PCA and Kernel PCA

- Classical PCA

Linear dimension reduction.

given $X = [x_1, \dots, x_n] \subset \mathbb{R}^D$

target $\tilde{x}_i \mapsto \tilde{x}_i$ (centered)

problem: find a projection to dimensional subspace W of $\mathbb{R}^D$, $\dim W = d < D$, s.t. the projected variance is maximal.

max $\frac{1}{n} \sum_{i=1}^{n} \left(\frac{\tilde{x}_i}{\|\tilde{x}_i\|} w \right)^2$

where $\{w_1, \dots, w_d\}$ an onb of W

$\Leftrightarrow \max_{\|w\|=1} \| \underbrace{\tilde{X}^T w}_{\in \mathbb{R}^n} \|^2$

Eq. to: max $\frac{1}{n} \frac{\tilde{x}_i^T w}{\|\tilde{x}_i\|} \tilde{x}_i^T w$

$\Leftrightarrow \max_{\|w\|=1} w^T \tilde{X} \tilde{X}^T w$

Note that $\frac{1}{n} \tilde{X} \tilde{X}^T = \frac{1}{n} \sum_{i=1}^{n} \tilde{x}_i \tilde{x}_i^T$

$= \frac{1}{n} \left(\frac{\tilde{x}_i}{\|\tilde{x}_i\|} \tilde{x}_i^T \right) w$

the problem is equivalent to

max $w^T \tilde{S} w$, $\tilde{S} := \frac{1}{n} \sum_{i=1}^{n} \tilde{x}_i \tilde{x}_i^T$

$\|w\|=1$ Sample cov matrix

w is found by top eigenvector of $\tilde{S}$ (eigenvalue = largest eigenvalue)

Eq. formul case: $1 \le d \le D (< n)$

max $\frac{1}{n} \sum_{i=1}^{n} \frac{\tilde{x}_i}{\|\tilde{x}_i\|} \tilde{x}_i^T w$

$\|w_1, \dots, w_d\| = 1$

i.e. $w_i^T w_j = \delta_{ij}$

Eq. to: find d eigenvector of $\tilde{S}$ by Lanczos-Finder iteration, priority such $w_1, \dots, w_d$ are called principal components (p.c.)

In matrix form

$X = \begin{bmatrix} x_1 & \dots & x_n \end{bmatrix} \in \mathbb{R}^{D \times n}$

$X_c = X H$, $H := I - \frac{1}{n} \mathbb{1} \mathbb{1}^T$ centering matrix

$\tilde{S} = \frac{1}{n} X_c X_c^T$ $D \times D$

$\operatorname{eig}(\tilde{S})$: $\tilde{S} = \tilde{U} \tilde{\Lambda} \tilde{U}^T$,

relates to SVD of X_c: $X_c = U \Sigma V^T$,

$\operatorname{rank}(X_c) = \operatorname{rank}(D, n)$ suppose $n \ge D$,

then $\tilde{S} = U (\frac{1}{n} \Sigma^2) U^T \Leftrightarrow \operatorname{eig}(\tilde{S})$.

The PC.

$U_d = \begin{bmatrix} u_1, \dots, u_d \end{bmatrix} \in \mathbb{R}^{D \times d}$ first d col. of U

The dimension reduced repr.:

$\tilde{X} := U_d^T X_c$

$\begin{array}{c} \begin{array}{ccc} & \overset{n}{\downarrow} & \end{array} \\ d \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \begin{bmatrix} \\ X_c \\ \\ \end{bmatrix} \end{array} = d \begin{bmatrix} \\ \tilde{X} \\ \\ \end{bmatrix}$

back to $\mathbb{R}^D$: $\tilde{X}_c = U_d \tilde{X}$

Check: $\tilde{X}_c = U_d U_d^T X_c$

$\begin{pmatrix} U_d U_d^T \end{pmatrix}_{D \times D}$ is an orthogonal projection to the subspace $U_d \operatorname{span}(\mathbb{R}^{D \times d})$.

eg. PCA can also apply to "Low's data", where also $d \ll n$ and $\operatorname{rank}(X_c) = d$ obtained from X viewed as low-rank data.

