Lecture Notes 4: Reproducing Kernel

Professor: Zhihua Zhang *Scribe: Tianyuan Liu, Zhiming Ding*

1 Reproducing Kernel

1.1 Basic Definitions (review)

Definition 1.1 *For $\mathcal{X} \subseteq \mathbb{R}^p$, a kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called positive definite (p.d.)* ***iff*** *$\forall n \in \mathcal{N}, \forall \{x_1, x_2, \ldots, x_n\} \subseteq \mathcal{X}$, $\mathbf{K} = [K(x_i, x_j)]_{n\times n}$ is positive semi-definite (p.s.d.).*

Definition 1.2 *For $\mathcal{X} \subseteq \mathbb{R}^p$, a kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called conditionally positive definite (c.p.d.)* ***iff*** *$\forall n \in \mathcal{N}, \forall \mathbf{a} \subseteq \mathbb{R}^n$ such that $\sum_{i=1}^n a_i = 0$ and $\mathbf{K}_n = [K(x_i, x_j)]_{n\times n}$:*

$$\mathbf{a}^T \mathbf{K}_n \mathbf{a} \geq 0$$

Definition 1.3 *A kernel K is called negative definite (n.d.)* ***iff*** *$-K$ is c.p.d.*

Note: In the following discussion, we only consider the case where K is symmetric.

Definition 1.4 *For a symmetric p.d. kernel K, we define its normalized kernel K_0 as follows:*

$$\forall x_i, x_j \in \mathcal{X}, K_0(x_i, x_j) = \begin{cases} 0 & if(K(x_i, x_i) = 0 \wedge K(x_j, x_j) = 0) \\ \frac{K(x_i,x_j)}{|K(x_i,x_i)|^{1/2}|K(x_j,x_j)|^{1/2}} & otherwise \end{cases}$$

Note:$K(x_i, x_j) \leq 1$ always holds since for any *p.d.* matrix $\mathbf{K}$:

$$|\mathbf{K}| = \begin{vmatrix} K(x_i, x_i) & K(x_i, x_j) \\ K(x_j, x_i) & K(x_j, x_j) \end{vmatrix} = K(x_i, x_i)K(x_j, x_j) - K^2(x_i, x_j) \geq 0$$

Moreover, whenever there is some i where $K(x_i, x_i) = 0$, elements in the ith row and column are all 0. Therefore we can always removes these lines so that we can only consider case 2.

It is clear that $\mathbf{K}_0 = \mathbf{D}^{-1/2}\mathbf{K}\mathbf{D}^{-1/2}$ where $\mathbf{D} = \text{diag}(K(x_i, x_i))$.

1.2 Properties

Proposition 1.1 *A function $f : \mathcal{X} \to \mathcal{F} \subseteq \mathbb{R}^r$, $\mathcal{X} \subseteq \mathbb{R}^p$. If $K(x, y) = f(x)^T f(y)$, then K is p.d.*

Proof:
Let $\mathbf{F}$ be a matrix:

$$\mathbf{F} = \begin{pmatrix} f_1(x_1) & \ldots & f_r(x_1) \\ f_1(x_2) & \ldots & f_r(x_2) \\ & \vdots & \\ f_1(x_n) & \ldots & f_r(x_n) \end{pmatrix} = \begin{pmatrix} f(x_1)^T \\ f(x_2)^T \\ \vdots \\ f(x_n)^T \end{pmatrix}$$

then $\mathbf{K} = \mathbf{F}\mathbf{F}^T$ is *p.s.d.*, so K is *p.d.*

Proposition 1.2 *Given a sequence of kernels* $\{K_n\}_{n\in\mathbb{N}}$ *where* $\lim_{n\to\infty} K_n(x,y) = K(x,y)$, *and a power series* $\sum_{n=0}^{\infty} a_n x^n$, *then the two series* $\sum_{n=0}^{\infty} a_n x^n$ *and* $\sum_{i=0}^{\infty} a_i K^i(x,y)$ *have the same Radius of convergence.*

This proposition is referred as pointwise limit.

Theorem 1.1 p.d. *kernels are closed under sum, direct product, tensor product,pointwise limit and composition with a power series* $\sum_{n=1}^{\infty} a_n x^n$ *with* $a_n \geq 0$ *for all* $n \in \mathbb{N}$

Proof:

We start with two kernel matrices, $\mathbf{K}_1$ and $\mathbf{K}_2$, generated from *p.d.* kernels K_1 and K_2 for an arbitrary set of m points. By assumption, these kernel matrices are *p.s.d.* Observe that for any $\mathbf{a} \in \mathbb{R}^m$,

$$(\mathbf{a}^T\mathbf{K}_1\mathbf{a} \geq 0) \wedge (\mathbf{a}^T\mathbf{K}_2\mathbf{a} \geq 0) \implies \mathbf{a}^T(\mathbf{K}_1 + \mathbf{K}_2)\mathbf{a} \geq 0.$$

