April 23, 2019

function approx.

true $f: \mathbb{R}^n \to \mathbb{R}$

observe $f(x)$ $x \in X$, f_X

$\hat{f}: \mathbb{R}^n \to \mathbb{R}$

Today: kernel methods

$k(x, y)$ = similarity between x & y $\quad x, y \in \mathbb{R}^n$

exp: $k(x,y) = s^2 \exp\left(-\frac{\|x-y\|^2}{l}\right)$

cubic splines ($n=1$):

$$k(x, y) = |x - y|^3$$

---

(1) Feature maps
kernel is inner product
b/w feature vectors

(2) Quadratic forms on function spaces

(3) Covariance of random process is given by a kernel
$\Rightarrow$ Gaussian Processes

kernel on some set S

$$k: S \times S \Rightarrow \mathbb{R}$$

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

linear regression

$$f(x) \simeq \hat{f}(x) = x^T c$$

$$\min_c \| X^T c - f_X \|_2^2$$

rows are data points $x \in X$

true $f(x)$ at $x \in X$

$$\hat{f}(x) = \Psi(x)^T c$$

$$\approx \sum_{j=1}^{N} c_j \psi_j(x)$$

ψ is the feature map

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

$$\Psi_{ij} = \psi_i(x_j)$$

if N (# of features) $\leq$ m (# of data points)

$\Rightarrow$ standard OLS

$N > m$

$$\boxed{\Psi^T} \; c = f_x$$

$$\min \; \|c\|_2^2$$

$$\text{s.t.} \quad \Psi^T c = f_x$$

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

$$\hat{f}(z) = \Psi(z)^T c$$

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

$$k(x, y) = \Psi(x)^T \Psi(y)$$

$$\boxed{\hat{f}(z) = k_{zX} K_{XX}^{-1} f_X}$$

$$[K_{xx}]_{ij} = k(x_i, x_j)$$

$$[k_{zx}]_i = k(z, x_i)$$

$$x_i, x_j \in X$$

Only need $k(x, y) = \psi(x)^T \psi(y)$

called "kernel trick"

If we start from kernel,

e.g. $k(x, y) = \exp(-\|x-y\|^2)$,

is there an associated feature map?

Yes! (Mercer's theorem)

if K_{XX} is positive semi-def. for any finite set X, then there is an associated feature map

---

$$\hat{f}(x) = \sum_i c_i \psi_i(x)$$

$$\min \|c\|^2$$

$$\text{s.t. } \underline{\Psi}^T c = f_X$$

implicitly, defined an inner product on functions with orthonormal basis $\{\psi_1, \ldots, \psi_N\}$

call this $\mathcal{H}$

$$\min \|\hat{f}\|_{\mathcal{H}}^2$$

$$\text{s.t.} \quad \hat{f}_x = f_x$$

point evaluation

$$\hat{f}(z)$$

$$k_z(x) \triangleq \sum_i \psi_i(z)\psi_i(x)$$

$$\langle \hat{f}, k_z \rangle_{\mathcal{H}}$$

$$= \left\langle \sum_i c_i \psi_i, \sum_j \psi_j(z)\psi_j \right\rangle$$

$$= \sum_{ij} c_i \psi_j(z) \langle \psi_i, \psi_j \rangle_{\mathcal{H}}$$

$$(\langle \psi_i, \psi_j \rangle_{\mathcal{H}} = \delta_{ij})$$

$$\simeq \sum c_i \psi_i(z)$$

$$\simeq \hat{f}(z)$$

More generally,

$$g(z) = \langle g, k_z \rangle_{\mathcal{H}}$$

$\Rightarrow \mathcal{H}$ is a reproducing kernel Hilbert space (RKHS)

Can we get the inner product w/o explicitly computing the feature map?

Main idea:

$$\left\langle \sum_i c_i k_{x_i}, \sum_j d_j k_{x_j} \right\rangle_H$$

$$k_{x_i}(z) = k(x_i, z)$$

$$= \sum_{ij} c_i d_j \langle k_{x_i}, k_{x_j} \rangle_H$$

$$= c^T K_{xx} d$$

(need to extend to infinite sums)