Then dim-red mapping $\mathbb{R}^D \to \mathbb{R}^d$

$\begin{cases} x_{new}^{(D)} \to \tilde{x}_{new}^{(d)} & \text{(index of } x_{new} \to x) \\ U_d & \text{old PCs dimensional by } \tilde{x}_{new} \end{cases}$

new sample $x^{(D)} \in \mathbb{R}^D$

$\tilde{x}_c^{(d)} = U_d^T \left(x^{(D)} - \frac{1}{n} X^{(D)} \mathbb{1} \right)$

PCA dimensional not part of $x^{(D)}$

$\tilde{x}^{(d)} = \underbrace{U_d U_d^T}_{\tilde{U}_d} \left(x^{(D)} - \frac{1}{n} X^{(D)} \mathbb{1} \right) + \frac{1}{n} X^{(D)} \mathbb{1}$ $\quad \in \mathbb{R}^D$

- Another view pt of classical PCA

Simple to notice: $\tilde{S}$, $X^{\times}$ $[x_1, \dots, x_n]_{D \times n}$

$X \tilde{S} = X X^T$ $D \times D$

Gram matrix: $G = X^T X$, where

$G_{ij} = X^T x_j$, $i, j = 1, \dots, n$

$X = U \Sigma V^T$, $(X \tilde{S}) = U \Sigma^3 U^T$,

$G = V \Sigma^2 V^T$

Q. Can we formulate PCA without X and G?

$G = V \Lambda V^T$, $\Lambda = \Sigma^2$.

Then also find d singular values/eigenvalues of X $\lambda_1, \dots, \lambda_d$,

$v_1, \dots, v_d$

can be obtained from G only.

We can also construct $\{u_i\} = V \Lambda^{1/2} V^T$

$V \Lambda = \begin{bmatrix} v_1, \dots, v_d \end{bmatrix}_{n \times d}$

$\Lambda_d = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{bmatrix} = \begin{bmatrix} \sqrt{\lambda_1} & & \\ & \ddots & \\ & & \sqrt{\lambda_d} \end{bmatrix}_{d \times d}$

for any new $x^{(D)} \in \mathbb{R}^D$,

$X_c^T x^{(D)} = V \Sigma U^T x^{(D)}$

then

$\left(V_d \Lambda_d^{-1/2} \right) \left(V^T x^{(D)} \right)$

$= \underbrace{U_d U_d^T}_{\text{this implies the}} V \Sigma U_d^T x^{(D)}$

$= U_d^T \Sigma U_d^T x^{(D)}$ dim-red repr.

where $U_d = \begin{bmatrix} u_1 \\ \vdots \\ u_d \end{bmatrix}_{d \times D}$ $U_d^T = \begin{bmatrix} u_1, \dots, u_d \end{bmatrix}_{D \times d}$

also

$X_c \left(V_d \Lambda_d^{-1/2} \right) = \quad \cdots \quad (*)$

$= \left(U \Sigma V^T \right) \left(V_d \Lambda_d^{-1/2} \right) V \Sigma U^T x^{(D)}$

$= U_d \Sigma V_d^T \Lambda_d^{-1/2} V^T x^{(D)}$ this implies the dimred $\tilde{x}^{(d)} = \tilde{x}^{(d)}$

- Kernel PCA

Recall the dual case of classical PCA.

max $\frac{1}{n} \sum_{i=1}^{n} \left(w^T (x_i - \bar{x}) \right)^2$ $\quad \tilde{x}_i = \frac{1}{n} \sum_{j=1}^{n} x_j$

$\|w\|^2 = w^T w = 1$

$\Leftrightarrow$ PCA: $\operatorname{eig}(\tilde{S})$ $\tilde{S} = \frac{1}{n} X_c X_c^T$ covariance matrix of n D-dim data.

