Reproducing Kernel Hilbert Spaces

John Duchi

Motivation

Can always break down risk in terms of

$$L(\widehat{h}) - \inf_{h} L(h) = \underbrace{L(\widehat{h}) - \inf_{h \in \mathcal{H}} L(h)}_{\text{estimation error}} + \underbrace{\inf_{h \in \mathcal{H}} L(h) - \inf_{h} L(h)}_{\text{approximation error}}$$

- Generalization and other convergence guarantees get at estimation error (via complexity bounds on $\mathcal{H}$, characteristics of risk L and loss ℓ, etc.)
- Approximation error requires understanding how expressive function class is

Motivation: nonlinear features

- Instead of using

$$\langle \theta, x \rangle$$

use

$$\langle \theta, \phi(x) \rangle$$

Example (Polynomials)

For $x \in \mathbb{R}$, use $\phi(x) = [1 \; x \; x^2 \; \cdots \; x^d]^T \in \mathbb{R}^{d+1}$

Example (Strings)

For x a string, let

$$\phi(x) = [\text{count of } a \in x]_{a \in \mathcal{S}}$$

Can we cut down on computation and control complexities?

Data representations

Theorem (Representer theorem)

Let

$$\widehat{L}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(\langle \theta, \phi(x_i)\rangle, y_i) + \varphi(\|\theta\|_2)$$

for any loss ℓ, non-decreasing regularizer $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$. Then w.l.o.g. any minimizer of $\widehat{L}_n$ can be taken of the form

$$\widehat{\theta} = \sum_{i=1}^{n} \alpha_i \phi(x_i)$$

- Extends to populatin ($n = \infty$) case too
- Key takeaway: future predictions are

$$\langle \theta, \phi(x)\rangle = \sum_{i=1}^{n} \alpha_i \langle \phi(x_i), \phi(x)\rangle$$

Polynomial features

For $x \in \mathbb{R}^k$, let

$$\phi(x) = \begin{bmatrix} 1 \\ \sqrt{2}x_1 \\ \vdots \\ \sqrt{2}x_k \\ [x_i x_j]_{i,j=1}^k \end{bmatrix} \in \mathbb{R}^{1+k+k^2}$$

Then

$$\phi(x)^T \phi(z) = (1 + x^T z)^2$$

More generally: for degree d,

$$\langle \phi(x), \phi(z) \rangle = (1 + x^T z)^d$$

Kernels: definitions

Definition (Positive definite function)

A function $\mathsf{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is *positive definite* if it is symmetric and for all $n \in \mathbb{N}$ and $x_1, \ldots, x_n \in \mathcal{X}$, the Gram matrix

$$K = \begin{bmatrix} \mathsf{k}(x_1, x_1) & \cdots & \mathsf{k}(x_1, x_n) \\ \vdots & \ddots & \vdots \\ \mathsf{k}(x_n, x_1) & \cdots & \mathsf{k}(x_n, x_n) \end{bmatrix}$$

is positive semidefinite, i.e. $\alpha^T K \alpha \geq 0$ for all $\alpha \in \mathbb{R}^n$.

A function k is a *kernel* if and only if it is a positive semidefinite function

Examples

- Inner products: $\mathsf{k}(x, z) = x^T z = \sum_{j=1}^d x_j z_j$
- Polynomials: $\mathsf{k}(x, z) = (1 + x^T z)^k$
- Min-kernel: $\mathsf{k}(x, z) = \min\{x, z\}$
- Sequence mis-match kernel: $\mathcal{X} = \Sigma^*$ is alphabet of all sequences over Σ
  - String $u \sqsubset x$ (u is a subsequence of x) if $\mathsf{len}(u) = k$ and there are $i_1, \ldots, i_k$

$$u = x_{i_1} x_{i_2} \cdots x_{i_k} = x(\mathbf{i}) \text{ for } \mathbf{i} = (i_1, \ldots, i_k)$$

  - Kernel:

$$\mathsf{k}(x, z) = \sum_{u \in \Sigma^*} \sum_{\mathbf{i}, \mathbf{j} : x(\mathbf{i}) = z(\mathbf{j}) = u} \lambda^{\mathrm{card}(\mathbf{i}) + \mathrm{card}(\mathbf{j})}$$

Construction of kernels

- Any product $\mathsf{k}(x, z) = f(x)f(z)$ is a kernel

$$K = uu^T \text{ for } u = [f(x_1) \ \cdots \ f(x_n)]$$

- Any sum: $\mathsf{k}(x, z) = \mathsf{k}_1(x, z) + \mathsf{k}_2(x, z)$ because $K = K_1 + K_2 \succeq 0$

Product kernels

For $A \in \mathbb{R}^{n \times n}, B \in \mathbb{R}^{m \times m}$ symmetric with $A = \sum_{i=1}^{n} \lambda_i u_i u_i^T$ and $B = \sum_{i=1}^{m} \nu_i v_i v_i^T$, Kronecker product

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{n1}B & \cdots & a_{nn}B \end{bmatrix}$$

has spectral decomposition

$$A \otimes B = \sum_{i=1}^{n} \sum_{j=1}^{m} \nu_i \lambda_j (u_i \otimes v_j)(u_i \otimes v_j)^T$$

