April 30, 2020

Kernel: $k(x,y) = k(y,x)$ $\in \mathbb{R}^d$ measure similarity

Feature maps: $\psi_1, \dots, \psi_m$ $\psi_j : \mathbb{R}^d \to \mathbb{R}$

$\Psi : \mathbb{R}^d \to \mathbb{R}^m$ $\Psi(x) = \begin{pmatrix} \psi_1(x) \\ \vdots \\ \psi_m(x) \end{pmatrix}$ $m > n$ = # of data points

Data: $X = \{x_1, \dots, x_n\}$ $x_i \in \mathbb{R}^d$

$\star$ $\min_c \|c\|_2^2$ s.t. $\Psi c = f_x$

$\Psi = \begin{bmatrix} \Psi(x_1)^T \\ \vdots \\ \Psi(x_n)^T \end{bmatrix}$ $n$ $m$

$\begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix}$

$\hat{f} = \sum_{j=1}^m c_j \psi_j$ $\hat{f} \approx f$

if $\psi_1, \dots, \psi_m$ are orthonormal basis for some $\star$ inner product space $H$

$\star$ $\min \|\hat{f}\|_H^2$ s.t. $\hat{f}_x = f_x$

$k_y \in H \quad k_y = \sum_j \underbrace{\psi_j(y)}_{\text{coeff}} \underbrace{\psi_j}_{\text{function}}$

$k_y(x) = \sum_j \psi_j(y)\psi_j(x)$

$= \langle \psi(y), \psi(x) \rangle_{\ell_2}$

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

$\langle k_y, k_x \rangle_H$

$= \left\langle \sum_i \psi_i(y)\psi_i, \sum_j \psi_j(x)\psi_j \right\rangle_H$

$= \sum_{ij} \psi_i(y)\psi_j(x) \langle \psi_i, \psi_j \rangle_H \quad \delta_{ij}$

$= \sum_j \psi_j(y)\psi_j(x)$

$k_y(x) = k(y,x) = \langle \psi(y), \psi(x) \rangle_{\ell_2} = \langle k_y, k_x \rangle_H$

RKHS: $g(x) = \langle g, k_x \rangle_H$ for any $g \in H$

Note: need to be careful in infinite dimensions

Note! $x_1, \ldots, x_n \quad (K_{XX})_{ij} = k(x_i, x_j)$

$K_{XX} = \Psi \Psi^T$ $n \left[ \Psi \right]_m$ positive semi-def.

RKHS $\Rightarrow$ kernels $k_y = \sum \psi_j(y)\psi_j$

Can we just start with a kernel?

We need an inner product: $\langle \cdot, \cdot \rangle_H$

$x_1, \dots, x_n \quad k_{x_i} \quad k_{x_i}(y) = k(x, y)$ given $k(x,y)$

Suppose $f = \sum_i a_i k_{x_i}, \quad g = \sum_j b_j k_{x_j}$

Propose: $\langle f, g \rangle_H = \left\langle \sum_i a_i k_{x_i}, \sum_j b_j k_{x_j} \right\rangle_H$

$$= \sum_{ij} a_i b_j \langle k_{x_i}, k_{x_j} \rangle_H$$

$$= \sum_{ij} a_i b_j k(x_i, x_j) \quad \text{proposal}$$

$$= a^T K_{xx} b$$

Works if $f, g$ finite lin. combos of $k_{x_i}$

Need to "complete" space with Cauchy sequences.

native space RKHS

$\langle f, f \rangle_H = f^T K_{xx} f \geq 0$ (need)

## Moore-Aronszajn theorem

If $K_{xx}$ is positive semidef for any finite collection of points $X$, then there exists a corresp. unique RKHS

Difficult to characterize functions in native space

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

$\hat{f}(z) = k_{zx} K_{xx}^{-1} f_x$ (last lecture)

$k_{zx} = [k(z, x_1) \cdots k(z, x_n)]$

Can derive error bounds $|\hat{f}(z) - f(z)|$ in terms of $\|f\|_H^2$, $\|\hat{f}\|_H^2$

# Gaussian Processes (GPs)

Multivariate normal: distribution for vectors

$$y \sim N(\mu, \Sigma) \qquad p(y) = \frac{\exp\left(-\frac{1}{2}(y-\mu)^T \Sigma^{-1}(y-\mu)\right)}{\sqrt{(2\pi)^d \det(\Sigma)}}$$

