STA 250: Theoretical Foundations for Machine Learning

Lecture 4: Kernel Methods

Dogyoon Song

Spring 2025, UC Davis

Motivation

So far, we discussed how to control the excess risk

$$R(f_{\hat{\theta}}) - R^* = \underbrace{\left(R(f_{\hat{\theta}}) - R(f_{\theta^*})\right)}_{\text{estimation error}} + \underbrace{\left(R(f_{\theta^*}) - R^*\right)}_{\text{approximation error}}$$

However...

- Our discussion focused exclusively on estimation error
- We ended up with a Rademacher complexity upper bound for linear models

Kernels provide a way to represent a class of functions $\mathcal{F}_\varphi = \{\langle \theta, \varphi(x) \rangle\}$ that can represent non-linear prediction functions (via non-linear feature map φ) while enjoying the uniform convergence bound akin to linear models

Agenda

- Kernels
- Three viewpoints for kernel methods
- Learning with kernels: Examples
- Additional topics related to kernels

Kernels

Definition

A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a (positive semidefinite) kernel if for every finite set $x_1, \ldots, x_n \in \mathcal{X}$, the associated *kernel matrix* $K \in \mathbb{R}^{n \times n}$ such that $K_{ij} = k(x_i, x_j)$ is positive semidefinite (PSD).

Examples:

- Linear kernel: $k(x, x') = \langle x, x' \rangle$
- Polynomial kernel: $k(x, x') = (1 + \langle x, x' \rangle)^p$
- Gaussian kernel: $k(x, x') = \exp\left(-\frac{\|x - x'\|_2^2}{2\sigma^2}\right)$

How to check?

- Check $K \succeq 0$ for all instances of kernel matrices
- "Kernel calculus": (a) $k(x, x') = f(x)f(x')$ is PSD, $\forall f$; (2) $K_1, K_2 \succeq 0 \implies K_1 + K_2 \succeq 0$ and $K_1 \circ K_2 \succeq 0$

Three viewpoints for kernel methods

Feature map $\varphi : \mathcal{X} \to \mathbb{R}^d$ (or $\mathcal{H}$)

- Properties of a single data point

Kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

- Similarity between data points

Reproducing kernel Hilbert space (RKHS) $\mathcal{H} : (\{f : \mathcal{X} \to \mathbb{R}\}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$

- Prediction function $f : \mathcal{X} \to \mathbb{R}$

Indeed, these three viewpoints are closely related and equivalent

Feature $\leftrightarrow$ kernel

Proposition

If $\varphi : \mathcal{X} \to \mathcal{H}$ is a feature map, then $k(x, x') = \langle \varphi(x), \varphi(x') \rangle$ is a kernel.

Proposition

For every kernel k, $\exists \mathcal{H}$ and $\varphi : \mathcal{X} \to \mathcal{H}$ such that $k(x, x') = \langle \varphi(x), \varphi(x') \rangle$.

Reproducing kernel Hilbert space

Definition

A Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$ is a RKHS if the evaluation functional is bounded for all $x \in \mathcal{X}$.

Proposition

Every RKHS $\mathcal{H}$ defines a unique kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ called the reproducing kernel of $\mathcal{H}$.

Proposition (More-Aronszajn)

For every kernel k, there exists a unique RKHS $\mathcal{H}$ with reproducing kernel k.

Representer theorem

Let

$$f^* \in \arg\min_{f \in \mathcal{H}} \left\{ L(\{(x_i, y_i, f(x_i)) : i \in [n]\}) + Q(\|f\|_{\mathcal{H}}^2) \right\}$$

where $L : (\mathcal{X} \times \mathcal{Y} \times \mathbb{R})^* \to \mathbb{R}$ is an arbitrary function and $Q : \mathbb{R} \to \mathbb{R}$ is strictly monotone increasing

Theorem

Let f^* be defined as above. Then $f^* \in \operatorname{span}(\{k(x_i, \cdot) : i \in [n]\})$.