CMPUT 654 Fa 23: Theoretical Foundations of Machine Learning **Fall 2023**

Lecture 20: Kernel Methods (November 21)

Lecturer: Csaba Szepesvári *Scribes: Kushagra Chandak*

Motivation. Suppose we have a feature map

$$\psi : \mathcal{X} \to \mathbb{R}^d$$

which is used to make predictions

$$f_w(x) = \langle w, \psi(x) \rangle \,,$$

where $w \in \mathbb{R}^d$. Note that if d is huge then computing the prediction $f_w(x)$ is expensive (linear in d). Can we do this efficiently? Ambitiously, can we replace $\mathbb{R}^d$ with some *Hilbert space* $\mathcal{W}$? Recall that a Hilbert space is a complete inner product space. The endowed inner product is a bilinear function $\langle \cdot, \cdot \rangle : \mathcal{W}^2 \to \mathbb{R}$ satisfying the following properties for any $u, v, u_1, u_2 \in \mathcal{W}$.

1. (Symmetric) $\langle u, v \rangle = \langle v, u \rangle$.
2. (Linear in both arguments) $\langle u_1 + u_2, v \rangle = \langle u_1, v \rangle + \langle u_2, v \rangle$ and $\langle \lambda u, v \rangle = \lambda \langle u, v \rangle$.
3. (Positive) $\langle u, u \rangle = 0$ iff $u = 0$

Note that the inner product induces the norm $\|u\|^2 = \langle u, u \rangle$. Some examples of Hilbert spaces:

1. $\mathcal{W} = \mathbb{R}^d$ with the inner product of the form $\langle u, v \rangle = u^\top Q v$ where Q is PSD ($Q \succeq 0$).
2. $\mathcal{W} = \ell_2(\mathbb{R}^{\mathbb{N}})$, where $\ell_2(\mathbb{R}^{\mathbb{N}}) \subset \mathbb{R}^N$ such that $\|u\|^2 = \sum_{i=1}^{\infty} u_i^2 < \infty$ for all $u \in \mathcal{W}$, with the inner product $\sum_{i=1}^{\infty} u_i v_i$.

With a Hilbert space structure, we can do computations like calculating ERM efficiently. We have input space $\mathcal{X}$ and output space $\mathcal{Y}$. The prediction we want to compute efficiently is $f_w(x) = \langle w, \psi(x) \rangle$. We also have a loss function $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$. For data $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathcal{Y}$, the regularized loss is defined as

$$Q(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(f_w(x_i), y_i) + \frac{\lambda}{2} \|w\|^2 \,.$$

We introduce another map, the *kernel*, $k : \mathcal{X}^2 \to \mathbb{R}$, using which we can write the ERM solution.

$$k(u, v) = \langle \psi(u), \psi(v) \rangle \,.$$

Now imagining that the Hilbert space $\mathcal{W}$ is $\mathbb{R}^d$, we can find the ERM solution:

$$0 = Q'(w) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial p} \ell(f_w(x_i), y_i) \psi(x_i) + \lambda w \,,$$

or

$$w = -\frac{1}{\lambda n}\sum_{i=1}^{n}\frac{\partial}{\partial p}\ell(f_w(x_i), y_i)\psi(x_i) = \sum_{i=1}^{n}\alpha_i\psi(x_i)\,.$$

Therefore f_w can be computed as

$$f_w(x) = \langle w, \psi(x)\rangle = \sum_{i=1}^{n}\alpha_i k(x_i, x) = \tilde{f}_\alpha(x), \quad \alpha \in \mathbb{R}^n\,.$$

Note that the penalty term $\|w\|^2$ can be written as

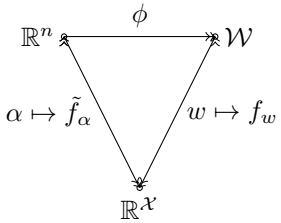
$$\|w\|^2 = \sum_{i,j=1}^{n}\alpha_i\alpha_j k(x_i, x_j) = \alpha^\top K\alpha\,,$$

where K is an $n \times n$ matrix composed from the kernel: $K = (k(x_i, x_j)_{i,j=1}^n$.

We can also optimize the loss in the α ($\mathbb{R}^n$) space:

$$\tilde{Q}(\alpha) = \frac{1}{n}\sum_{i=1}^{n}\ell(\tilde{f}_\alpha(x_i), y_i) + \frac{\lambda}{2}\alpha^\top K\alpha$$

We can go between the α space ($\mathbb{R}^n$) and the $\mathcal{W}$ space. Going from $\mathbb{R}^n$ to $\mathcal{W}$ is simpler, and can be done using the map $\phi : \mathbb{R}^n \to \mathcal{W}$ defined as $\alpha \mapsto \sum_{i=1}^n \alpha_i\psi(x_i)$.



We know that $\phi(\mathbb{R}^n) \subset \mathcal{W}$. So if we can show that $\arg\min_{w\in\mathcal{W}} Q(w) \subset \phi(\mathbb{R}^n)$ then minimizing $Q(w)$ would be same as minimizing $\tilde{Q}(\alpha)$.

Switching to the "α" representation is called the *kernel trick*.

Definition 20.1 (Positive definite kernel[1])**.** Let $k : \mathcal{X}^2 \to \mathbb{R}$ be symmetric. k is positive definite if for all $n \in \mathbb{N}$, $x_{1:n} \in \mathcal{X}^n$ and $\alpha \in \mathbb{R}^n$, $\sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j k(x_i, x_j) \geq 0$.

Let k be a symmetric positive definite kernel. Let $\mathcal{H}_0 \subseteq \mathbb{R}^{\mathcal{X}}$ be defined by

$$\mathcal{H}_0 = \left\{x \mapsto \sum_{i=1}^{n}\alpha_i k(x, x_i) : n \in \mathbb{N}, \alpha \in \mathbb{R}^n\right\}.$$

Suppose we define a function on $\mathcal{H}_0$ (which we will claim to be an inner product)

$$\langle\sum_{i=1}^{n}\alpha_i k(x_i, \cdot), \sum_{j=1}^{n}\beta_j k(x_j, \cdot)\rangle = \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\beta_j k(x_i, x_j)\,.$$

Claim 20.2. *$\mathcal{H}_0$ is a pre-Hilbert space (no completeness).*

Theorem 20.3. *For every symmetric positive definite k, $\exists!(\mathcal{H}, \langle\cdot,\cdot\rangle \subseteq \mathbb{R}^{\mathcal{X}}$ Hilbert space such that $\mathcal{H}_0 \subseteq \mathcal{H}$ is dense. For any function $\sum\alpha_i k(x_i, \cdot), \beta_j k(x_j, \cdot) \in \mathcal{H}_0$, $\langle\alpha_i k(x_i, \cdot), \beta_j k(x_j, \cdot)\rangle = \sum_{i,j}\alpha_i\beta_j k(x_i, x_j)$.*

The space $\mathcal{H}$ is called a *reproducing kernel Hilbert space* (RKHS). A RKHS has the *reproducing kernel* property which follows from the construction:

$$f(x) = \langle f, k(x, \cdot)\rangle\,.$$

[1]Technically, positive semi-definite