Kernel PCA : Mercer's Thm

Given $x_1, \dots, x_m \in \mathbb{R}^n$ i.i.d. random vectors

$$\mu = \mathbb{E}[x] \quad \text{and} \quad C = \mathbb{E}[(x-\mu)(x-\mu)^T]$$

Principal Components of $x \in \mathbb{R}^N$ are entries of

$$y = U^T(x-\mu) = \begin{bmatrix} - & u_1^T & - \\ - & u_2^T & - \\ & \vdots & \\ - & u_N^T & - \end{bmatrix} (x-\mu)$$

__View 1__ $\mathbb{E}[y] = 0$ and $\mathbb{E}[yy^T] = \Lambda$

__View 2__ Direction u_j maximizes $\mathrm{var}(y^{(j)})$ s.t. $u_j \perp \{u_1, u_2, \dots, u_{j-1}\}$

$$u_j = \underset{\substack{\|v_j\|=1 \\ v_j \perp \{u_1, \dots, u_{j-1}\}}}{\arg\max} \mathbb{E}\left[|v_j^T(x-\mu)|^2\right]$$

PCA can only "find" detect linear features in the data.

Nonlinear Component Analysis

To identify "nonlinear" structure in the data, we can "add" new variables:

$$x_1 = \begin{pmatrix} x_1^{(1)} \\ x_1^{(2)} \end{pmatrix}, \quad \ldots, \quad x_m = \begin{pmatrix} x_m^{(1)} \\ x_m^{(2)} \end{pmatrix}$$

$$\Downarrow \qquad\qquad\qquad \Downarrow$$

Add $\quad x_k^{(3)} = (x_k^{(1)})^2, \quad x_k^{(4)} = x_k^{(1)} x_k^{(2)}, \quad x_k^{(5)} = (x_k^{(2)})^2$

$\Rightarrow$ Can find $r^2 = (x^{(1)})^2 + (x^{(2)})^2$

Let $\varphi = \mathbb{R}^n \to \mathbb{R}^d$ be a dictionary of features that "lift" the data into a higher-dimensional space ($d \gg n$).

$$\varphi(\overset{\in \mathbb{R}^n}{x}) = [\varphi_1(x), \varphi_2(x), \ldots, \varphi_d(x)]^T$$

$$\mu = \sum_{j=1}^{m} \varphi(x_j), \qquad C = \frac{1}{m-1} \sum_{j=1}^{m} (\varphi(x_j) - \mu)(\varphi(x_j) - \mu)^T$$

We can run PCA in the new higher-dim feature space to detect nonlinear structure:

$$C = U \Lambda U^T \Rightarrow \Psi(x_j) = U^T \varphi(x_j)$$

Diagonalize $d \times d$ cov. matrix

principle components of mapped data in feature space

Kernel PCA

$$C = \frac{1}{m-1} \sum_{j=1}^{m} (\varphi(x_j) - \mu)(\varphi(x_j) - \mu)^T = \frac{1}{m-1} \underset{d \times m}{B} \underset{m \times d}{B^T}$$

To compute nonzero eigenpairs:

$$\lambda \neq 0 \qquad \underset{d \times d}{Cu = \lambda u} \quad \overset{u = \frac{1}{\sqrt{m-1}} Bv}{\Longleftrightarrow} \quad \underset{m \times m}{\frac{1}{m-1} B^T B v = \lambda v}$$

To compute the principal components:

$$u^T(\varphi(x_j) - \mu) = \frac{1}{\sqrt{m-1}} v^T B^T (\varphi(x_j) - \mu)$$

$$= \frac{1}{\sqrt{m-1}} v^T B^T B$$

The Kernel Matrix

To avoid working in the d-dimensional space

$$(B^T B)_{ij} = [\varphi(x_i) - \mu]^T [\varphi(x_j) - \mu]$$

$$= \sum_{k=1}^{d} (\varphi_k(x_j) - \mu)(\varphi_k(x_i) - \mu)$$

For simplicity, take $\mu = 0$ (not necessary):

$$(B^T B)_{ij} = \sum_{k=1}^{d} \varphi_k(x_j)\varphi_k(x_i)$$

Define the kernel $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$

$$K(x,y) = \sum_{k=1}^{d} \varphi_k(x)\varphi_k(y)$$

Instead of starting w/ $\varphi : \mathbb{R}^n \to \mathbb{R}^d$, we start with a self-adjoint semi-definite Hilbert-Schmidt kernel $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, continuous on $\mathbb{R}^n \times \mathbb{R}^n$. Then, we have Mercer's Thm:

$$K(x,y) = \sum_{k=1}^{d \to \mathbb{N} \cup \{\infty\}} \lambda_n u_n(x) u_n(y)$$

converges pointwise, absolutely & uniformly.

Mercer's theorem allows us to write

$$K(x_i, x_j) = \sum_{k=1}^{\sqrt{d} \to \mathbb{N} \cup \{\infty\}} \lambda_n u_n(x_i) u_n(x_j) = (B^T B)_{ji}$$

Dictionary $\varphi(x) = [\underbrace{\sqrt{\lambda_1} u_1(x)}_{\varphi_1(x)}, \underbrace{\sqrt{\lambda_2} u_2(x)}_{\varphi_2(x)}, \ldots, \underbrace{\sqrt{\lambda_d} u_d(x)}_{\varphi_d(x)}]$

<u>Kernel PCA</u> (mean $u = 0$)

Form $\overset{m \times m}{(K)_{ij}} = k(x_i, x_j) \qquad 1 \le i, j \le m$

Compute $\frac{1}{\sqrt{m-1}} K v_\ell = \lambda v_\ell \qquad \ell = 1, 2, 3, \ldots$

Project (ℓth principal component) $c_\ell = \frac{1}{\sqrt{m-1}} v_\ell^T K \qquad \ell = 1, 2, 3, \ldots$

$\Rightarrow$ Implicitly run PCA in ∞-dim. feature space via "kernel trick."

$\Rightarrow$ Kernel PCA "finds" nonlinear structures in $\mathbb{R}^N$ that are well approximated by kernel eigfuns associated w/largest eigenvalues

=> Different Kernels lead to different biases based on eigenvalue decay and "dominant" eigenspaces.