6주차 발표 요약

#01 차원 축소 개요

- 차원 축소: 다차원 데이터 세트의 차원을 축소해 새로운 차원의 데이터 셋을 생성하는 것 →매우 많은 다차원의 피처를 차원 축소해 피처 수를 줄이면 더 직관적으로 데이터를 해석할 수 있음
- 피처 선택과 피처 추출: 데이터의 특징을 잘 나타내는 주요 피처만 선택 vs 또 다른 공간으로 매핑해 저차원의 중요 피처로 압출해서 추출(완전히 다른 값이라 볼 수 있음.
- 차원 축소 알고리즘: PCA, SVD, NMF→ 이미지 데이터에서 활용, 텍스트 문서에서 활용

+추가자료

차원의 저주: 데이터의 차원이 늘어나면 데이터 분석이나 학습이 점점 어려워지는 현상을 말함.

이를 위한 접근방법: 투영(모든 훈련샘플을 부분 공간에 수직으로 투영3D→2D or 스위스 롤의 경우 롤을 펼쳐서 투영해야 함.) 매니폴드 학습(훈련 샘플이 놓여있는 매니폴드를 모델링)

#02 PCA

- 여러 변수 간 에 존재하는 상관관계를 이용해 이를 대표하는 주성분(Principal Component)을 추출해 차원을 축소하는 기법
- 가장 높은 분산을 가지는 데이터의 축을 기준으로 축소

1. 가장 큰 데이터 변동성(Variance)을 기반으로 첫 번째 벡터 축을 생성
2. 두 번째 축은 이 벡터 축에 직각이 되는 벡터(직교 벡터)를 축으로 함.
3. 세 번째 축은 다시 두 번째 축과 직각 이 되는 벡터를 설정하는 방식으로 축을 생성.
4. 이렇게 생성된 벡터 축에 원본 데이터를 투영하면 벡터 축의 개수만큼의 차원으로 원본 데이터가 차원 축소

- PCA 차원 축소 방법(선형대수 관점)
  입력 데이터의 공분산 행렬(개별 분산값을 대각원소로 하는 대칭행렬)을 고유값 분해하고, 이렇게 구한 고유벡터에 입력 데이터를 선형 변환하는 것
  고유벡터: PCA의 주성분 벡터로서 입력 데이터의 분산이 큰 방향을 나타냄
  고윳값(eigenvalue): 바로 이 고유벡터의 크기를 나타내며, 동시에 입력 데이터의 분산을 나타냄
  <선형변환>

: 특정 벡터에 행렬 A를 곱해 새로운 벡터로 변환하는 것
특정 벡터를 하나의 공간에서 다른 공간으로 투영하는 개념
<고유벡터>
: 고유벡터는 행렬 A를 곱하더라도 방향이 변하지 않고 그 크기만 변하는 벡터를 지칭
Ax = ax(A는 행렬, x는 고유벡터, a는 스칼라값)
고유벡터는 여러 개가 존재하며, 정방 행렬은 최대 그 차원 수만큼의 고유벡터를 가질 수 있음

입력 데이터의 공분산 행렬이 고유벡터와 고유값으로 분해될 수 있으며,
이렇게 분해된 고유벡터를 이용해 입력 데이터를 선형 변환하는 방식이 PCA다

- 붓꽃 실습 포인트

-사이킷런은 PCA 변환을 위해 PCA 클래스를 제공. PCA 클래스는 생성 파라미터로 n_components(PCA로 변환 할 차원의 수)를 입력받음

-explained_variance_ratio_ 속성은 전체 변동성에서 개별 PCA 컴포넌트별로 차지하는 변동성 비율을 제공

-4개의 속성이 2개로, 속성 개수가 50% 감소한 것을 고려한다면 PCA 변환 후에도 원본 데이터의 특성을 상당 부분 유지하고 있음

- 신용카드 실습 포인트

-높은 상관도를 가진 속성들은 소수의 PCA만으로 이 속성들의 변동성을 수용할 수 있음

-전체 23개 속성의 약 1/4 수준인 6개의 PCA 컴포넌트만으로도 원본 데이터를 기반으로 한 분류 예측 결과보다 약 1~2%정도의 예측 성능 저하만 발생 →PCA의 뛰어난 압축 능력

+추가자료: 차원을 축소하지 않고 PCA계산한 뒤훈련 세트의 분산을 95 %로 유지하는 데 필요한 최소한의 차원 수 계산

#03 LDA(선형 판별 분석법)

- PCA와 유사점: 입력 데이터 세트를 저차원 공간에 투영해 차원을 축소하는 기법임
- PCA와 차이점: LDA는 지도학습의 분류(Classification)에서 사용하기 쉽도록 개별 클래스를 분별할 수 있는 기준을 최대한 유지하면서 차원을 축소.
  PCA는 입력 데이터의 변동성의 가장 큰 축을 찾았지만, LDA는 입력 데이터의 결정 값 클래스를 최대한으로 분리할 수 있는 축을 찾음
- 방법: 클래스 내부와 클래스 간 분산 행렬을 구하고, 행렬을 고유벡터로 분해. 고유값이 가장 큰 순으로 차수만큼 추출, 새롭게 입력데이터 변환
- 붓꽃 실습 포인트:

