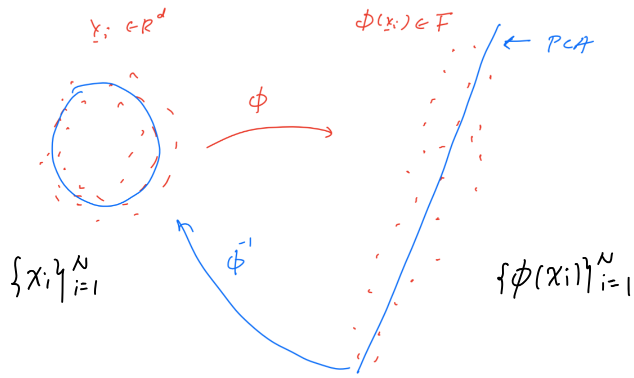
Kernel PCA – making PCA non-linear



PCA on $\{\phi(\mathbf{x}_i)\}_{i=1}^N$ in the feature space F:

- Suppose for the moment that the dataset is centered in F,

$$\sum_{i=0}^{N} \phi(\mathbf{x}_i) = 0.$$

- The sample covariance matrix is

$$C = \frac{1}{N}\sum_{i=1}^{N} \phi(\mathbf{x}_i)\phi(\mathbf{x}_i)^T.$$

- The principal components in F are the eigenvectors of C:

$$C\mathbf{v}_j = \lambda_j \mathbf{v}_j, \quad j = 1, \cdots, p$$

where the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots$.

Finding the eigenvectors of C in F: — | = scalar

- Using the definition of the covariant matrix, we have

$$\sum_{i=1}^{N} \underbrace{\left(\frac{a_{ji}}{N}\cdot\frac{1}{\lambda_j}\right)}_{a_{ji}} v_j = v_j \qquad \underbrace{\overbrace{\frac{1}{N}\sum_{i=1}^{N} \phi(\mathbf{x}_i)\phi(\mathbf{x}_i)^T}^{C}\mathbf{v}_j} = \lambda_j \mathbf{v}_j. \tag{1}$$

- $\mathbf{v}_j$ is a linear combination of the features $\{\phi(\mathbf{x}_i)\}_{i=1}^N$:

$$\mathbf{v}_j = \sum_{i=1}^{N} a_{ji}\phi(\mathbf{x}_i). \tag{2}$$

$j = 1, 2, \cdots, p$

Next, we find a_{ji}. $\{a_{ji}\}_{i=1}^N \rightarrow v_j$

Substitute (2) into (1): $\underbrace{\frac{1}{N}\sum_{i=1}^{N}\phi(x_i)\phi(x_i)^T}_{C}\underbrace{\left(\sum_{\ell=1}^{N} a_{j\ell}\,\phi(x_\ell)\right)}_{v_j} = \lambda_j \underbrace{\sum_{\ell=1}^{N} a_{j\ell}\,\phi(x_\ell)}_{v_j}$

$$\frac{1}{N}\sum_{i=1}^{N}\phi(x_i)\sum_{\ell=1}^{N} a_{j\ell}\,\phi(x_i)^T\phi(x_\ell) = \lambda_j \sum_{\ell=1}^{N} a_{j\ell}\,\phi(x_\ell)$$

Multiply both side by $\phi(x_k)^T$, $k = 1, 2 \cdots N$

$$\frac{1}{N}\sum_{i=1}^{N}\phi(x_k)^T\phi(x_i)\sum_{\ell=1}^{N} a_{j\ell}\,\phi(x_i)^T\phi(x_\ell) = \lambda_j \sum_{\ell=1}^{N} a_{j\ell}\,\phi(x_k)^T\phi(x_\ell)$$

(N linear equations for a_{ji}, $i = 1, 2, \cdots, N$)

$\underset{\text{kernel}}{k}(x_i, x_k) := \phi(x_i)^T\phi(x_k)$ inner product in F

$$K = \begin{pmatrix} k(x_1, x_1) & & \\ & \ddots & \\ & & k(x_N, x_N) \end{pmatrix}$$

