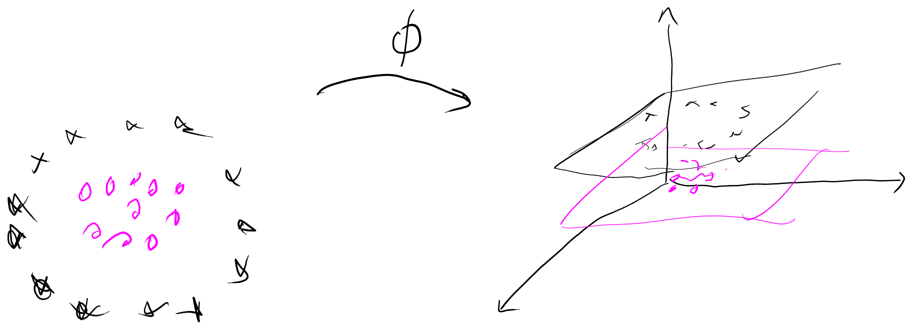
Kernel PCA



$$\phi : \mathbb{R}^2 \longrightarrow \mathbb{R}^3$$
$$(x,y) \longmapsto (x, y, x^2+y^2)$$

Consideremos um conj^{to} com N observações,

$\{x_n\}_{n=1}^N$, $x_i \in \mathbb{R}^D$, $\sum_{i=0}^N x_i = 0$

$S_{D\times D}$ cov. amostral $\qquad S u_i = \lambda_i u_i$

$\parallel$

$\frac{1}{N}\sum_{i=1}^N x_i x_i^T \qquad \lambda_i \geq 0$, $u_i \perp u_j$, $i \neq j$

$\| u_i \| = 1$

kPCA $\qquad \phi : \mathbb{R}^D \rightarrow \mathbb{R}^M \qquad M > D$

$x_i \longmapsto \phi(x_i)$ "projecção"

Assumimos que $\phi(x_i)$ estão centrados em 0

$$\sum_{i=1}^N \phi(x_i) = 0$$

$$C_{M\times M} = \frac{1}{N}\sum_{i=1}^{N} \phi(x_i)\,\phi(x_i)^T = \sum_{i=1}^{M} \lambda_i v_i v_i^T$$

$$\lambda_i \geq 0 \;, \quad v_i \perp v_j$$
$$\|v_i\| = 1$$

onde λ_i são os valores prop. de C

e v_i vect. prop. assoc. v.p. λ_i

i.e., $$C v_i = \lambda_i v_i$$

Se $\lambda_i \neq 0$,

$$C v_i = \frac{1}{N}\sum_{j=1}^{N} \phi(x_j)\,\underbrace{\phi(x_j)^T v_i}_{\in \mathbb{R}} = \lambda_i v_i$$

$$\Rightarrow v_i = \sum_{j=1}^{N} \underbrace{\left(\frac{1}{N\lambda_i}\,\phi(x_j)^T v_i\right)}_{=:\,a_{ij}} \phi(x_j)$$

$$\Rightarrow v_i = \sum_{j=1}^{N} a_{ij}\,\phi(x_j)$$

Ou seja, os vect. prop. de C assoc. a v.p. $\neq 0$

são combinações linear dos $\phi(x_n)$.

$$\lambda_i v_i = \lambda_i \sum_{j=1}^{N} a_{ij}\, \phi(x_j)$$

$$\parallel$$

$$C v_i = \frac{1}{N} \sum_{j=1}^{N} \phi(x_j) \left(\phi(x_j)^T v_i \right)$$

$$= \frac{1}{N} \sum_{j} \left(\phi(x_j)\, \phi(x_j)^T \right) \underbrace{v_i}_{= \sum a_{ij} \phi(x_j)}$$

$$= \frac{1}{N} \sum_{n} \phi(x_n)\, \phi(x_n)^T \sum_{j} a_{ij}\, \phi(x_j)$$

$$\lambda_i \sum_{j} a_{ij}\, \phi(x_j) = \frac{1}{N} \sum_{n} \phi(x_n)\, \phi(x_n)^T \sum_{j} a_{ij}\, \phi(x_j)$$

