Kernel PCA

훈련 집합 $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^n \in \mathbb{R}^d$와 변환 $\phi : \mathbb{R}^d \to \mathbb{R}^m$에 대하여

$$\phi(X) = \begin{bmatrix} \phi(\boldsymbol{x}^1)^T \\ \vdots \\ \phi(\boldsymbol{x}^n)^T \end{bmatrix}$$

의 주성분 분해에 대하여 알아보자. 식을 단순화하기 위하여, 우선 $\phi(X)$ 대신 일반적인 행렬을 사용하여 계산한다. 앞으로 등장할 행렬 A는 $\phi(X)$라 생각하면 된다.

커널 (또는 커널 함수) k는

$$k(\boldsymbol{x}, \boldsymbol{y}) = \phi(\boldsymbol{x}) \cdot \phi(\boldsymbol{y})$$

로 정의된다.

1. 공분산 행렬과 커널 행렬

$A = (a_{ij})$를 $n \times m$ 행렬이라 하고 A의 i 행을 전치한 행렬(벡터)를 A^i라 하자.

$$A^i = \begin{bmatrix} a_{i1} \\ \vdots \\ a_{im} \end{bmatrix}, \quad A = \begin{bmatrix} A^{1T} \\ \vdots \\ A^{nT} \end{bmatrix}$$

또, 두 행렬 C와 K를 각각

$$C = A^T A, \quad K = AA^T$$

라 하자. $\frac{1}{n}C$와 K는 각각 **공분산 행렬**(covariance matrix)과 **커널 행렬**(kernel matrix)이라 한다. 행렬 C는 다음과 같이 표현할 수도 있다.

$$C = \sum_{i=1}^{n} A^i A^{iT} \tag{1}$$

C의 고유 값을 $\lambda_1, \ldots, \lambda_m$, 대응되는 고유 벡터를 $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m \in \mathbb{R}^m$이라 하자. C가 대칭 행렬이므로 고유 벡터들은 정규 직교 기저를 이룬다.

따라서 A^i는 고유 벡터들의 일차 결합으로 표현된다. $A^i = \sum_{k=1}^{m} \alpha_k^i \boldsymbol{v}_k$라 하고 α_k^i를 구하면

$$\alpha_k^i = A^i \cdot \boldsymbol{v}_k = A^{iT} \boldsymbol{v}_k \tag{2}$$

를 얻는다. 여기서

$$\boldsymbol{\alpha}_k = \begin{bmatrix} \alpha_k^1 \\ \vdots \\ \alpha_k^n \end{bmatrix} = \begin{bmatrix} A^{1T} \boldsymbol{v}_k \\ \vdots \\ A^{nT} \boldsymbol{v}_k \end{bmatrix} \in \mathbb{R}^n$$

라 하면

$$\boldsymbol{\alpha}_k = A \boldsymbol{v}_k \tag{3}$$

임을 알 수 있다.

고유 방정식 $C\boldsymbol{v}_k = \lambda_k \boldsymbol{v}_k$에 식 (1)을 대입하고 식 (2)를 적용하면

$$\lambda_k \boldsymbol{v}_k = \sum_{i=1}^{n} A^i A^{iT} \boldsymbol{v}_k = \sum_{i=1}^{n} \alpha_k^i A^i$$

와 같다. 위 식의 오른쪽 항은 $A^T \boldsymbol{\alpha}_k$와 같으므로

$$\lambda_k \boldsymbol{v}_k = A^T \boldsymbol{\alpha}_k \tag{4}$$

를 얻는다.

◇ 식 (2) 대신 식 (3)을 $\boldsymbol{\alpha}_k$의 정의로 사용해도 동일한 결과를 얻는다.

◇ 두 행렬 C와 K가 서로 쌍대(dual) 형식이듯 두 식 (3)과 (4)도 서로 쌍대 형식이다. A와 A^T의 역할을 바꾸면 C와 K의 역할이 바뀌고 식 (3)과 (4)의 역할이 바뀐다.

◇ $\boldsymbol{\alpha}_k$의 정의를 $c_k\boldsymbol{\alpha}_k = A\boldsymbol{v}_k$, $c_k \in \mathbb{R}$로 바꾸면 식 (4)는 $\lambda_k / c_k \boldsymbol{v}_k = A^T\boldsymbol{\alpha}_k$로 바뀐다. 고유값 λ_k를 식 (3)과 (4)가 나누어 가질 수 있다.

커널 행렬 K에 $\boldsymbol{\alpha}_k$를 곱하여 식 (3)을 적용하면

$$K\boldsymbol{\alpha}_k = AA^TA\boldsymbol{v}_k = AC\boldsymbol{v}_k = \lambda_k A\boldsymbol{v}_k = \lambda_k\boldsymbol{\alpha}_k$$

를 얻는다. 따라서 λ_k와 $\boldsymbol{\alpha}_k$는 각각 K의 고유값과 고유 벡터이다.

여기서 생각해 볼 점은 기저 $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$에 대한 A^i의 성분이 $\alpha_1^i, \ldots, \alpha_m^i$이라는 것, 다시 말하면, 커널 행렬의 고유 벡터의 성분이 $\phi(\boldsymbol{x}^i)$의 성분이라는 것이다.

우리의 목표는 PCA에 있으므로 $\lambda_1 \geq \cdots \geq \lambda_m$이라 가정하자. A^i 곧, $\phi(\boldsymbol{x}^i)$의 k 차원 주성분은 $(\alpha_1^i, \ldots, \alpha_k^i)$임을 알 수 있다.

