April 28, 2020 Kernel methods

$$-\|x_i - x_j\|_2^2 = -\|x_i\|_2^2 - \|x_j\|_2^2 + 2x_i^T x_j$$

$x_i^T x_j \approx$ how similar x_i, x_j are

node2vec
t-SNE
PCA

$k(x_i, x_j) = x_i^T x_j$ (similarity)

$k(x_i, x_j) = -\|x_i - x_j\|_2^2$

Example: squared exponential / Gaussian

$$k(x_i, x_j) = s^2 \exp\left(-\frac{1}{2}\|x_i - x_j\|_2^2 / l^2\right)$$

Example: polynomial (degree 2)

$$k(x_i, x_j) = (c + x_i^T x_j)^2$$

$$(c + y^T z)^2 = \left(c + \sum_{k=1}^{d} y_k z_k\right)^2$$

$$= c^2 + 2c \sum y_k z_k + \left(\sum y_k z_k\right)^2$$

$$= c \cdot c + \sum (\sqrt{2c}\, y_k)(\sqrt{2c}\, z_k) \quad \sum_{k=1}^{d} y_k^2 z_k^2 + \sum_{i<j} (\sqrt{2} y_i y_j)(\sqrt{2} z_i z_j)$$

$$\phi(x) = (c, \sqrt{2c}\, x_1, \ldots, \sqrt{2c}\, x_d, x_1^2, \ldots, x_d^2, \sqrt{2} x_1 x_2, \ldots, \sqrt{2} x_{d-1} x_d)$$

$$= \phi(y)^T \phi(z) \qquad \phi : x \in \mathbb{R}^d \to \phi(x) \in \mathbb{R}^{1+d+d+\binom{d}{2}}$$

$= k(y, z)$

ϕ is called a feature map

What about Gaussian?

$$k(x_i, x_j) = \exp\left(-\frac{1}{2}\|x_i - x_j\|_2^2\right) \qquad r = \|x_i - x_j\|_2$$

$$\exp\left(-\tfrac{1}{2} r^2\right)$$

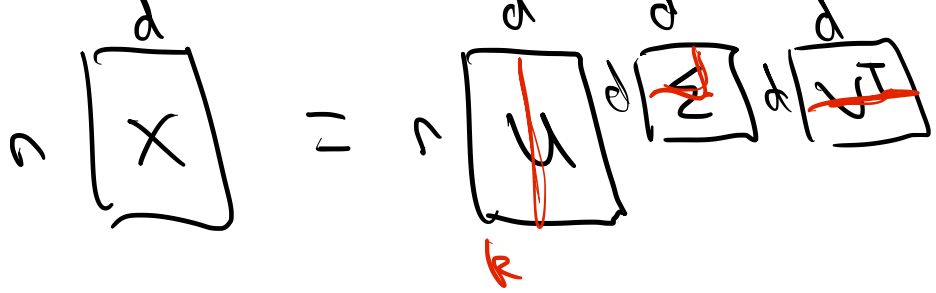
$$= 1 - \frac{r^2}{2} + \frac{r^4}{8} - \frac{r^6}{48} + \cdots$$

Key idea: $k(x_i, x_j)$ often easy to compute

$k(x_i, x_j) = x_i^T x_j$ $\quad k(x_i, x_j) = \exp(-\|x_i - x_j\|_2^2)$

Key idea: if our method only relies on $x_i^T x_j$ as measure of similarity, then we can think of using $k(x_i, x_j)$ "kernel trick"

Example: kernel PCA



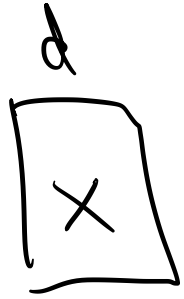
columns of X (features) have zero mean

PCs: V_k

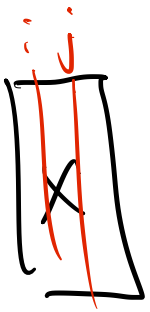
Projection: $XV_k = U_k \Sigma_k$

$$X^T X = V \Sigma U^T U \Sigma V^T = V \Sigma^2 V^T$$

PCA: first k eigenvectors of $X^T X$ sample covariance



$$(X^T X)_{ij} = x_i^T x_j$$



$X^T X \Rightarrow K \quad K_{ij} = k(x_i, x_j)$ drop-in replacement?

$k(x, y) = k(y, x)$ symmetric

$K \succeq 0$

$\bar{V}_k$ = first k eigenvectors of $\bar{K}$

$$\bar{K} = \left(I - \frac{1}{n} 1 1^T\right) K \left(I - \frac{1}{n} 1 1^T\right)$$

zero mean

Function approx.