GPs: "distribution" for functions $\{f : \mathbb{R}^d \to \mathbb{R}\}$

1. mean field $\mu : \mathbb{R}^d \to \mathbb{R}$
2. kernel function

Definition: $f \sim GP(\mu, k)$ if for any $x_1, \dots, x_n$

$$f_x \sim N(\mu_x, K_{xx})$$

$$f_x = \begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix} \qquad \mu_x = \begin{pmatrix} \mu(x_1) \\ \vdots \\ \mu(x_n) \end{pmatrix} \qquad (K_{xx})_{ij} = k(x_i, x_j)$$

Bayesians: start with prior, observe data, get posterior

Multivariate normal

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \sim N\left(0, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right)$$

$$p(y_2 \mid y_1) \sim N\left(\Sigma_{21}\Sigma_{11}^{-1} y_1,\ \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)$$

GP: prior $m(z) = 0$ observe $X = x_1, \ldots, x_n$, $f_X = \begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix}$

$\Rightarrow$ conditioning on $X$ ... gives new GP

$$f_{Y \mid X, f_X} \sim N\left(\hat{m}_Y, \hat{K}_{YY}\right)$$

$$\hat{m}(z) = k_{zX} K_{XX}^{-1} f_X \qquad \star \quad \hat{m} = \underset{f \text{ s.t. } f_X = f_X}{\operatorname{argmin}} \|f\|_H^2$$

$[k(z,x_1) \cdots k(z,x_n)]$

$$\hat{k}(z,z') = k(z,z') - k_{zX} K_{XX}^{-1} k_{Xz} \qquad k_{Xz} = k_{zX}^T$$

$x_i$ demo features, $f(x_i)$ election results

$$f_{Y|X} \sim N(\hat{\mu}_Y, \hat{K}_{YY})$$

now we get quantifiable uncertainty

Another: time series: $x_i$ timestamps $f(x_i)$ observed values

Example: noisy observations

$f \sim GP(0, k)$ observe: $x_i, f(x_i) + \varepsilon_i, \varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$

Given $\sigma, X, f_x + \varepsilon$, posterior is

$$f_{Y|X,\sigma} \sim N(\tilde{\mu}_Y, \tilde{K}_{YY})$$

$\begin{pmatrix} \tilde{\mu}(y_1) \\ \vdots \\ \tilde{\mu}(y_1) \end{pmatrix}$

$$\tilde{\mu}(z) = k_{zx}(K_{xx} + \sigma^2 I)^{-1} f_x$$

$$\tilde{k}(z, z') = k(z, z') - k_{zx}(K_{xx} + \sigma^2 I)^{-1} k_{xz}$$

$K_{XX} + \sigma^2 I$ instead of $K_{XX}$

$$\hat{\mu} = \operatorname{argmin} \|\hat{f}\|_H^2 \quad \text{s.t. } \hat{f}_X = f_X$$

$$\tilde{\mu} = \operatorname*{argmin}_{\hat{f}} \sigma^2 \|\hat{f}\|_H^2 + \|\hat{f}_X - f_X\|_2^2$$

observe: $y = f_X + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I)$

$$y \sim N(0, K_{XX} + \sigma^2 I) \qquad \Sigma = K_{XX} + \sigma^2 I$$

$$\max_\sigma \mathcal{L}(\sigma \mid y) = p_\sigma(y) = \frac{\exp\left(-\frac{1}{2} y^T \Sigma^{-1} y\right)}{\sqrt{(2\pi)^d \det(\Sigma)}}$$

$$\log \mathcal{L} \Rightarrow \underbrace{-\frac{1}{2} y^T \Sigma^{-1} y}_{\text{fidelity}} - \frac{d}{2}\log(2\pi) - \underbrace{\frac{1}{2}\log\det(\Sigma)}_{\text{complexity}}$$

$$\frac{\partial}{\partial \sigma} \log\det(\Sigma) = \operatorname{trace}(\Sigma^{-1} 2\sigma I)$$

$$\frac{\partial}{\partial \sigma} y^T \Sigma^{-1} y = -y^T \Sigma^{-1} (2\sigma I) \Sigma^{-1} y$$

$$\Sigma = K_{xx} + \sigma^2 I$$

Annoying computations!!