LHS $\phi(x_{\cancel{j}k})^T$

$$\Longrightarrow \quad \lambda_i \sum_{j} a_{ij}\, \phi(x_k)^T \phi(x_j) = \frac{1}{N} \sum_{n} \phi(x_k)^T \phi(x_n) \sum_{j} a_{ij}\, \phi(x_n)^T \phi(x_j)$$

Seja $\quad k(x_n, x_m) = \phi(x_n)^T \phi(x_m)$

obtemos

$$\frac{1}{N} \sum_{n} k(x_k, x_n) \sum_{j} a_{ij}\, k(x_n, x_j) = \lambda_i \sum_{j} a_{ij}\, k(x_k, x_j)$$

$$K = \begin{bmatrix} k(x_i, x_j) \end{bmatrix} \qquad a_i = \begin{bmatrix} a_{1i} \\ a_{2i} \\ \vdots \\ a_{Ni} \end{bmatrix}$$

donde

$$K^2 a_i = \lambda_i N K a_i$$

As soluções de $K a_i = (\lambda_i N) a_i$

são soluç de $K^2 a_i = \lambda_i N K a_i$

Ou seja, a_i são vectores próprios de K assoc. v.p. $\lambda_i N$

$$1 = v_i^T v_i = \sum_n a_{in} \phi(x_n)^T \sum_j a_{ij} \phi(x_j)$$

$$= \sum_{n,j} a_{in} a_{ij} \underbrace{\phi(x_n)^T \phi(x_j)}_{k(x_n, x_j)}$$

$$= a_i^T \underbrace{(K a_i)}_{= \lambda_i N a_i} = \lambda_i N a_i^T a_i = \lambda_i N \|a_i\|^2$$

$$\lambda_i N \|a_i\|^2 = 1$$

As "projecções" de x sobre o vect. próp. v_i

$$y_i(x) = \phi(x)^T v_i = \sum_j a_{ij} \underbrace{\phi(x)^T \phi(x_j)}_{k(x, x_j)} = \sum_j a_{ij} k(x, x_j)$$

i.e., a "proj." de x sobre α^i é comb. linear de $k(x, x_j)$

Em $\mathbb{R}^D$, tem D vectores prop. ortonormados

Em $\mathbb{R}^M$ vamos ter c.p. em número $> D$

No entanto o número de c.p. em $\mathbb{R}^M$ é $< N$

porque $\text{car}(K) \leq N$ $\quad K_{N \times N}$

$$\text{car}(K) = \text{car}(C)$$

NOTA. Assumimos que $\{\phi(x_i)\}_{i=1}^N$ estavam centrados na média 0, o que em geral pode não acontecer

$$\tilde{\phi}(x_n) = \phi(x_n) - \frac{1}{N}\sum_{j=1}^N \phi(x_j)$$

$$(\tilde{K})_{n,m} = \tilde{\phi}(x_n)^T \tilde{\phi}(x_m) = \left(\phi(x_n) - \frac{1}{N}\sum_j \phi(x_j)\right)^T\left(\phi(x_m) - \frac{1}{N}\sum_i \phi(x_i)\right)$$

$$= \phi(x_n)^T \phi(x_m) - \frac{1}{N}\sum_j \phi(x_j)^T \phi(x_m) - \frac{1}{N}\sum_i \phi(x_n)^T \phi(x_i) +$$

$$+ \frac{1}{N^2}\sum_j\sum_i \phi(x_j)^T \phi(x_i)$$

$$= k(x_n, x_m) - \frac{1}{N}\sum_j k(x_j, x_m) - \frac{1}{N}\sum_i k(x_n, x_i) + \frac{1}{N^2}\sum\sum k(x_j, x_i)$$

$$\tilde{K} = K - 1\!\!1_{1/N} K - K 1\!\!1_{1/N} + 1\!\!1_{1/N} K 1\!\!1_{1/N} = K - 2\, 1\!\!1_{1/N} K + 1\!\!1_{1/N} K 1\!\!1_{1/N}$$

$$1\!\!1_{1/N} = \begin{bmatrix} 1/N & 1/N & \cdots & 1/N \\ \vdots & & & \\ 1/N & \cdots & & 1/N \end{bmatrix}$$