This shows that $\mathbf{K}_1 + \mathbf{K}_2$ is *p.s.d.* and thus that $K_1 + K_2$ is *p.d.*

To show closure under direct product, we will use the fact that for any *p.s.d.* matrix $\mathbf{K}$ there exists $\mathbf{M}$ such that $\mathbf{K} = \mathbf{M}\mathbf{M}^T$. The existence of $\mathbf{M}$ is guaranteed as it can be generated via singular value decomposition(SVD) of $\mathbf{K}$, or by Cholesky decomposition.The kernel matrix associated to $K_1 \odot K_2$ is $[(\mathbf{K}_1)_{ij}(\mathbf{K}_2)_{ij}]$. For any $\mathbf{a} \in \mathbb{R}^m$, expressing $\mathbf{K}_{ij}$ in terms of the entries of $\mathbf{M}$, we can write

$$\begin{aligned}
\sum_{i,j=1}^{m} a_i a_j (\mathbf{K}_1)_{ij}(\mathbf{K}_2)_{ij} &= \sum_{i,j=1}^{m} a_i a_j [\sum_{k=1}^{m} \mathbf{M}_{ik}\mathbf{M}_{jk}(\mathbf{K}_2)_{ij}] \\
&= \sum_{k=1}^{m} [\sum_{i,j=1}^{m} a_i a_j \mathbf{M}_{ik}\mathbf{M}_{jk}(\mathbf{K}_2)_{ij}] \\
&= \sum_{k=1}^{m} [\sum_{i,j=1}^{m} (a_i\mathbf{M}_{ik})(a_j\mathbf{M}_{jk})(\mathbf{K}_2)_{ij}] \\
&= \sum_{k=1}^{m} \mathbf{b}^{(k)^T}\mathbf{K}_2\mathbf{b}^{(k)} \geq 0,
\end{aligned}$$

with $\mathbf{b}^{(k)} = \begin{bmatrix} a_1\mathbf{M}_{1k} \\ \vdots \\ a_m\mathbf{M}_{mk} \end{bmatrix}$. This shows that *p.d.* kernels are closed under direct product.

Definition 1.5 *Two kernels* $K_1(x_i, x_j) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}, \{x_1, x_2, \cdots, x_n\} \subseteq \mathcal{X}$ *and* $K_2(y_i, y_j) : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}, \{y_1, y_2, \cdots, y_m\} \subseteq \mathcal{Y}$, *their tensor product*

$$K((x_i, y_i), (x_j, y_j)) = K_1 \otimes K_2 = \left[(\mathbf{K}_1)_{ij}\mathbf{K}_2\right]$$

We define $\hat{K}_1((x_1, y_1), (x_2, y_2)) = K_1(x_1, x_2)$ and $\hat{K}_2((x_1, y_1), (x_2, y_2)) = K_2(y_1, y_2)$, the tensor product of K_1 and K_2 is *p.d.* as the product of the two *p.d.* kernels $\hat{K}_1$ and

$\hat{K}_2$. This shows $K((x_i, y_i), (x_j, y_j)) = (K_1 \otimes K_2)((x_i, y_i), (x_j, y_j)) = K_1(x_i, y_i)K_2(x_j, y_j)$ is positive definite symmetric.

Another idea to prove the closure under tensor product is using SVD. Since K_1 and K_2 are *p.d.*, their kernel matrices $\mathbf{K}_1$ and $\mathbf{K}_2$ are *p.s.d.* By SVD, $\mathbf{K}_1 = \mathbf{U}_1\Lambda_1{\mathbf{U}_1}^T$, $\mathbf{K}_2 = \mathbf{U}_2\Lambda_2{\mathbf{U}_2}^T$, then we have

$$\mathbf{K} = (\mathbf{U}_1 \otimes \mathbf{U}_2)(\Lambda_1 \otimes \Lambda_2)({\mathbf{U}_1}^T \otimes {\mathbf{U}_2}^T)$$

This shows that $K_1 \otimes K_2$ is *p.d.*

Let K_n, $(n \in \mathbb{N})$ be a sequence of *p.d.* kernels with pointwise limit K. Let $\mathbf{K}$ be the kernel matrix associated to K and $\mathbf{K}_n$ the one associated to K_n for any $n \in \mathbb{N}$. Observe that

$$\forall n, \mathbf{c}^T\mathbf{K}_n\mathbf{c} \geq 0 \implies \lim_{n\to\infty} \mathbf{c}^T\mathbf{K}_n\mathbf{c} \geq 0$$

This shows the closure under pointwise limit.