$x_1, \dots, x_n \in \mathbb{R}^d \qquad f: \mathbb{R}^d \to \mathbb{R}$

demographic data in county $\xrightarrow{f}$ election (% Biden / Trump)

$f \approx \hat{f} \qquad \hat{f}(x) = x^T c^*$

$$c^* = \min_c \| Xc - f_X \|_2^2$$

X: $n \times d$ with rows $x_1^T, \dots, x_n^T$

f_X: $n \times 1$ with entries $f(x_1), \dots, f(x_n)$

Basis functions $\phi_1, \dots, \phi_d \qquad \phi_j(x) = x_j$

$$\hat{f} = \sum c_j \phi_j \qquad \hat{f}(x) = \sum c_j \phi_j(x) = \sum c_j x_j$$
$$= c^T x = x^T c$$

More complex space

$\psi_j : \mathbb{R}^d \to \mathbb{R}$ $\quad j = 1, \dots, m$

$\hat{f} = \sum_{j=1}^{m} c_j \psi_j$ $\qquad \hat{f}(x) = \sum_{j=1}^{m} c_j \psi_j(x)$

$$\min_c \| \Psi c - f_x \|_2^2$$

$$\Psi = \begin{bmatrix} \psi_1(x_1) & \cdots & \psi_m(x_1) \\ \vdots & & \vdots \\ \psi_1(x_n) & \cdots & \psi_m(x_n) \end{bmatrix} \quad n \times m$$

Key idea: ψ_j can be nonlinear (feature maps)
still get OLS

- $n \geq m \Rightarrow c^* = (\Psi^T \Psi)^{-1} \Psi^T f_x$
- $n \leq m \Rightarrow$ HW1 $\Psi c^* = f_x$

$$\min_c \|c\|_2^2 \quad \text{s.t.} \quad \Psi c = f_x \Rightarrow c^* = \Psi^T (\Psi\Psi^T)^{-1} f_x$$

$$\hat{f} = \sum c_j^* \psi_j \qquad \hat{f}(z) = \sum c_j^* \psi_j(z) = \psi(z)^T c^*$$

$$= \psi(z)^T \Psi^T (\Psi\Psi^T)^{-1} f_x$$

$\psi(z)^T = [\psi_1(z) \cdots \psi_m(z)]$ (1 × m)

Ψ^T (m × n)

K_{xx} $(K_{xx})_{ij} = \psi(x_i)^T \psi(x_j) = k(x_i, x_j)$

$\triangleq K_{zx}$ $(K_{zx})_j = \sum_\ell \psi_\ell(z)\psi_\ell(x_j) = \psi(z)^T \psi(x_j) = k(z, x_j)$

$$\hat{f}(z) = k_{zx} K_{xx}^{-1} f_x \qquad K_{xx} > 0$$

Vector space H with orthonormal basis $\psi_1, \ldots, \psi_m$

Approx. f with $f \in H$ $\langle \psi_i, \psi_j \rangle_H = \begin{cases} 1 & i = j \\ 0 & o/w \end{cases}$

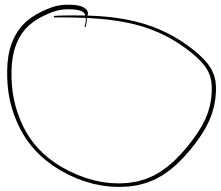
$$\hat{f} = \sum c_j \psi_j$$

$$\|\hat{f}\|_H^2 = \left\langle \sum_j c_j \psi_j, \sum_i c_i \psi_i \right\rangle_H$$

$$= \sum_{ij} c_i c_j \langle \psi_i, \psi_j \rangle_H$$

$$= \sum_i c_i^2 \langle \psi_i, \psi_i \rangle_H = \sum_i c_i^2 = \|c\|_2^2$$

$$\min_{\hat{f}} \|\hat{f}\|_H^2 \quad \text{s.t.} \quad \hat{f}_x = f_x$$

example:  $\langle \psi_i, \psi_j \rangle_H = \int_{S^1} \psi_i(x) \psi_j(x) dx$

Point approx

$k_y \in H \quad k_y(x) = \sum_{j=1}^{m} \psi_j(y)\,\psi_j(x)$

$k_y = \sum \psi_j(y)\,\psi_j$

$g \in H \quad g = \sum a_i \psi_i \quad g(y) = \sum a_i \psi_i(y)$

$$\langle g, k_y \rangle_H = \left\langle \sum a_i \psi_i, \sum \psi_j(y)\,\psi_j \right\rangle_H$$

$$= \sum_{ij} a_i \psi_j(y) \langle \psi_i, \psi_j \rangle_H$$

$$= \sum_{i} a_i \psi_i(y) \underbrace{\langle \psi_i, \psi_i \rangle_H}_{1}$$

$$= g(y)$$

"reproducing"

(need to be more careful in infinite dims)

H is a Reproducing Kernel Hilbert Space (RKHS)
if $g(y) = \langle g, k_y \rangle_H$ for any $g \in H$

feature maps orthonormal basis $\psi_1, \ldots, \psi_m$ for $\mathcal{H}$

$\Rightarrow$ RKHS

Need: inner product $\langle \psi(x_i), \psi(x_j) \rangle_{\ell_2}$

Other direction: RKHS $\Rightarrow$ kernel function

kernel $\Rightarrow$ RKHS