$a_j = (a_{j1}, \cdots a_{jN})^T$

$$\frac{1}{N}\sum_{i=1}^{N} k(x_k, x_i)\sum_{\ell=1}^{N} a_{j\ell}\, k(x_i, x_\ell) = \lambda_j \sum_{\ell=1}^{N} a_{j\ell}\, k(x_k, x_\ell)$$

$$\frac{1}{N}\sum_{\ell=1}^{N} k(x_k, x_i)\sum_{i=1}^{N} a_{j\ell}\, k(x_i, x_\ell) = \lambda_j \sum_{\ell=1}^{N} a_{j\ell}\, k(x_k, x_\ell)$$

$$\frac{1}{N}\sum_{\ell=1}^{N}\Big[\underbrace{\sum_{i=1}^{N} k(x_i, x_\ell)k(x_k, x_i)}_{(KK)_{k\ell}}\Big]\underset{(a_j)_\ell}{a_{j\ell}} = \lambda_j \sum_{\ell=1}^{N} \underset{(K)_{k\ell}}{k(x_k, x_\ell)}\ \underset{(a_j)_\ell}{a_{j\ell}}$$

$$\frac{1}{N}(KK)\underline{a}_j = \lambda_j K \underline{a}_j \leftarrow \text{Matrix form for } N \text{ equations}$$

drop K: $K\underline{a}_j = N\lambda_j \underline{a}_j \Rightarrow \underline{a}_j$ is an eigenvector of K,

---

Goal: $\underline{v}_j$ - PC of $\{\phi(\underline{x}_i)\}$ in F

$$\underline{v}_j = \sum_{i=1}^{N} a_{ji}\phi(\underline{x}_i)$$

$\{a_{ji}\}_{i=1}^{N}$ can be obtained by solving $K\underline{a}_j = N\lambda_j \underline{a}_j$

Normalize $\underline{v}_j$: $1 = \|\underline{v}_j\|_2^2 = \underline{v}_j^T\underline{v}_j = \sum_{k=1}^{N} a_{jk}\phi(\underline{x}_k)^T \sum_{\ell=1}^{N} a_{j\ell}\phi(\underline{x}_\ell)$

$$= \sum_{k=1}^{N}\sum_{\ell=1}^{N} a_{jk}a_{j\ell} \underbrace{\phi(\underline{x}_k)^T\phi(\underline{x}_\ell)}_{K(\underline{x}_k,\,\underline{x}_\ell) = (K)_{k\ell}}$$

$$= \underline{a}_j^T K \underline{a}_j$$

$$\Rightarrow \underline{a}_j^T K \underline{a}_j = 1 \qquad \Rightarrow N\lambda_j \underline{a}_j^T \underline{a}_j = 1$$

$$\left(1 = \underline{a}_j^T K \underline{a}_j = N\lambda_j \underline{a}_j^T \underline{a}_j \right)$$

Solve $\underline{a}_j$ from $K\underline{a}_j = N\lambda_j \underline{a}_j$ with $N\lambda_j \underline{a}_j^T \underline{a}_j = 1$ (Normalization condition)

Summary:

- The principal components in the feature space F are given by

$$\mathbf{v}_j = \sum_{i=1}^{N} a_{ji}\phi(\mathbf{x}_i), \quad j = 1, \cdots, p$$