Assume that K is a *p.d.* kernel, then, for any $n \in \mathbb{N}$, K^N and a_nK_n are *p.d.* by closure under product. For any $N \in \mathbb{N}$, $\sum_{n=0}^{N} a_nK^n$ is *p.d.* by closure under sum of a_nK_n. At last, we have $\sum_{n=0}^{\infty} a_nK^n$ is *p.d.* by closure under the pointwise limit.

Theorem 1.2 *Let $\mathcal{X}$ be a nonempty set, $x_0 \in \mathcal{X}$, and let $\boldsymbol{\phi} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric kernel. Kernel K satisfies:*

$$K(x, y) = \boldsymbol{\phi}(x, x_0) + \boldsymbol{\phi}(y, x_0) - \boldsymbol{\phi}(x, y) - \boldsymbol{\phi}(x_0, x_0)$$

Then K is p.d. ***iff*** *$\boldsymbol{\phi}$ is n.d.*

Proof:

We first prove the sufficiency: $(\Rightarrow)$

Since K is *p.d.*, it is *c.p.d.* Thus, $\forall\ \{x_1, x_2, \ldots, x_n\} \subseteq \mathcal{X}$ and $\forall\ \mathbf{a}$ such that $\sum_{i=1}^{n} a_i = 0$ we have:

$$\begin{aligned} & \sum_{i,j} a_i a_j K(x_i, x_j) \\ = & \sum_{i,j} a_i a_j [\boldsymbol{\phi}(x_i, x_0) + \boldsymbol{\phi}(x_j, x_0) - \boldsymbol{\phi}(x_i, x_j) - \boldsymbol{\phi}(x_0, x_0)] \\ = & -\sum_{i,j} a_i a_j \boldsymbol{\phi}(x_i, x_j) \geq 0 \end{aligned}$$

which implies $-\boldsymbol{\phi}$ is *c.p.d.* so that $\boldsymbol{\phi}$ is *n.d.*

Then we move to prove the necessity: $(\Leftarrow)$

$\forall\ \{a_1, a_2, \ldots, a_n\} \subseteq \mathbb{R}$, we let $a_0 = -\sum_{i=1}^{n} a_i$ such that $\sum_{i=0}^{n} a_i = 0$. Since $\boldsymbol{\phi}$ is *n.d.*, we have:

$$
\begin{aligned}
& \sum_{i,j=0}^{n} a_i a_j \boldsymbol{\phi}(x_i, x_j) \\
= & \sum_{i,j=1}^{n} a_i a_j \boldsymbol{\phi}(x_i, x_j) + \sum_{i=1}^{n} a_i a_0 \boldsymbol{\phi}(x_i, x_0) + \sum_{j=1}^{n} a_0 a_j \boldsymbol{\phi}(x_0, x_j) + a_0^2 \boldsymbol{\phi}(x_0, x_0) \\
= & -\sum_{i,j=1}^{n} a_i a_j K(x_i, x_j) \leq 0
\end{aligned}
$$

Theorem 1.3 *Let $\mathcal{X}$ be a nonempty set and $\boldsymbol{\phi} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric kernel. Then $\boldsymbol{\phi}$ is n.d.* ***iff*** *$K = \exp(-t\boldsymbol{\phi})$ is p.d. for all $t > 0$.*

Proof:

We first prove the necessity: ($\Leftarrow$)
Since we have:

$$\lim_{t \to 0^+} \frac{1 - \exp(-t\boldsymbol{\phi})}{t} = \boldsymbol{\phi}$$

according to Theorem 1.1, $\boldsymbol{\phi}$ is *n.d.*

Then we prove the sufficiency: ($\Rightarrow$)
Let $t = 1$ and

$$\boldsymbol{\psi}(x, y) = -\boldsymbol{\phi}(x, y) - \boldsymbol{\phi}(x_0, x_0) + \boldsymbol{\phi}(x, x_0) + \boldsymbol{\phi}(y, x_0)$$

such that $\boldsymbol{\psi}$ is *p.d.*, and then we have:

$$
\begin{aligned}
& \exp(-\boldsymbol{\phi}) \\
= & \exp(\boldsymbol{\psi}(x, y)) \exp(\boldsymbol{\phi}(x_0, x_0)) \exp(-\boldsymbol{\phi}(x, x_0)) \exp(-\boldsymbol{\phi}(y, x_0))
\end{aligned}
$$

