 높은 분산을 가지는 데이터의 축을 찾아 이 축으로 차원을 축소하는데, 이것이 PCA의 주성분이 됨.
- PCA는 제일 먼저 가장 큰 데이터 변동성을 기반으로 첫 번째 벡터 축을 생성하고, 두 번째 축은 이 벡터 축에 직각이 되는 벡터를 축으로 함. 세 번째 축은 다시 두 번째 축과 직각이 되는 벡터를 설정하는 방식으로 축을 생성함. 이렇게 생성된 벡터 축에 원본 데이터를 투영

하면 벡터 축의 개수만큼의 차원으로 원본 데이터가 차원 축소됨.
- PCA를 선형대수 관점에서 해석 해보면, 입력 데이터의 공분산 행렬을 고유값 분해하고, 이렇게 구한 고유 벡터에 입력 데이터를 선형 변환하는 거임. 이 고유 벡터가 PCA의 주성분 벡터로서 입력 데이터의 분산이 큰 방향을 나타냄. 고유값은 바로 이 고유 벡터의 크기를 나타내며, 동시에 입력 데이터의 분산을 나타냄.

● 선형변환
특정 벡터에 행렬 A를 곱해 새로운 벡터로 변환. 이를 특정 벡터를 하나의 공간에서 다른 공간으로 투영하는 개념으로 볼 수 있으며, 이 경우 이 행렬을 바로 공간으로 가정하는 것임.

● 공분산
보통 분산은 한 개의 특정한 변수의 데이터 변동을 의미하나, 공분산은 두 변수 간의 변동을 의미함.
공분산 행렬은 여러 변수와 관련된 공분산을 포함하는 정방형 행렬임.

● 고유 벡터
행렬 A를 곱하더라도 방향이변하지 않고 그 크기만 변하는 벡터.
고유 벡터는 여러 개 존재하며, 정방 행렬은 최대 그 차원 수만큼의 고유 벡터를 가질 수 있음.

● 공분산 행렬
정방 행렬이며 대칭 행렬임.
정방 행렬 : 열과 행이 같은 행렬
대칭 행렬 : 정방 행렬 중에서 대각 원소를 중심으로 원소 값이 대칭되는 행렬. 대칭 행렬은 항상 고유 벡터를 직교 행렬로, 고유값을 정방 행렬로 대각화 할 수 있음.

PCA 수행 스텝
1. 입력 데이터 세트의 공분산 행렬을 생성함.
2. 공분산 행렬의 고유 벡터와 고유값을 계산함.
3. 고유값이 가장 큰 순으로 K개만큼 고유 벡터를 추출함.
4. 고유값이 가장 큰 순으로 추출된 고유벡터를 이용해 새롭게 입력 데이터를 변환함.

*사이킷런은 PCA 변혼을 위해 PCA 클래스를 제공함. PCA 클래스는 생성 파라미터로 n_components를 입력받음.

*PCA는 차원 축소를 통해 데이터를 쉽게 인지하는 데 활용할 수 있지만, 이보다 더 활발하게 적용되는 영역은 컴퓨터 비전 분야임 .특히 얼굴 인식의 경우 Eigen-face라고 불리는 PCA 변환으로 원본 얼굴 이미지를 변환해 사용하는 경우가 많음.

6.3

LDA

- 선형 판별 분석법으로 불리며, PCA와 매우 유사함.
- PCA와 유사하게 입력 데이터 세트를 저차원 공간에 투영해 차원을 축소하는 기법이지만, 중요한 차이는 LDA는 지도학습의 분류에서 사용하기 쉽도록 개별 클래스를 분별할 수 있는 기준을 최대한 유지하면서 차원을 축소함.
- 입력 데이터의 결정 값 클래스를 최대한 분리할 수 있는 축을 찾음.
- 특정 공간상에서 클래스 분리를 최대화하는 축을 찾기 위해 클래스 간 분산과 클래스 내부 분산의 비율을 최대화하는 방식으로 차원을 축소함.
  -> 클래스 간 분산은 최대한 크게 가져가고, 클래스 내부 분산은 최대한 작게 가져가는 분산임.

