파이썬 머신러닝 완벽 가이드 **6**장 차원축소

- 차원축소?
  - 매우 많은 피처로 구성된 다차원 세트의 차원을 축소해 새로운 차원의 데이터 세트를 생성
- 다차원 데이터 세트의 문제점
  - 차원이 증가할수록 데이터 포인트 간의 거리가 기하급수적으로 멀어지고, 희소한Sparse 구조를 가져 예측 신뢰도가 떨어진다.
  - 다중공선성 문제(독립변수 간의 상관관계가 높은 것)로 예측 성능 저하
    - 회귀분석의 전제 가정 위배 : 독립변수간 상관관계는 높으면 안된다
- 차원 축소의 분류
  - 피처(특성) 선택 : 특정 피처에 종속성이 강한 불필요 피처는 아예 제거 + 데이터 특징 잘 나타내는 주요 피처만 선택
  - 피처(특성) 추출 : 기존 피처를 저차원의 중요 피처로 압축하여 추출 ⇒ 기존 피처와 완전히 다른 새로운 값이 됨
    - 단순 압축이 아닌, 피처를 합축적으로 더 잘 설명할 수 있는 또 다른 공간으로 매칭하여 추출
- 차원 축소의 활용
  - 이미지 데이터에서 잠재된 특성을 피처로 도출해 함축적 형태의 이미지 변환과 압축 수행
  - 텍스트 문서의 숨겨진 의미 추출. 문서 내 단어들의 구성에서 숨겨져 있는 시맨틱Semantic 의미나 토픽topic을 잠재 요소로 간주하고 이를 찾아낸다.

- PCA
  - 변수 간 상관관계를 이용해 이를 대표하는 주성분을 추출해 차원을 축소하는 방법
  - PCA의 주성분: 정보 유실을 최소화하기 위해 가장 높은 분산을 가지는 데이터를 찾아, 이 축으로 차원을 축소한다. 즉, 분산이 데이터의 특성을 가장 잘 나타내는 것으로 간주
  - 공분산/공분산행렬/정방행렬/대칭행렬
  - 주성분
  - 공분산
    - 공분산 행렬은 항상 고유벡터를 직교행렬로, 고유값을 정방 행렬로 대각화 할 수 있음
  - 고유값
  - 고유벡터
  - 선형변환
    - 특정 벡터에 행렬을 곱해 새로운 벡터로 변환하는 것

- LDA
    - PCA와 유사하지만, 지도 학습의 분류에서 사용하기 쉽도록 개별 클래스를 분별할 수 있는 기준을 최대한 유지하면서 차원 축소
    - PCA는 입력 데이터의 변동성의 가장 큰 축을 찾았지만, LDA는 입력 데이터의 결정 값 클래스를 초대한 분리할 수 있는 축을 찾음

    - 수행
        - 입력 데이터의 결정값 클래스 별로 개별 피처의 평균 벡터를 기반으로 글래스 내부,클래스 간 분산 행렬을 구한다
        - 클래스 내부 분산 행렬, 클래스 간 분산행렬 두 행렬을 고유벡터로 분해한다
        - 고유값이 가장 큰 순으로 k개 만큼 추출한다
        - 고유값이 가장 큰 순으로 추출된 고유 벡터를 이용해 새롭게 입력 데이터를 변환
-

- SVD
  - PCA와 유사. 정방 행렬뿐만 아니라 행과 열의 크기가 다른 행렬에도 적용 가능
  - Full SVD
  - compact svd
    - 일반적
  - truncated svd
    - 사이파이에서만 지원됨
    - 검증 수행 순서
      - 임의의 원본행렬을 nomal svd로 분해
      - 다시 truncated SVD로 분해
      - turncated SVD로 분해된 행렬의 내적 계산
  - 다시 행렬 복원하기
    - U∗∑∗V^T=A
  - SVD는 텍스트의 토픽 모델링 기법인 LSA기반 알고리즘
- NMF
  - 원본 행렬 내의 모든 원소값이 모두 양수(0 이상)라는 게 보장되면, 두 개의 기반 양수 행렬로 분해될 수 있는 기법
  - Truncated SVD와 같이 낮은 랭크를 통한 행렬 근사 방식의 변형
  - W×H≈V