STAT538A: Statistical Learning: the Way of the Kernel	Winter 2020
Lecture 1	
Lecturer: Zaid Harchaoui	*Scribe: Lang Liu*

1 Kernel and Hilbert Space

We first give some important definitions.

Definition 1 (Empirical Mean Vector)**.** *If $x_1, \dots, x_n \in \mathbb{R}^d$, the* empirical mean vector *is defined as:*

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

And the mean vector *is defined as:*

$$\mu = \mathbb{E}(X).$$

Definition 2 (Cauchy Sequence)**.** *Given a metric space $(\mathcal{X}, d)$, a sequence $(x_k)_{k\geq 1} \subset \mathcal{X}$ is called a Cauchy sequence, if for every $\varepsilon > 0$ there exists a positive integer N such that the distance $d(x_m, x_n) < \varepsilon$ for all $m, n > N$.*

Definition 3 (Hilbert Space)**.** *Let $\mathcal{H}$ be a vector space equipped with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$. Let $\|\cdot\|_{\mathcal{H}} : \mathcal{H} \to R$ be the associated norm defined as*

$$\|f\|_{\mathcal{H}} = \sqrt{\langle f, f \rangle_{\mathcal{H}}}, \quad f \in \mathcal{H}.$$

The vector space $\mathcal{H}$ is a Hilbert space if it is complete with respect to this norm, i.e., every Cauchy sequence in $\mathcal{H}$ has a limit in $\mathcal{H}$.

Hilbert space is a generalization of Euclidean space which may be infinite dimensional. Indeed, every Euclidean space is a Hilbert space. This generalization allows us to work with spaces of functions. The next definition identifies a particular type of function spaces that we will use throughout this course.

Definition 4 (Reproducing Kernel Hilbert Space)**.** *Let $\mathcal{X}$ be a set and $\mathcal{H}$ be a Hilbert space of functions defined on $\mathcal{X}$. We call $\mathcal{H}$ a Reproducing Kernel Hilbert Space (RKHS) if there exists a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that, for every $x \in \mathcal{X}$, $k(x, \cdot) \in \mathcal{H}$ and the* reproducing property

$$\langle f, k(x, \cdot) \rangle_{\mathcal{H}} = f(x)$$

holds for every $f \in \mathcal{H}$. We refer to k as a reproducing kernel for the space $\mathcal{H}$, and call $\phi(x) := k(x, \cdot)$ the feature map from $\mathcal{X}$ to $\mathcal{H}$.

Definition 5 (Empirical Mean Element)**.** *Assume $\phi : \mathcal{X} \to \mathcal{H}$ is the feature map in an RKHS $\mathcal{H}$, then the* empirical mean element *is defined as*

$$\hat{\mu}_{\mathcal{H}} = \frac{1}{n}\sum_{i=1}^{n} \phi(x_i).$$

Note that $\mathcal{H}$ is a vector space, so the empirical mean element $\hat{\mu}_{\mathcal{H}} \in \mathcal{H}$.

Remark It follows from the reproducing property that $\forall f \in \mathcal{H}$,

$$\langle \hat{\mu}_{\mathcal{H}}, f\rangle_{\mathcal{H}} = \langle \frac{1}{n}\sum_{i=1}^{n}\phi(x_i), f\rangle_{\mathcal{H}} = \frac{1}{n}\sum_{i=1}^{n} f(x_i),$$

which is computationally tractable.

Example 6. *Here are two examples of reproducing kernels*

- Linear kernel*:* $k(x, y) = x^\top y$ *for* $x, y \in \mathbb{R}^d$*.*
- Gaussian kernel*:* $k(x, y) = exp(-\frac{\|x-y\|_2^2}{\sigma^2})$ *for* $x, y \in \mathbb{R}^d$*, where* $\sigma > 0$*.*
- Bag of vectors kernel*: let* $\mathcal{X} := \{\{v_1, \dots, v_p\} : v_i \in \mathbb{R}^d \text{ for } i \in [p] := \{1, \dots, p\}, p \in \mathbb{N}\}$*. The bag of vectors kernel is defined by:*

$$k(\{u_1, \dots, u_p\}, \{v_1, \dots, v_q\}) = \sum_{i=1}^{p}\sum_{j=1}^{q} u_i^T v_j.$$

