6장 차원축소

01 차원축소 개요

차원축소: 매우 많은 피처로 구성된 다차원 데이터 세트의 차원을 축소해 새로운 차원의 데이터 세트를 생성하는 것
차원이 증가할수록 데이터 포인트 간의 거리가 기하급수적으로 멀어지고 sparse한 구조를 가짐, 예측 신뢰도가 떨어지고 상관관계가 높을 가능성이 커 다중 공선성 문제의 문제
차원 축소는 피처 선택과 피처 추출로 나눌 수 있음
피처 선택: 특정 피처에 종속성이 강한 불필요한 피처는 아예 제거하고 데이터의 특징을 잘 나타내는 주요 피처만 선택하는 것
피처 추출: 기존 피처를 저차원의 중요 피처로 압축해서 추출하는 것, 기존 피처를 단순 압축이 아닌 피처를 함축적으로 더 잘 설명할 수 있는 또 다른 공간으로 매핑해 추출하는 것
함축적인 특성 추출은 기존 피처가 전혀 인지하기 어려웠던 잠재적 요소를 추출하는 것 의미
차원 축소 알고리즘은 매우 많은 픽셀로 이뤄진 이미지 데이터에서 잠재된 특성을 피처로 도출해 함축적 형태의 이미지 변환과 압축을 수행
이미지 분류 등의 분류 수행시에 과적합 영향력이 작아져 오히려 원본 데이터로 예측하는 것보다 예측 성능 더 끌어 올릴 수 있음, 문서 내 단어들 구성에서 숨겨져 있는 시맨틱 의미나 토픽을 잠재 요소로 간주하고 이를 찾아낼 수 있음(SVD와 NMF는 이러한 시맨틱 토픽 모델링 위한 기반 알고리즘)

02 PCA(principal component analysis)

PCA 개요

PCA: 여러 변수 간에 존재하는 상관관계를 이용해 이를 대표하는 주성분을 추출해 차원을 축소하는 기법
가장 높은 분산을 가지는 데이터의 축을 찾아 이 축으로 차원을 축소하는데, PCA의 주성분이 됨 (즉 분산이 데이터의 특성을 가장 잘 나타내는 것을 간주)

제일 먼저 가장 큰 데이터 변동성을 기반으로 첫 번째 벡터 축 생성하고 두번째 축은 이 벡터 축에 직각이 되는 벡터를 축으로 함, 세번째 축은 다시 두 번째 축과 직각이 되는 벡터를 설정하는 방식으로 축을 생성
선형대수 관점- 입력 데이터의 공분산 행렬을 고유값 분해하고 고유벡터에 입력 데이터를 선형 변

환하는 것, 고유벡터가 PCA의 주성분 벡터로서 입력 데이터의 분산이 큰 방향 나타냄, 고유값은 이 고유벡터의 크기를 나타내며 동시에 입력 데이터의 분산을 나타냄

고유벡터는 행렬A를 곱하더라도 방향이 변하지 않고 크기만 변하는 벡터 지칭
고유벡터는 행렬이 작용하는 힘의 방향과 관계가 있어서 행렬을 분해하는데 사용
대칭행렬은 항상 고유벡터를 직교행렬로, 고유값을 정방행렬로 대각화할 수 있음

PCA

1. 입력 데이터 세트의 공분산 행렬 생성
2. 공분산 행렬의 고유벡터와 고유값 계산
3. 고유값이 가장 큰 순으로 k개 만큼 고유벡터 추출
4. 고유값이 가장 큰 순으로 추출된 고유벡터를 이용해 새롭게 입력 데이터를 변환

여러속성을 PCA로 압축하기 전 각 속성값을 동일한 스케일로 변환하는 것 필요

03 LDA(linear discriminant analysis)

LDA 개요

LDA는 지도학습의 분류에서 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 차원을 축소
클래스 간 분산은 최대한 크게 가져가고 클래스 내부의 분산은 최대한 작게 가져가는 방식
PCA와 유사하나 가장 큰 차이점은 공분산 행렬이 아니라 위에 설명한 클래스 간 분산과 클래스 내부 분산 행렬을 생성한 뒤 이 행렬에 기반해 고유벡터를 구하고 입력 데이터를 투영한다는 점

LDA

1. 클래스 내부와 클래스 간 분산 행렬을 구합니다. 이 두개의 행렬은 입력 데이터의 결정 값 클래스 별로 개별 피처의 평균 벡터를 기반으로 구합니다.
2. 클래스 내부 분산 행렬을 Sw, 클래스 간 분산 행렬을 Sb 라고 하면 두 행렬을 고유벡터로 분해.
3. 고유값이 가장 큰 순으로 k개 추출
4. 고유값이 가장 큰 순으로 추출된 고유벡터를 이용해 새롭게 입력 데이터를 변환

04 SVD(singular value decomposition)

SVD 개요

PCA와 유사한 행렬 분해 기법 이용
PCA는 정방행렬만을 고유벡터로 분해 할 수 있지만 SVD는 정방행렬 뿐만 아니라 행과 열의 크기가 다른 행렬에도 적용 가능

truncated SVD는 ∑의 대각원소 중 상위 몇 개만 추출해서 여기에 대응하는 U와 V의 원소도 함께 제거해 더욱 차원을 줄인 형태로 분해하는 것

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 양수라는 게 보장되면 다음과 같이 좀 더 간단하게 두 개의 기반 양수 행렬로 분해될 수 있는 기법을 지칭

행렬분해는 일반적으로 SVD와 같은 행렬 분해 기법을 통칭