일반적으로 LDA를 구하는 스텝은 PCA와 유사하나 가장 큰 차이점은 공분산 행렬이 아니라 위에 설명한 클래스 간 분산과 클래스 내부 분산 행렬을 생성하 뒤, 이 행렬에 기반해 고유 벡터를 구하고 입력 데이터를 투영한다는 점임.

LDA 수행 스텝

1. 클래스 내부와 클래스 간 분산 행렬을 구함. 이 두 개의 행렬은 입력 데이터의 결정 값 클래스별로 개별 피처의 평균 벡터를 기반으로 구함.
2. 클래스 내부 분산 행렬은 Sw, 클래스 간 분산 행렬을 Sb라고 하면 다음 식으로 두 행렬의 고유벡터로 분해할 수 있음.(p.416 식)
3. 고유값이 가장 큰 순으로 K개(LDA변호나 차수만큼) 추출함.
4. 고유값이 가장 큰 순으로 추출된 고유벡터를 이용해 새롭게 입력 데이터를 변환함.

6.4

SVD

- PCA와 유사한 행렬 분해 기법을 이용함.
- PCA의 경우 정방행렬만을 고유벡터로 분해할 수 있지만, SVD는 정방행렬뿐만 아니라 행과 df의 크기가 다른 행렬에도 적용할 수 있음.
- SVD는 특이값 분해로 불리며, 행렬 U와 V에 속한 벡터는 특이 벡터이며, 모든 특이 벡터는 서로 직교하는 성질을 가짐.
- sigma는 대각 행렬이며, 행렬의 대각에 위차한 값만 0이 아니고 나머지 위치의 값은 모두 0임. sigma가 위치한 0이 아닌 값이 바로 행렬 A의 특이값임.

● Truncated SVD

sigma의 대각원소 중에 상위 몇 개만 추출해서 여기에 대응하는 U와 V의 원소도 함께 제거해 더욱 차원을 줄인 형태로 분해하는 것임.

일반적으로 넘파이나 사이파이 라이브러리를 이용함.

-> 이렇게 분해하면, 인위적으로 더 작은 차원의 U, sigma, Vt로 분해하기 때문에 원본 행렬을 정확하게 다시 원복할 수 없음. 하지만, 데이터 정보가 압축되어 분해됨에도 불구하고 상

당한 수준으로 원본 행렬을 근사할 수 있음.
원래 차원의 차수에 가깝게 잘라낼수록 원본 행렬에 더 가깝게 복원할 수 있음.

사이킷런의 TruncatedSVD 클래스는 사이파이 svds와 같이 Truncated SVD 연산을 수행해 원본 행렬을 분해한 U, Sigma, Vt 행렬을 반환하지는 않음. 시ㅏ이킷런의 TruncatedSVD 클래스는 PCA 클래스와 유사하게 fit()와 transform()을 호출해 원본 데이터를 몇 개의 주요 컴포넌트로 차원을 축소해 변환함. 원본 데이터를 Truncated SVD 방식으로 분해된 U*Sigma 행렬에 선형 변혼해 생성함.

6.5

NMF는 Truncated SVD와 같이 낮은 랭크를 통한 행렬 근사 방식의 변형임.
NMR는 원본 행렬 내의 모든 원소 값이 모두 양수라는 게 보장되면 좀 더 간단하게 두 개의 기반 양수 행렬로 분해될 수 있는 기법을 지칭함.

● 행렬 분해
일반적으로 SVD와 같은 행렬 분해 기법을 통칭함.
W 행렬과 H 행렬은 일반적으로 길고 나는 행렬 W와 작고 넓은 행렬 H로 분해됨.
이렇게 분해된 행렬은 잠재 요소를 특성으로 가지게 됨 .

NMF도 SVD와 유사하게 이미지 압축을 통한 패턴 인식, 텍스트의 토픽 모델링 기법, 문서 유사도 및 클러스터링에 잘 사용됨.
영화 추천과 같은 추천 영역에 활발하게 적용됨.