Definition 7 (Positive Semi-Definite Kernel)**.** *A function* $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *is called a positive semi-definite (PSD) kernel if it satisfies:*

- *Symmetry:* $\forall x, y \in \mathcal{X}, k(x, y) = k(y, x)$*.*
- *Positive semi-definiteness: for every* $n \in \mathbb{N}, \alpha_1, \dots, \alpha_n \in \mathbb{R}$*, and* $x_1, \dots, x_n \in \mathcal{X}$*,*

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j k(x_i, x_j) \geq 0.$$

It is said to be strictly positive definite if the above equality implies $\alpha_1 = \alpha_2 = \dots = \alpha_n = 0$*.*

Example 8. *Let* $\mathcal{X} = \mathbb{R}^d$*, the linear kernel* $k(x, y) = x^\top y$ *is a PSD kernel.*

- *It is symmetric since* $x^\top y = (x^\top y)^\top = (y)^\top x$*.*
- *PSD follows from*

$$\begin{aligned}\sum_i\sum_j \alpha_i\alpha_j k(x_i, x_j) &= \sum_i\sum_j \alpha_i\alpha_j x_i^T x_j \\ &= \left(\sum_i \alpha_i x_i\right)^\top \left(\sum_i \alpha_i x_i\right) \\ &\geq 0.\end{aligned}$$

An argument analogous to Example 8 shows that the reproducing kernel of an RHKS is PSD.

Proposition 9. *Let* $\mathcal{H}$ *be an RKHS of functions defined on* $\mathcal{X}$ *with reproducing kernel* $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$*, then* k *is PSD.*

Here listed below is a table of common kernel functions:

Kernel Function	$k(\mathbf{x}, \mathbf{y})$	**Domain** $\mathcal{X}$
Dirac	$\mathbb{1}_{\mathbf{x}=\mathbf{y}}$	$\{1, 2, \ldots, m\}$
Discrete	$\sum_{s \in \mathcal{X}} w_s \#_s(\mathbf{x}) \#_s(\mathbf{y})$ with $w_s > 0$ for all s	$\{s_1, s_2, \ldots, s_m\}$
Linear	$\langle \mathbf{x}, \mathbf{y} \rangle$	$\mathbb{R}^d$
Polynomial	$(\langle \mathbf{x}, \mathbf{y} \rangle + c)^p$	$\mathbb{R}^d$
Gaussian	$\exp(-\sigma \|\mathbf{x} - \mathbf{y}\|_2^2),\ \sigma > 0$	$\mathbb{R}^d$
Laplacian	$\exp(-\sigma \|\mathbf{x} - \mathbf{y}\|_1),\ \sigma > 0$	$\mathbb{R}^d$
Rational quadratic	$(\|\mathbf{x} - \mathbf{y}\|_2^2 + c^2)^{-\beta},\ \beta > 0, c > 0$	$\mathbb{R}^d$
Exponential	$\exp(\sigma \langle \mathbf{x}, \mathbf{y} \rangle),\ \sigma > 0$	compact sets of $\mathbb{R}^d$
Matérn	$\frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}\|\mathbf{x}-\mathbf{x}'\|_2}{\sigma} \right) K_\nu \left(\frac{\sqrt{2\nu}\|\mathbf{x}-\mathbf{x}'\|_2}{\sigma} \right)$	$\mathbb{R}^d$
Poisson	$1/(1 - 2\alpha \cos(\mathbf{x} - \mathbf{y}) + \alpha^2),\ 0 < \alpha < 1$	$([0, 2\pi), +)$

Figure 1: Well-known Kernel Functions.

Next theorem establishes the correspondence between Reproducing Kernel Hilbert Spaces and positive semi-definite kernels.

Theorem 10 (Aronszajn, 1950)**.** *Let $\mathcal{X}$ be a metric space and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a PSD kernel, there exists a unique Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ of functions on $\mathcal{X}$ satisfying the following conditions:*

(i) for all $x \in \mathcal{X}, \phi(x) := k(x, \cdot) \in \mathcal{H}$,

(ii) the span of the set $\{\phi(x) | x \in \mathcal{X}\}$ is dense in $\mathcal{H}$, and

(iii) for all $f \in \mathcal{H}$ and $x \in \mathcal{X}, f(x) = \langle \phi(x), f \rangle_{\mathcal{H}}$.

In particular, $\mathcal{H}$ is an RKHS.