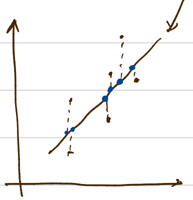
CH06 차원축소

01. 차원축소 dimention reduction

- 많은피처: 다차원 → 축소 → 하나의차원으로
  - sparse : data간 거리↑
  - 피처간 상관관계↑ : 다중공선성문제
- 차원축소 = 피처수↓
- 직관적(시각화), 처리능력↑
- 피처선택 selection : 제거
- 피처추출 extration : (압축) → 완전히 새로운 값 — 잠재변수고려한, Latent Factor
- 이미지, 텍스트 처리에 Good.
- PCA, SVD, NMF
  - ↳ Semantic Topic modeling

02. PCA : Principal Component Analysis

- 피처간 상관관계 → 주성분 principal components (축)
  - → 축소(투영)



* 변동성이 가장 큰 축
  - → 직교벡터
  - → 직교벡터
  - ⋮ (피처수만큼 추가)
  - 이후 주성분으로 투영

(선형대수) 주성분 = 고유벡터

$$C = P\Sigma P^T$$

입력data → 고유벡터 + 고유값 (고유벡터의크기)

이걸로 입력data를 선형 변환

1. input → 공분산행렬
2. 고유벡터, 고유값 계산
3. ↳ 큰 순 추출
4. input 변환

- PCA. 여러 속성 계산, 스케일링 필요
- PCA(n_component=) → 변환할 피처 수 (n개 → 2개 피처로나옴, 2명)
  - ↳ fit, transform, fit_transform (input_data)
- PCA.explained_variance_ratio_
  - : 전체 변동성에서 컴포넌트별 변동성 비율

03 LDA : Linear Discriminant Analysis

: 선형 판별 분석법

- 분류 특화 → 분별기준 최대한 유지
  - PCA : 최대 변동성 가지는 축
  - LDA : 최대한 분리可 하게하는 축
  - ↑ 클래스간 분산↑ & 클래스내부 분산↓
- 분산행렬 → 고유 vector
  - PCA : 공분산행렬 → 고유 vector
- sklearn.discriminant_analysis.LinearDiscriminantAnalysis

04 SVD: Singular Value Decomposition

- PCA : n×n 행렬 → 고유vector
- SVD : m×n → 고유vector

$$A = U\Sigma V^T$$

(m×m, m×n, n×n)

U, V는 특이벡터, 서로 직교

Σ는 대각행렬.

A의 특이값 = Σ의 0이아닌값

- Truncated SVD : Σ의 대각원소 中, 상위 일부data 사용
  - 원복어려움. A 내적으로 대충 복원 可
- scipy.sparse.linalg.svds
  - scipy.linalg.svd : 희소행렬만지원
- 비전, 이미지 압축에 사용

05 NMF: Non-negative Matrix Factorization

- ≒ SVD. Low-Rank Approximation, 행렬 분해
- 원본 행렬의 모든 원소 >0 이면, 간단한 행렬 곱셈 가능
- 이미지 압축, 클러스터링, 추천에 사용

06

- 피처의 개수를 줄여 데이터를 잘 설명할 수 있는
  잠재적인 요소를 추출하는것
- PCA: 변동성 → 축 → 투영
  고유벡터 → 선형변환
- LDA: 분리가능성 → 축
- SVD/NMF: 고차원행렬 → 행렬분해 → 저차원행렬
  잠재요소추출