Error analysis

Approx $f \in \mathcal{H}$ using X

$$\min \|\hat{f}\|_{\mathcal{H}}^2$$

$$\text{s.t. } \hat{f}_X = f_X$$

$$|\hat{f}(y) - f(y)| < ?$$

for $y \notin X$

Define $\tilde{f}$ by

$$\min \|\tilde{f}\|_{\mathcal{H}}^2$$

$$\text{s.t. } \tilde{f}_X = f_X$$

$$\tilde{f}(y) = f(y)$$

$$\|\hat{f}\|_H^2 \le \|\tilde{f}\|_H^2 \le \|f\|_H^2$$

$$X' \triangleq X \cup \{y\}$$

$$e \triangleq \tilde{f} - \hat{f}$$

$$e_{X'} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ e(y) \end{bmatrix} \qquad e(y) = \tilde{f}(y) - \hat{f}(y)$$

$$\langle e, \hat{f} \rangle_H = 0$$

$$\|e\|_H^2 + \|\hat{f}\|_H^2 = \|\tilde{f}\|_H^2$$

$$\|e\|_H^2 \le \|f\|_H^2 - \|\hat{f}\|_H^2$$

$$\|e\|_H^2 = d^T K_{X'X'} d$$

$$\left(d = K^{-1}_{x'x'} e_{x'}\right)$$ (can show)

$$= e^T_{x'} K^{-1}_{x'x'} K_{x'x'} K^{-1}_{x'x'} e_{x'}$$

$$= e^T_{x'} K^{-1}_{x'x'} e_{x'}$$

$$= \left(K^{-1}_{x'x'}\right)_{yy} e(y)^2$$

Putting the algebra together...

$$e(y)^2 \leq \frac{1}{\left(K^{-1}_{x'x'}\right)_{yy}} \left(\|f\|^2_H - \|\hat{f}\|^2_H\right)$$

$$k(y,y) - k_{yx} K^{-1}_{xx} k_{xy}$$

no x'

$k_{yx}(x_i) = k(y, x_i)$

$x_i \in X$ "

$k_{xy}(x_i) = k(x_i, y)$

Gaussian Processes

GP on a set S ($\mathbb{R}^d$)

with mean field m

covariance kernel k

$f \sim GP(m, k)$

for any $X \subset S$ finite

$$f_X \sim N(\mu_X, K_{XX})$$

multivariate normals

$$Y \sim N(0, K)$$

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}$$

Theorem

$$Y_2 | Y_1 = y_1 \sim N(K_{12}K_{22}^{-1} y_1, K_{11} - K_{12}K_{22}^{-1}K_{21})$$

idea with GPs:

- observe set of points X

- condition on X
- get a new GP

Observe f_X

$$\hat{f} \sim GP(\hat{\mu}, \hat{k})$$

$$\hat{\mu}(z) = k_{zX} K_{XX}^{-1} f_X + \mu$$

↳ same as least squares featurization

$$\hat{k}(y, z) = k(y, z) - k_{yX} K_{XX}^{-1} k_{Xz}$$

$$\left.\begin{aligned} k_{yX}(x) &= k(y, x) \\ k_{zX}(x) &= k(z, x) \end{aligned}\right\} x \in X$$

$$\hat{k}(y, y) = k(y, y) - k_{yX} K_{XX}^{-1} k_{Xy}$$

↳ same as in error bound

Two ways to think about error in kernel methods for function approx.

① $\min \|\hat{f}\|_{\mathcal{H}}^2$

s.t. $\hat{f}_X = f_X$

worst-case error bound

② Bayesian inference

Assume $f \sim GP(\mu, k)$ (prior)

Observe X, f_X

Update to get $\hat{f}$ (posterior)

MLE with GPs

k parameterized by some θ

$$\mathcal{L}(\theta) = \log(p(X;\theta))$$

$$p(y) = \frac{1}{\sqrt{\det(2\pi K)}} \cdot$$

$$\exp\left(-\frac{1}{2}(y-m)K^{-1}(y-m)\right)$$

$$\log p(y)$$

$$= -\frac{1}{2}(y-m)^T K^{-1}(y-m) \quad \text{fidelity}$$

$$-\frac{1}{2}\log\det K \quad \text{complexity}$$

$$-\frac{n}{2}\log(2\pi) \quad \text{constant}$$

$$\frac{d}{d\theta} \log(p(y))$$

$$\frac{d}{d\theta} \left[\text{fidelity} \right]$$

$$= \frac{1}{2}(y-\mu)^T c^T \left(\frac{d}{d\theta} K\right) c$$

$$c = K^{-1}(y-\mu)$$

requires solving linear system

$$\frac{d}{d\theta} \left[\text{complexity} \right]$$

$$= \text{tr}\left(K^{-1} \frac{d}{d\theta} K\right)$$

also expensive

Efficient optimization

active area of research

(at Cornell)