1 Kernels

1. let $(x, y) \in \mathbb{R}^+ \times \mathbb{R}^+$, where $\mathbb{R}^+ = \{x \in \mathbb{R}; x \geq 0\}$, the "french positive" real numbers.

   (a) Verify that $\min(x, y) = \int_0^\infty \mathbb{I}_{t\leq x}\mathbb{I}_{t\leq y} dt$ where $\mathbb{I}_A = \begin{cases} 1 & \text{if A is true} \\ 0 & \text{otherwise} \end{cases}$

   When $x \leq y$,

$$\begin{aligned}\int_0^\infty \mathbb{I}_{t\leq x}\mathbb{I}_{t\leq y} dt &= \int_0^x \mathbb{I}_{t\leq x}\mathbb{I}_{t\leq y} dt + \int_x^y \mathbb{I}_{t\leq x}\mathbb{I}_{t\leq y} dt + \int_y^\infty \mathbb{I}_{t\leq x}\mathbb{I}_{t\leq y} dt \\ &= \int_0^x 1 \cdot 1 dt + \int_x^y 0 \cdot 1 dt + \int_y^\infty 0 \cdot 0 dt = x\end{aligned}$$

   By the same way, when $y \leq x$, $\int_0^\infty \mathbb{I}_{t\leq x}\mathbb{I}_{t\leq y} dt = y$.
   Therefore, $\min(x, y) = \int_0^\infty \mathbb{I}_{t\leq x}\mathbb{I}_{t\leq y} dt$

   (b) Use the previous question to show that $K(x, y) = \min(x, y)$ is a pd kernel over $\mathbb{R}^+$

   $K(x, y) = \min(x, y) = \int_0^\infty \mathbb{I}_{t\leq x}\mathbb{I}_{t\leq y} dt = \min(y, x) = K(y, x)$ symmetric

$$\sum_{i=1}^n \sum_{j=1}^n \alpha_i\alpha_j \min(x, y) = \int_0^\infty \sum_{i=1}^n \alpha_i \mathbb{I}_{t\leq x} \sum_{j=1}^n \alpha_j \mathbb{I}_{t\leq y} dt = \int_0^\infty (\sum_{i=1}^n \alpha_i \mathbb{I}_{t\leq x})^2 dt \geq 0$$

   (c) Show that $\max(x, y)$ is not a pd kernel over $\mathbb{R}^+$.

When $x \leq y$, $\int_0^\infty \mathbb{I}_{t\geq x}\mathbb{I}_{t\geq y} dt = \int_y^\infty 1 \cdot 1 dt = t|_y^\infty \neq \max(x, y)$
The Gram Matrix

$$M(x, y) = \begin{bmatrix} \max(x, x) & \max(x, y) \\ \max(x, y) & \max(y, y) \end{bmatrix} = \begin{bmatrix} x & y \\ y & y \end{bmatrix} = xy - y^2 \leq 0$$

When $y \leq x$, it is the same. $M(x, y) = xy - x^2 \leq 0$
Therefore, $\max(x, y)$ can not be a p.d. kernel over $\mathbb{R}^+$

2. Consider a probability space $(\Omega, \mathcal{A}, P)$

   (a) Define for any two events A and B, $K_1(A, B) = P(A \cap B)$ where $A \cap B$ is the intersection between the events A and B Verify that K_1 is positive definite. Hint: $P(A) = E[\mathbb{I}_A]$

   $K_1(A, B) = P(A \cap B) = P(B \cap A) = K_1(B, A)$ symmetric
   $P(A) = E[\mathbb{I}_A]$; $P(B) = E[\mathbb{I}_B]$; $P(A \cap B) = E[\mathbb{I}_A\mathbb{I}_B]$

$$\sum_{i=1}^n \sum_{j=1}^n \alpha_i\alpha_j E[\mathbb{I}_{A_i}\mathbb{I}_{A_j}] = E[\sum_{i=1}^n \sum_{j=1}^n \alpha_i\alpha_j \mathbb{I}_{A_i}\mathbb{I}_{A_j}] = E[(\sum_{i=1}^n \alpha_i \mathbb{I}_{A_i})^2] \geq 0$$

   (b) Define for any two events A and B, $K_2(A, B) = P(A \cap B) - P(A)P(B)$ Verify that K_2 is positive definite.

