1. 주성분 분석(PCA)을 이용한 데이터 표현과 특징 추출

1. PCA는 무엇인가?

- PCA (Pricinpal Component Analysis) : 차원 축소 기법
    - 차원 축소 : 데이터 복잡성을 줄이고, 계산 효율성을 향상.
    - 주성분 추출 : 데이터의 변동성을 가장 많이 설명하는 방향 벡터들.
    - 데이터 시각화 : 복잡한 데이터를 2D, 3D로 축소하여 시각화 가능.
    - 특징 선택 : 가장 중요한 특징이 무엇인지를 결정하는 데 사용됨.
    - 노이즈 제거 : 노이즈는 주성분의 작은 변화로 표시되어 제거됨.

2. 결합 확률과 공분산

- 확률 개념을 알아야 PCA 이해가 가능.
- **표본공간** : 통계실험에서 발생가능한 모든 결과를 모은 공간.
- **결합분포** : 표본공간에서 발생 가능한 무수히 많은 확률 변수들 중 **여러 개의 확률 변수를 동시에 다루는 것**을 말함.
    - 확률변수의 표기 : $X, Y, Z...$ 또는 $X_1, X_2, X_3, ...$
- **확률 벡터** : 확률변수들의 순서쌍.
- **기댓값** : 각 사건이 벌어졌을 때의 **이득 값과 그 사건이 벌어질 확률을 곱한 것**을 **전체** 사건에 대해 합한 값.

  X가 확률분포 $f(x)$를 가지는 확률변수라 하자.

  X가 이산형인 경우 X의 평균 또는 기댓값은

  $$\mu = E[X] = \Sigma_x x f(x) = \Sigma_x x p(X = x)$$

X가 연속인 경우 X의 평균 또는 기댓값은

$$\mu = E[X] = \int_{-\infty}^{\infty} xf(x)dx$$

- **분산** : 확률 변수 X의 분포가 평균을 중심으로 밀집된 정도.

  X가 확률분포 $f(x)$를 가지는 확률변수라 하자. X의 기댓값 $\mu = E[X]$에 대하여

  X가 이산형인 경우 X의 분산은

  $$E[(x-\mu)^2] = \Sigma_x (x-\mu)^2 f(x)$$

  X가 연속인 경우 X의 분산은

  $$E[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx$$

  - 분산의 변형

    $$\sigma^2 = E[(X-\mu)^2] = E[X^2] - \mu^2$$

- **표준편차** : 분산을 보기 편한 단위로 변환한 척도로, 분산의 제곱근으로 계산.

  $\sqrt{E[(x-\mu)^2]}$

- **공분산** : **2개의 확률변수의 선형 관계**를 나타낸다.

  두 확률변수 X, Y에 대하여 $E[X] = \mu_X, E[Y] = \mu_Y$라 하자. 이 때

  $$E[(X-\mu_X)(Y-\mu_Y)] = E[XY] - \mu_X \mu_Y$$

  를 두 확률변수 X, Y의 공분산이라 부르고, $Cov(X, Y)$로 표시한다.

- **상관계수** : 두 변수 사이의 통계적 관계를 표현하기 위한 수치.

$$\begin{aligned}\rho_{XY} &= \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \\ &= \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sqrt{E[(X-\mu_X)^2]}\sqrt{E[(Y-\mu_Y)^2]}} \\ &= \frac{E[XY]-\mu_X\mu_Y}{\sqrt{E[X^2]-\mu_X^2}\sqrt{E[Y^2]-\mu_Y^2}}\end{aligned}$$

- 공분산 행렬 : 확률변수 $X_1, ...X_n$에 대하여 다음의 n x n 행렬을 공분산 행렬이라고 한다.

$$\Sigma = \begin{pmatrix} Var(X_1) & Cov(X_1, X_2) \cdots & Cov(X_1, X_n) \\ Cov(X_2, X_1) & Var(X_2) \quad \cdots & Cov(X_2, X_n) \\ \vdots & \vdots \quad\quad \ddots & \vdots \\ Cov(X_n, X_1) & Cov(X_3, X_2) \cdots & Var(X_n) \end{pmatrix}$$

3. 주성분분석 (PCA)

- 차원 축소 : 원본 데이터의 특성을 보존하면서 차원을 줄이는 방법.
  - 두 데이터 사이의 관계를 설명하는 정사영을 찾는 방법 : 분산이 큰 직선을 선택한다.
- 샘플에서의 평균, 분산 계산
  - 샘플 평균 : $\frac{1}{n}\Sigma_{i=1}^{n}x_i$
  - 샘플 분산 : $\frac{1}{n-1}\Sigma_{i=1}^{n}(x_i - \mu)^2$
  - 샘플 공분산 : $\frac{1}{n-1}\Sigma_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$

4. PCA응용 : Eigenface와 영상인식응용