Claim: max $\frac{1}{n} \sum_{i=1}^{n} \left(\tilde{\phi}(x_i)^T w \right)^2$ $\quad \tilde{\phi}(x_i) = \phi(x_i) - \frac{1}{n} \sum_{j=1}^{n} \phi(x_j)$

$\|w\|^2 = 1$ instead of $\tilde{\phi}(x_1), \dots, \tilde{\phi}(x_n)$

$\|\tilde{\phi}(x_1), \dots, \tilde{\phi}(x_n)\| = 1$ a subspace of the ϕ of the xs.

(Why? Dis ex., $\tilde{\phi}(x_1), \dots, \tilde{\phi}(x_n)$ $\Rightarrow$ $\exists$ ϕ the orthogonal projection to the $\tilde{\phi}$-dimensional subspace, $\tilde{\phi} \in \mathbb{R}^D$ the new coordinates in the projected variables.)

Then we write

$\tilde{\phi} = \frac{1}{n} \sum_{i=1}^{n} \tilde{\phi}(x_i)$ $\quad \phi \in \mathbb{R}^D$, $w \in \mathbb{R}^D$.

then

$\langle \tilde{\phi}, \tilde{\phi}(x_i) \rangle_{\mathcal{H}} = \frac{1}{n} \sum_{j=1}^{n} \phi(x_j)^T \tilde{\phi}(x_i)$

$= \left(\tilde{K} \mathbb{1}_n \right)_i$, $\quad i = 1, \dots, n$

where $\tilde{K}_{ij} = \langle \tilde{\phi}(x_i), \tilde{\phi}(x_j) \rangle_{\mathcal{H}} = \tilde{\phi}(x_i)^T \tilde{\phi}(x_j)$

Then we write

$\left(\tilde{K}_{ij} = \langle \tilde{\phi}(x_i) - \tilde{\phi}_n, \tilde{\phi}(x_j) - \tilde{\phi}_n \rangle_{\mathcal{H}} \right.$

$= K(x_i, x_j) - \langle \tilde{\phi}_n, \tilde{\phi}(x_j) \rangle - \tilde{\phi}_n \rangle_{\mathcal{H}}$

$= \langle \phi(x_i) - \frac{1}{n} \sum_{k=1}^{n} \phi(x_k), \phi(x_j) - \frac{1}{n} \sum_{k=1}^{n} \phi(x_k) \rangle$

$- \langle \tilde{\phi}_n, \tilde{\phi}(x_j) \rangle_{\mathcal{H}}$

$= \langle \phi(x_i), \phi(x_j) \rangle - \frac{1}{n} \sum_{k=1}^{n} K(x_i, x_k) - \frac{1}{n} \sum_{k=1}^{n} K(x_k, x_j)$

$+ \frac{1}{n^2} \sum_{k,l=1}^{n} K(x_k, x_l)$

$\langle \tilde{\phi}_n, \tilde{\phi}(x_j) \rangle_{\mathcal{H}} = \frac{1}{n} \sum_{k=1}^{n} \langle \tilde{\phi}(x_k), \tilde{\phi}(x_j) \rangle_{\mathcal{H}}$

$\frac{1}{n} \sum_{k=1}^{n} K(x_k, x_j)$ required as above

$\mathbb{1}_n$: $x^{(n)} \Leftrightarrow \left(\langle \tilde{\phi}_n, \tilde{\phi}(x^{(n)}) \rangle_{\mathcal{H}} \right)_{n \times 1} \in \mathbb{R}^n$

- The dimred input by kernel PCA

suppose $K := \left(K(x_i, x_j) \right)_{i,j=1}^{n}$ general kernel case

$\tilde{K} = H K H$ $\quad$ no being done.