Calculamos K usando $\phi(x_i)$;

$$(K)_{i,j} = k(x_i, x_j) = \phi(x_i)^T \phi(x_j)$$

que usamos para obter $\tilde{K}$

$$\tilde{K} a_i = \lambda_i a_i$$

"Projectamos" então os dados $x \in \mathbb{R}^D$ em $\mathbb{R}^M$

$$y_i(x) = \sum_j a_{ij} K(x, x_j)$$

Obs. PCA é um caso particular do KPCA
$$k(x, x') = x^T x'$$

Vários exemplos $k(x, x')$:

$$k(x, x') = \exp\left(-\|x - x'\|^2 / 0.1\right)$$
$$K(x, x') = \exp\left(-\|x - x'\|^2 / 2\sigma^2\right)$$
$$K(x, x') = \left(x^T x'\right)^d$$
$$K(x, x') = \left(x^T x' + c\right)^d$$

Resumo p/ KPCA

1) Obter a matriz kernel K dos dados $\{x_n\}_{n=1}^N$
centrados em 0

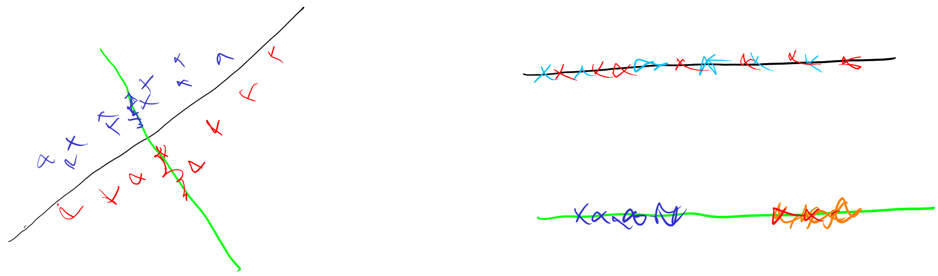
2) Calcular $\tilde{K}$ (matriz de Gram)

3) Obter val. p. & vect. p. $\tilde{K} a_k = \lambda_k N a_k$

4) $y_k(x) = \phi(x)^T v_k = \sum_{i=1}^N a_{ki} k(x, x_i)$

LDA

linear discriminant analysis



$$C = C_1 \cup C_2 \qquad C_1 \cap C_2 = \emptyset$$

média da classe i

$$\mu_i = \frac{1}{N_i} \sum_{x \in C_i} x \quad ; \quad N_i = \# C_i$$

média amostral

$$\mu = \frac{1}{N} \sum_{x \in C} x \quad ; \quad N = \# C$$

S_B matriz de dispersão inter-classes (between scatter matrix)

k = nº de classes

$$S_B = \sum_{i=1}^{k} N_i (\mu_i - \mu)(\mu_i - \mu)^T \qquad \text{car } S_B \leq k-1$$

S_W matriz dispersão intra-classe (within scatter-matrix)

$$S_W = \sum_{i=1}^{k} \sum_{x \in C_i} (x - \mu_i)(x - \mu_i)^T$$

$$= \Sigma_1 + \Sigma_2 + \dots + \Sigma_k$$

Σ_i é a matriz de dispersão de C_i

O método de Fisher consiste em maximizar

$$J(w) = \frac{w^T S_B w}{w^T S_w w}$$

é o mesmo que $\max w^T S_B w$

sujeito $w^T S_w w = k'$

Usando multiplicadores de Lagrange

$$L(w,\lambda) = w^T S_B w - \lambda (w^T S_w w - k')$$

$$= w^T (S_B - \lambda S_w) w + \lambda k'$$

$$\nabla_w L = 2(S_B - \lambda S_w) w = 0$$

$$S_B w = \lambda S_w w$$

se $S_w = \Sigma_1 + \Sigma_2 + \dots + \Sigma_k$ for invertível,

$$S_w^{-1} S_B w = \lambda w$$

Ou seja w é vect. próp. de $S_w^{-1} S_B$ assoc. λ

Como $S_B w = \lambda S_w w$ então

$$w^T S_B w = \lambda \underbrace{(w^T S_w w)}_{k' \text{ cons}^{te}}$$

Maximizar $w^T S_B w$ é o mesmo que maximizar λ

w é um vect. próp. generalizado (no 2º sentido)