- Product kernel $\mathsf{k}(x, z) = \mathsf{k}_1(x, z) \cdot \mathsf{k}_2(x, z)$, $K = K_1 \odot K_2$ (Hadamard/elementwise product) is sub-matrix of Kronecker

Examples

- Inner products: $\mathsf{k}(x, z) = x^T z = \sum_{j=1}^{d} x_j z_j$
- Polynomials: $\mathsf{k}(x, z) = (1 + x^T z)^k$
- Gaussian-like kernel:

$$\mathsf{k}(x, z) = \exp(\langle x, z \rangle) = \sum_{k=0}^{\infty} \frac{\langle x, z \rangle^k}{k!}$$

The three views of kernel methods

Hilbert spaces

Note: we are lazy and usually work with *real* Hilbert spaces

Definition (Hilbert space)

A vector space $\mathcal{H}$ is a *Hilbert space* if it is a complete inner product space.

Definition (Inner product)

A bi-linear mapping $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is an *inner product* if it satisfies

- Symmetry: $\langle f, g \rangle = \langle g, f \rangle$
- Linearity: $\langle \alpha f_1 + \beta f_2, g \rangle = \alpha \langle f_1, g \rangle + \beta \langle f_2, g \rangle$
- Positive definiteness: $\langle f, f \rangle \geq 0$ and $\langle f, f \rangle = 0$ if and only if $f = 0$

This gives Euclidean norm

$$\|f\|_{\mathcal{H}} := \sqrt{\langle f, f \rangle}.$$

Examples

1. Euclidean space $\mathbb{R}^d$, $\langle u, v \rangle = \sum_{j=1}^d u_j v_j$
2. Square-summable sequences:

$$\ell_2 := \left\{ u \in \mathbb{R}^{\mathbb{N}} \mid \sum_{j=1}^{\infty} u_j^2 < \infty \right\}$$

   with $\langle u, v \rangle = \sum_{j=1}^{\infty} u_j v_j$
3. Square integrable functions against *any* probability distribution p:

$$\langle f, g \rangle := \int f(x) g(x) p(x) dx$$

   or, more generally,

$$\langle f, g \rangle := \mathbb{E}_P[f(X) g(X)]$$

Fun example

Let

$$\mathsf{k}(x, z) = \exp\left(-\frac{\|x - z\|_2^2}{2\sigma^2}\right)$$

Feature maps and kernels

Definition (Feature mapping)

Given a Hilbert space $\mathcal{H}$, a *feature mapping* $\phi : \mathcal{X} \to \mathcal{H}$, $\phi(x) \in \mathcal{H}$

Theorem

Any feature mapping defines a valid kernel.

Reproducing kernel Hilbert spaces

We want to be sure we can *evaluate* or prediction function $f(x)$, where $f \in \mathcal{H}$ for some $\mathcal{H}$

Example

Hilbert space $L^2([0,1]) = \{f : [0,1] \to \mathbb{R} \mid \|f\|_2 < \infty\}$. If $f(x) = g(x)$ almost everywhere, then $\|f - g\|_2 = 0$

Definition

For Hilbert space $\mathcal{H}$ a linear functional $L : \mathcal{H} \to \mathbb{R}$ is *bounded* if

$$|L(f)| \leq M \|f\|_{\mathcal{H}} \quad \text{for all } f \in \mathcal{H}$$

Evaluation functionals

For Hilbert space $\mathcal{H}$ of $f : \mathcal{X} \to \mathbb{R}$, the evaluation functional

$$L_x(f) := f(x).$$

Example

For $\mathcal{X} = \mathbb{R}^d$, $\mathcal{H} = \{f_c \mid c \in \mathbb{R}^d\}$ where $f_c(x) = \langle c, x \rangle$, then $L_x(f_c) = \langle c, x \rangle$

Example (Unbounded evaluation)

Let $\mathcal{H} = L^2([0, 1])$, then $L_x(f) = f(x)$ is *unbounded*.

Reproducing Kernel Hilbert Spaces

Definition (RKHS)

A *reproducing kernel Hilbert space* is any Hilbert space $\mathcal{H}$ for which the evaluation functional L_x is bounded for each $x \in \mathcal{X}$

RKHSs define kernels

Theorem

Let $\mathcal{H}$ be an RKHS of $f : \mathcal{X} \to \mathbb{R}$. Then there is a unique $\mathsf{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *associated to $\mathcal{H}$ with*

$$\mathsf{k}(x, \cdot) \in \mathcal{H}$$

where the k *is reproducing for $\mathcal{H}$: for all $f \in \mathcal{H}$*

$$\langle f, \mathsf{k}(x, \cdot) \rangle = f(x)$$

Proof (continued)

Kernels define RKHSs

Theorem (Moore-Aronszajn)

Let $\mathsf{k} : \mathcal{X} \to \mathcal{X} \to \mathbb{R}$*. Then there is a unique RKHS* $\mathcal{H}$ *with reproducing kernel* k

Proof: Let $\mathcal{H}_0$ be all linear combinations $f(x) = \sum_{i=1}^{n} \alpha_i \mathsf{k}(x, x_i)$

Kernels define RKHSs: inner products

Kernels define RKHSs: completeness

Reading and bibliography

1. N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3): 337–404, May 1950
2. A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, 2004
3. G. Wahba. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, Philadelphia, 1990
4. N. Cristianini and J. Shawe-Taylor. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004