2. 예측(prediction)

임의의 벡터 $\boldsymbol{x} \in \mathbb{R}^d$에 대하여 $\phi(\boldsymbol{x}) \in \mathbb{R}^m$을 정규 직교 기저 $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$의 일차 결합으로 표현하여 주성분을 구해보자. $\phi(\boldsymbol{x}) = \sum_{k=1}^{m} \gamma_k \boldsymbol{v}_k$라 하면 $\gamma_k = \phi(\boldsymbol{x}) \cdot \boldsymbol{v}_k = \phi(\boldsymbol{x})^T\boldsymbol{v}_k$이 성립한다. 식 (4)를 대입하면

$$\gamma_k = \phi(\boldsymbol{x})^T\boldsymbol{v}_k = \frac{1}{\lambda_k}\phi(\boldsymbol{x})^T\phi(X)^T\boldsymbol{\alpha}_k = \frac{1}{\lambda_k}(\phi(X)\phi(\boldsymbol{x})) \cdot \boldsymbol{\alpha}_k$$

를 얻는다. $\phi(X)\phi(\boldsymbol{x})$는

$$\phi(X)\phi(\boldsymbol{x}) = \begin{bmatrix} \phi(\boldsymbol{x}_1) \cdot \phi(\boldsymbol{x}) \\ \vdots \\ \phi(\boldsymbol{x}_n) \cdot \phi(\boldsymbol{x}) \end{bmatrix} = \begin{bmatrix} k(\boldsymbol{x}_1, \boldsymbol{x}) \\ \vdots \\ k(\boldsymbol{x}_n, \boldsymbol{x}) \end{bmatrix}$$

이다.

3. zero-centered 커널

행렬 A의 j 열의 평균을 μ_j라 하자. 곧, $\mu_j = \frac{1}{n}\sum_{i=1}^{n} a_{ij}$라 하자. 또, A의 각 열에서 그 열의 평균을 뺀 행렬을 $\widetilde{A}$라 하자.

$$\widetilde{A} = \begin{bmatrix} a_{11} - \mu_1 & \cdots & a_{1m} - \mu_m \\ \vdots & \ddots & \vdots \\ a_{n1} - \mu_1 & \cdots & a_{nm} - \mu_m \end{bmatrix}$$

커널 행렬 $\widetilde{K} = \widetilde{A}\widetilde{A}^T$를 K를 써서 표현해보자.

모든 성분이 $\frac{1}{n}$인 $n \times n$행렬을 J_n이라 하면 $\widetilde{A} = A - J_nA$을 확인할 수 있다. 따라서

$$\widetilde{K} = K - KJ_n - J_nK + J_nKJ_n$$

을 얻는다.

4. 커널

자주 사용되는 커널은

Polynomial 커널

$$k(\boldsymbol{x},\boldsymbol{y}) = (\boldsymbol{x} \cdot \boldsymbol{y} + c)^p$$

RBF(radial basis function) 커널

$$k(\boldsymbol{x},\boldsymbol{y}) = \exp(-\gamma \|\boldsymbol{x}-\boldsymbol{y}\|^2)$$

Hyperbolic tangent 커널

$$k(\boldsymbol{x},\boldsymbol{y}) = \tanh(\boldsymbol{x} \cdot \boldsymbol{y} + c)$$

등이 있다. 이들의 경우 feature map ϕ는 무엇인지 알아보자.

논의를 단순화하기 위하여

$$\boldsymbol{x} = (x_1, x_2),\ \ \boldsymbol{y} = (y_1, y_2)$$

라 하자.

Polynomial 커널에서 $p = 2$, $c = 0$이라 하고

$$\phi(\boldsymbol{x}) \cdot \phi(\boldsymbol{y}) = (\boldsymbol{x} \cdot \boldsymbol{y})^2$$

을 만족하는 feature map ϕ를 찾아보자.

$$\begin{aligned}(\boldsymbol{x} \cdot \boldsymbol{y})^2 &= (x_1 y_1 + x_2 y_2)^2 \\ &= x_1^2 y_1^2 + x_2^2 y_2^2 + 2x_1 x_2 y_1 y_2 \\ &= (x_1^2, \sqrt{2}\, x_1 x_2, x_2^2) \cdot (y_1^2, \sqrt{2}\, y_1 y_2, y_2^2)\end{aligned}$$

이므로

$$\phi(\boldsymbol{x}) = (x_1^2, \sqrt{2}\, x_1 x_2, x_2^2)$$

이라 하면 된다. 일반적인 경우도 비슷하다.

RBF 커널에서 $\gamma = 1/2$이라 하고 ϕ를 찾아보자.

$$\exp(-\frac{1}{2}\|\boldsymbol{x}-\boldsymbol{y}\|^2) = \exp(-\frac{1}{2}\|\boldsymbol{x}\|^2)\exp(-\frac{1}{2}\|\boldsymbol{y}\|^2)\exp(x_1 x_2 y_1 y_2)$$

$$\begin{aligned}\exp(x_1 x_2 y_1 y_2) &= 1 + x_1 x_2 y_1 y_2 + \frac{1}{2} x_1^2 x_2^2 y_1^2 y_2^2 + \cdots \\ &= (1, x_1 x_2, x_1^2 x_2^2, \ldots) \cdot (1, y_1 y_2, y_1^2 y_2^2, \ldots)\end{aligned}$$

이므로

$$\phi(x_1, x_2) = e^{-\frac{1}{2}(x_1^2 + x_2^2)}(1, x_1 x_2, x_1^2 x_2^2, \ldots)$$

임을 알 수 있다.