-LDA는 PCA와 다르게 비지도학습이 아닌 지도학습임.즉, 클래스의 결정 값이 변환 시에 필요

#04 SVD

PCA와 유사한 행렬 분해 기법 이용

- 정방행렬만 고유벡터로 분해할 수 있는 PCA와 달리 SVD는 정방행렬뿐만 아니라 m*n 크기의 행렬 특이값 분해
- Σ의 대각원소 중 상위 몇 개만 추출해 대응하는 U와 V의 원소도 함께 제거해 더욱 차원 줄인 형태로 분해하는 것
- svd 예제

-랜덤한 4*4 행렬 생성 (행렬의 개별 로우끼리의 의존성 없애기 위해 랜덤 생성)

-Sigma 행렬은 다시 0을 포함한 대칭행렬로 변환한 뒤 내적 수행해야 함 주의!

-Sigma의 0에 대응되는 U, Sigma, Vt의 데이터 제외하고 복원

Sigma: 앞의 2개 요소만 0 아니므로 U 행렬 중 선형 두 개의 열만 추출

Vt: 선행 두 개의 행만 추출

-희소행렬로만 지원돼서 scipy.linalg.svd 말고 scipy.sparse.linalg.svds 이용해야 함

-사이킷런 TruncatedSVD 클래스:

TruncatedSVD 변환 역시 PCA와 유사하게 변환 후에 품종별로
클러스터링 가능할 정도로 각 변환 속성으로 뛰어난 고유성 가지고 있음

#05 NMF

- Truncated SVD와 같이 낮은 랭크 통한 행렬 근사(Low-Rank Approximation) 방식의 변형
- "하나의 객체정보를 음수를 포함하지 않은 두 개의 부분 정보로 인수분해하는 방법"
  목적 - 공통 특성만을 갖고 정보 줄이는 것
- 증배 갱신 규칙 (Multiplicative Update Rules)

1. 네 개의 갱신 행렬 생성
2. 행렬 W 갱신하기 위해 행렬 W의 모든 값을 식 (a) 내의
   대응하는 값과 곱하고 식 (b) 내의 대응하는 값으로 나눔
3. 행렬 H를 갱신하기 위해 행렬 H 내의 모든 값을 식 (c) 내의
   대응하는 값과 곱하고 식 (d) 내의 대응하는 값으로 나눔

4. 위 과정을 행렬 R과 행렬 W*H의 차이가 0이 될 때까지 반복

- 사용 영역:

-이미지 압축 통한 패턴 인식

-텍스트의 토픽 모델링 기법

-문사 유사도 및 클러스터링

-추천 영역

#06 추가: 다른 차원축소 기법들

- 랜덤 투영: 랜덤한 선형 투영을 사용해 고차원 데이터를 저차원으로 변환 sklearn.random_projection
- 다차원 스케일링 (MDS, Multidimensional Scaling):샘플 간의 거리(distance) 를 최대한 보존하면서 차원
- Isomap:각 샘플을 가까운 이웃과 연결해 그래프를 만들고, 그 그래프상의 지오데식 거리(geodesic distance) 를 유지하면서 차원 축소
- t-SNE:비슷한 샘플은 가깝게, 다른 샘플은 멀리 배치되도록 확률적 방식으로 임베딩

#07 추가 캐글 노트북: Dimensionality Reduction for Beginners

- 활용되는 차원 축소 기법

1. PCA (주성분 분석)
2. MDS (다차원 척도법)
3. t-SNE

- 전처리: 라벨 인코딩(원핫으로) 스탠다드 스케일링으로 단위 통일.
- 기법 1: PCA (주성분 분석)
  분석 목적: 유방암 데이터의 30개 Feature를 2개의 '주성분'으로 축소함. 이후 히트맵으로 주성분이 의미하는 바와 관계 파악.
- 기법 2: MDS (Multi-Dimensional Scaling)
  분석 목적: 유방암 데이터의 30개 Feature를 2개의 'MDS 차원'으로 축소.
  이유: 고차원 데이터 내에 숨어있는 저차원 구조(Manifold)를 찾아 2차원으로 시각화.
  → 거리 보존이 핵심
- 기법 3: t-SNE
  분석 목적: 유방암 데이터의 30개 Feature를 2개의 't-SNE 차원'으로 축소.

이유: 군집(클러스터)을 찾는 데 최적화된 기법. →2차원(2D)에서도 '가까운 이웃'이 되도록 뭉치게 만듦.MDS(전체적 '거리' 보존)와 달리, t-SNE는 '지역적 구조(Local structure)' 보존에 초점.

0차원의 유방암 데이터는 세 가지 기법 모두 2차원 시각화를 통해 성공적으로 그룹 분리가 가능했는데, 목적에 따라 가장 적합한 기법이 다르다.

- PCA: 가장 빠르고, 데이터의 전반적인 구조와 주요 성분을 확인할 때
- MDS: 샘플 간의 상대적인 거리 관계가 중요할 때
- t-SNE: 데이터의 숨겨진 군집을 시각적으로 명확하게 확인하고 싶을 때