Week 6

Sunday, August 10, 2025 10:52 PM

1. 차원축소

- 변수선택 : 목적에 부합하는 소수변수 선택 (해석용이) ↕ (상관관계 고려)
- 변수 추출 : 변환으로 새로운 변수 추출

Supervised Feature Selection : IG, Stepwise, LLA, ...

Feature Extraction : Partial Least Square

Unsupervised Feature Selection : PCA loading

Feature Extraction : PCA, Wavelets, AutoEncoder.

2. PCA

: n개 관측치와 p개 변수로 구성된 데이터 → 직교하는 k개 변수로 구성된 데이터로 요약.

기존변수의 선형조합

원래 데이터 분산을 최대한 보존하는 축으로 데이터 사영.

→ 차원축소 · 시각화 · 해석 ~ 초기 단계 사용.

$$Z_p = \alpha_p^T X = \alpha_{p1}X_1 + \alpha_{p2}X_2 + \cdots + \alpha_{pp}X_p$$

- X : 원래 변수
- α_j : j번째 Loading
- Z_p : 사영 변환 후 변수.

3. 수학적 배경

공분산 Covariance $\quad Cov(X) = \frac{1}{n}(X-\bar{X})(X-\bar{X})^T$

- 공분산 행렬 대각성분 = 각 변수의 분산 & 비대각 = 대응 두변수의 cov
- 데이터의 총분산은 공분산행렬 대각성분의 합 $\quad tr[Cov(X)] = Cov_{11} + Cov_{22} + \cdots + Cov_{pp}$

사영 Projection

$$\vec{x} = p\vec{a} \qquad p = \frac{\vec{b}^T\vec{a}}{\vec{a}^T\vec{a}}$$

고유값 EigenValue & 고유벡터 EigenVector

$Ax = \lambda x$ 일때,
- λ : 고유값
- x : 고유벡터 ~ 선형변환에 의해 방향 변하지 않는 벡터.

$(A - \lambda I)x = 0$

4. Algorithm ($\bar{x}_i = 0$ 인 데이터 가정)

: to find α that produces the largest variance of Z → 구하는 새로운 PC

$= \alpha^T X$ → 원변수의 선형결합

$$\mathrm{Max}\ Var(Z) = Var(\alpha^T X)$$
$$= \alpha^T Var(X)\alpha = \alpha^T \underbrace{\Sigma}_{\text{Cov. Matrix}} \alpha$$
$$= \alpha^T \underbrace{E}_{\text{EigenVec}} \underbrace{\Lambda}_{\text{Eigenvalue Matrix}} E^T \alpha$$

$\downarrow \ \beta = E^T\alpha \quad \|\beta\| = 1$

$$\mathrm{Max}\ \beta^T\Lambda\beta = \mathrm{Max}\ \lambda_1\beta_1^2 + \lambda_2\beta_2^2 + \cdots + \lambda_m\beta_m^2$$

$$s.t.\ \beta_1^2 + \beta_2^2 + \cdots + \beta_m^2 = 1, \quad \lambda_1 > \lambda_2 > \cdots > \lambda_m.$$

$\Rightarrow \beta_1 = 1$ 이고 나머지 0일때 Max, optimal value $= \lambda_1$, $\alpha = e_1$

$[1\ 0\ 0 \cdots 0]$

① 데이터 정규화 (원하는 경우 corr matrix)

② 각 변수 cov matrix 계산.

③ Cov matrix → eigenval & vector 계산

④ EigenVal 및 해당 EigenVector 순서대로 나열.

⑤ 정렬된 EigenVector 기반 각 변수 변환.

=> $\beta_1 = 1$ 이고 나머지 0일때 Max, optimal value = λ, $\alpha = e_1$

$[1 \ 0 \ 0 \cdots 0]$

5. PCA

공분산행렬의 EigenValue = 각 주성분의 분산.

ex) $Cov(Z) = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ $\quad Var(Z_1) = 2 = \lambda_3$, $Var(Z_2) = 1 = \lambda_2$, $Var(Z_3) = 1 = \lambda_1$

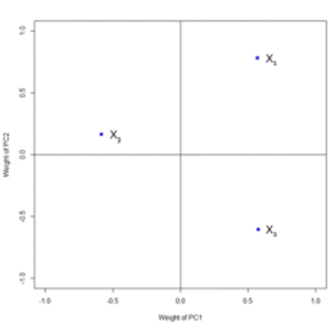
첫번째 주성분의 분산 비율 $= \dfrac{\lambda_3}{\lambda_1 + \lambda_2 + \lambda_3} = 0.5$

→ 원래 데이터의 50% 설명.

주성분 개수 선택

Elbow point / 일정 비율이상 분산비보존.

6. PCA Loading Plot.



: 주성분 만드는데 있어 기존변수가 어떻게 기여했는가.

X_1 : PC1 PC2 사영값 커짐. 중요변수

X_2 : 증가시 PC1 증가 PC2 감소

X_3 : 증가시 PC1 감소 PC2 증가

7. 한계

∘ Non Gaussian / 다중 Gaussian 자료시 적용 어려움.

→ 커널 PCA, LLE (Locally Linear Embedding)

∘ 범주 정보 고려 X, 범주간 구별이 잘되도록 변환는 아님.