- The coefficients a_{ji} are computed from the eigenvectors of the kernel matrix K,

$$K\mathbf{a}_j = N\lambda_j\mathbf{a}_j, \quad j = 1, \cdots, p$$

with the constraint

$$N\lambda_j\mathbf{a}_j^T\mathbf{a}_j = 1.$$

project $\phi(\underline{x}_i)$ onto $\underline{v}_j$: $\underbrace{(\phi(\underline{x}_i)^T\underline{v}_j)}_{\text{principal component score}}\underline{v}_j$

$$\phi(\underline{x}_i)^T\underline{v}_j = \phi(\underline{x}_i)^T \underbrace{\sum_{\ell=1}^{N} a_{j\ell}\phi(\underline{x}_\ell)}_{\underline{v}_j} = \sum_{\ell=1}^{N} a_{j\ell} \underbrace{\phi(\underline{x}_i)^T\phi(\underline{x}_\ell)}_{K(\underline{x}_i,\underline{x}_\ell)} = \sum_{\ell=1}^{N} \underset{(\underline{a}_j)_\ell}{a_{j\ell}} K(\underline{x}_i, \underline{x}_\ell)$$

To find the principal score along $\underline{v}_j$, all we need are the coefficients $\{a_{ji}\}_{i=1}^{N}$ and the kernels $K(\underline{x}, \underline{x}_i)$

- So far, we have assumed the features $\{\phi(\mathbf{x}_i)\}_{i=1}^N$ are centered. In general, however, the features $\{\phi(\mathbf{x}_i)\}_{i=1}^N$ may not be centered, even if $\{\mathbf{x}_i\}_{i=1}^N$ is centered.
- Center the features:

$$\tilde{\phi}_i = \phi_i - \frac{1}{N}\sum_{k=1}^N \phi_k,$$

where $\phi_i = \phi(\mathbf{x}_i)$. Then the kernel matrix of the centered features is given by

$N \times N$ matrix of all 1's

$$\tilde{K} = K - \frac{1}{N}\mathbf{1}K - \frac{1}{N}K\mathbf{1} + \frac{1}{N^2}\mathbf{1}K\mathbf{1}$$

where $\mathbf{1}$ is the $N \times N$ matrix of all 1's and K is the kernel matrix of $\{\phi(\mathbf{x}_i)\}_{i=1}^N$.

$\tilde{K} = \left(\tilde{\phi}(\underline{x}_i)^T \tilde{\phi}(\underline{x}_j)\right)_{N\times N}$, $K = \left(\phi(\underline{x}_i)^T \phi(\underline{x}_j)\right)_{N\times N}$

Summary of kernel PCA:

1) Pick a kernel

$$k(\mathbf{x}_i, \mathbf{x}_j): \quad R^d \times R^d \to R.$$

2) Construct the kernel matrix K:

$$(K)_{ij} = k(\mathbf{x}_i, \mathbf{x}_j), \quad i, j = 1, \cdots, N$$

Construct the centered kernel matrix $\tilde{K}$:

$$\tilde{K} = K - \frac{1}{N}\mathbf{1}K - \frac{1}{N}K\mathbf{1} + \frac{1}{N^2}\mathbf{1}K\mathbf{1}$$

3) Compute the first p eigenvectors of $\tilde{K}$:

$$\tilde{K}\mathbf{a}_j = N\lambda_j\mathbf{a}_j, \quad j = 1, \cdots, p$$

with the normalization constraint $N\lambda_j \mathbf{a}_j^T\mathbf{a}_j = 1$.

4) Given a data point $\mathbf{x}$, its principal component scores along the principal component axis in the feature space are

$$y_j = \sum_{i=1}^N a_{ji}\tilde{k}(\mathbf{x}, \mathbf{x}_i), \quad j = 1, \cdots, p$$

where

$$\tilde{k}(\mathbf{x}, \mathbf{x}_i) = k(\mathbf{x}, \mathbf{x}_i) - \frac{1}{N}\sum_{l=1}^N k(\mathbf{x}, \mathbf{x}_l) - \frac{1}{N}\sum_{n=1}^N k(\mathbf{x}_n, \mathbf{x}_i) + \frac{1}{N^2}\sum_{n=1}^N\sum_{l=1}^N k(\mathbf{x}_n, \mathbf{x}_l).$$

In this algorithm, we projected the data $\{\phi(\mathbf{x}_i)\}$ onto p principal components in the feature space.