$\tilde{K} = V \Lambda V^T$, $\operatorname{eig}(\tilde{K})$

$\tilde{g}(\tilde{K}) = V_d \Lambda_d^{-1/2} V_d^T$ dimensional normal of $\tilde{K}$

note. $V_d = \begin{bmatrix} v_1, \dots, v_d \end{bmatrix}_{n \times d}$ $\quad V_d V_d^T = I_d$

$\Lambda_d^{-1/2} V_d^T \tilde{g}_d = V_d \Lambda_d^{-1/2} V_d^T = V_d \Lambda_d^{-1/2} V_d$

Compute $\tilde{K}_{new} = \left(\tilde{K}(x_i, x^{(n)}) \right)_{i=1}^{n}$ as above

$x^{(n)}$ is the dimred input $\Leftrightarrow$

$\tilde{\phi}(x_i) = \frac{1}{n} \sum_{j=1}^{n} \phi(x_j)$

$\langle \tilde{\phi}(x_i), \tilde{\phi}(x_j) \rangle_{\mathcal{H}} = \frac{1}{n} \sum_{k=1}^{n} \langle \tilde{\phi}_k, \tilde{\phi}(x_i) \rangle_{\mathcal{H}} \langle \tilde{\phi}_k, \tilde{\phi}(x_j) \rangle_{\mathcal{H}}$

$= \frac{1}{n} \left(\tilde{K} \mathbb{1}_n \right)_i \left(\tilde{K} \mathbb{1}_n \right)_j$

$= \tilde{K}_{ij} \left(\frac{1}{\sqrt{\sum_{k=1}^{n} \tilde{K}_{ik}^2}} \tilde{K} \mathbb{1}_n \right)_i$

$\frac{1}{\sqrt{\sum_{k=1}^{n} \tilde{K}_{ik}^2}} = \tilde{g}(\tilde{K})$

The problem is to show for $x^{(n)} \in$

$\tilde{\phi}(x^{(n)}, x_i) \Leftrightarrow \tilde{K}_{ij}^{(n)} \frac{\tilde{g}(\tilde{K})_{ii}}{\tilde{g}(\tilde{K})_{jj}}$ $\quad i, j = 1, \dots, n$

which same required form $\tilde{g}(\tilde{K})$

But $\tilde{K}(x^{(n)}, x_i)$ is nonlinear, this involves slow optimization problem.

- Density by kernel projections

In kernel ridge regression, we consider $(X, Y) \sim P_{XY}$ training data.

$\hat{P}_X^{(n)} := \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$, $\quad \hat{P}_{XY}^{(n)} := \frac{1}{n} \sum_{i=1}^{n} \delta_{(x_i, y_i)}$

If we set $\tilde{P}^{(n)} \in \mathcal{H}$, and variable

$x \mapsto \tilde{\phi}^{(n)}$ be as above, $\phi(x) \in \mathcal{H}$ $\quad \phi \in \mathcal{H}$ given $\{x_1, \dots, x_n\}$.

for each $y = 1, \dots, D$

$\tilde{P}_Y^{(n)} = \left[\tilde{P}_Y^{(n)} \right]^T = \left[\tilde{C}_{YY}^{(n)} \right]^{-1} \left[\tilde{C}_{YY}^{(n)} \right]^{-1}$, $\quad \tilde{P}^{(n)} = \left[\tilde{C}_{YY}^{(n)} \right]^{-1}$

$\tilde{P}_Y^{(n)} \left(\tilde{\phi}^{(n)} \right) = \frac{1}{n} \sum_{i=1}^{n} \left(\tilde{\phi}^{(n)} \left(x^{(n)}, x_i \right) \right)^T \tilde{P}^{(n)}$

$= \frac{1}{n} \sum_{i=1}^{n} \left(\tilde{K}^{(n)} \right)_{ij} \left(\tilde{P}^{(n)} \right)_{ij}$

$= \frac{1}{n} \left(\tilde{K}^{(n)} \right)^T \left(\tilde{P}^{(n)} \right) \tilde{K}_{new}$

Concatenate all $y = 1, \dots, D$

$\tilde{g}^{(n)} \quad \times \quad \tilde{g}(\tilde{K}) \tilde{K}_{new} \quad \cdots (*)$

$(*)$ actually $(*)$ the dimred PCA density