Denote $g(x) = \exp(-\boldsymbol{\phi}(x, x_0))$, $K_1(x, y) = \exp(\boldsymbol{\psi}(x, y)) \exp(\boldsymbol{\phi}(x_0, x_0))$ and $K_2(x, y) = \exp(-\boldsymbol{\phi}(x, x_0)) \exp(-\boldsymbol{\phi}(y, x_0)) = g(x)^T g(y)$. Notice that $\exp(\boldsymbol{\phi}(x_0, x_0))$ is a constant, K_1 is *p.d.* On the other hand, according to Proposition 1.1 K_2 is *p.d.* Since $\exp(-\boldsymbol{\phi}) = K_1 \odot K_2$, we have $\exp(-\boldsymbol{\phi})$ is *p.d.*

Theorem 1.4 *Let F be a probability measure on the half line of $\mathbb{R}$ such that $0 < \int_0^\infty s dF(s) < \infty$ and $\mathcal{L}_F(u) = \int_0^\infty \exp(-su) dF(s)$, $u \in \mathbb{R}_+$, then*
$\boldsymbol{\phi} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is n.d. ***iff*** *$\int_0^\infty \exp(-ts\boldsymbol{\phi}) dF(s)$ is p.d for all $t > 0$.*

(Proof is similar to Theorem 1.3)

1.3 Examples

1.3.1 Polynomial Kernel

Definition 1.6 *A polynomial kernel is defined as:*

$$K(x,y) = (x^T y + \alpha)^d,\ x, y \in \mathbb{R}^p$$

where $\alpha > 0$ *and* $d \in \mathbb{N}$.

Note:A polynomial kernel maps p-dimensional nonlinear data to $\binom{p+d}{d}$-dimensional linear representation.

1.3.2 RBF Gaussian Kernel

Definition 1.7 *A RBF Gaussian kernel is defined as:*

$$K(x,y) = \exp(-\frac{||x-y||^2}{2\sigma^2})$$

Note:a RBF Gaussian kernel maps arbitrary data to ∞-dimensional representation. (Can be regarded as a Taylor expansion)

1.3.3 1-Norm Kernel

Definition 1.8 *A 1-Norm kernel is defined as:*

$$K(x,y) = \exp(-||x-y||_1)$$

where $||x||_1$ *is the 1-norm of* x.

1-Norm kernel is *p.d.* We can prove it by Theorem 1.4.
Proof:

$$\exp(-tz^{1/2}) = \int \exp(-tsz)\mathrm{d}F(s)$$

where $f(s) = \sqrt{\frac{t}{2\pi}}u^{-3/2}\exp(-\frac{t}{2u})\mathrm{d}u$

Take $t = 1$ and $z = ||x-y||_1^2$, we have:

$$\exp(-||x-y||_1) = \int \exp(-s||x-y||_1^2)\mathrm{d}F(s)$$

Then we only need to show $||x-y||_1^2$ is *n.d.*

1.4 Empirical Feature Function

Definition 1.9 (Empirical Feature function) *The empirical feature function $\boldsymbol{\phi}$ associated to a* p.d. *kernel K is a feature mapping that $K(x_i, x_j) = \boldsymbol{\phi}(x_i)^T \boldsymbol{\phi}(x_j)$.*

Given a training sample $\{x_1, x_2, \ldots, x_m\} \subseteq \mathcal{X}, \Psi : \mathcal{X} \to \mathbb{R}^m$ is defined for all $x \in \mathcal{X}$ by

$$\Psi(x) = \begin{bmatrix} K(x, x_1) \\ \vdots \\ K(x, x_m) \end{bmatrix}$$

Then Φ can be defined as follows using the empirical kernel map Ψ:

$$\begin{aligned} \Phi &= \mathbf{K}^{\dagger\frac{1}{2}} \Psi(x) \\ &= \mathbf{K}^{\dagger\frac{1}{2}} \begin{bmatrix} \Psi(x_1) & \cdots & \Psi(x_m) \end{bmatrix} \\ &= \mathbf{K}^{\dagger\frac{1}{2}} \mathbf{K} \end{aligned}$$

Thus,

$$\Phi^T \Phi = \mathbf{K}\mathbf{K}^{\dagger\frac{1}{2}} \mathbf{K}^{\dagger\frac{1}{2}} \mathbf{K} = \mathbf{K}\mathbf{K}^{\dagger}\mathbf{K} = \mathbf{K}$$

where $\mathbf{K}^{\dagger\frac{1}{2}} = \mathbf{U}\Lambda^{\frac{1}{2}}\mathbf{U}^T, \mathbf{K} = \mathbf{U}\Lambda\mathbf{U}^T$.

1.5 Reproducing Kernel Hilbert Space

Theorem 1.5 (Reproducing Kernel Hilbert space) *Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a p.d. kernel. Then, there exists a Hilbert space $\mathcal{H}$ and a mapping $\boldsymbol{\phi} \in \mathcal{H}$ such that:*

$$\forall x, y \in \mathcal{X}, K(x, y) = < \boldsymbol{\phi}(x), \boldsymbol{\phi}(y) >$$