$K_2(A, B) = P(A \cap B) - P(A)P(B) = E[\mathbb{I}_A\mathbb{I}_B] - E[\mathbb{I}_A]E[\mathbb{I}_B] = Cov[\mathbb{I}_A, \mathbb{I}_B]$

$$\sum_{i=1}^n \sum_{j=1}^n \alpha_i\alpha_j Cov[\mathbb{I}_{A_i}, \mathbb{I}_{A_j}] = Cov[\sum_{i=1}^n \alpha_i \mathbb{I}_{A_i}, \sum_{j=1}^n \alpha_j \mathbb{I}_{A_j}] = Var[\sum_{i=1}^n \alpha_i \mathbb{I}_{A_i}] \geq 0$$

2 Kernels and RKHS

1. Define the RKHS over $\mathbb{R}^d$ $K(x,y) = x^T y + c$ where $c > 0$.

   (a) What is the RKHS associated with the kernel K? no proof is required.

   $$\mathcal{H} = \{f: \mathbb{R}^d \mapsto \mathbb{R};\ f_{w,w_0}(x) = w^T x + w_0; \quad w \in \mathbb{R}^d, w_0 \in \mathbb{R}\}$$

   (b) What is the inner product in this RKHS? no proof required.

   $$\langle f_{v,v_0}, f_{w,w_0}\rangle_{\mathcal{H}} = v^T w + \frac{1}{c} v_0 w_0 \Rightarrow \langle f_{v,v_0}, f_{v,v_0}\rangle = \|f_{v,v_0}\|^2_{\mathcal{H}} = \|v\|^2 + \frac{v_0^2}{c}$$

   (c) Verify the reproducing property
   $\mathcal{H}$ contains all the functions $k(\cdot, x_i) : t \mapsto k(t,x) = t^T x + c = f_t(x)$

   $$\langle f_{w,w_0}, k(\cdot, x)\rangle = \langle f_{w,w_0}, f_{x,c}\rangle = x^T w + \frac{1}{c} c w_0 = w^T x + w_0 = f_w(x)$$

   $\therefore \langle f, k(\cdot, x)\rangle_{\mathcal{H}} = f(x)$ for each $f \in \mathcal{H}$, $x \in \mathcal{X}$

2. Define the RKHS over $\mathbb{R}^d$ $K(x,y) = (x^T y)^2$ The RKHS associated with the kernel K is $\{f_S; f_S(x) = x^T S x\}$ where S is a symmetric (d,d) matrix. The inner product is $< f_{S_1}, f_{S_2} >=< S_1, S_2 >_F$

   (a) Verify the reproducing property.
   $\mathcal{H}$ contains all the functions $k(\cdot, x_i) : t \mapsto k(t,x) = (t^T x)(t^T x) = x^T \cdot (tt^T) \cdot x = f_t(x)$

   $$\langle f_S, k(\cdot, x)\rangle_{\mathcal{H}} = \langle f_S, f_{xx^T}\rangle_{\mathcal{H}} = \langle S, xx^T\rangle_{\mathcal{F}} = \text{trace}[Sxx^T] = \text{trace}[x^T S x] = x^T S x = f_S(x)$$

   $\therefore \langle f_S, k(\cdot, x)\rangle_{\mathcal{H}} = f_S(x)$ for each $f \in \mathcal{H}$, $x \in \mathcal{X}$

   (b) Why do we require that S is symmetric?
   Only if $\underset{(d,d)}{S}$ is a symmetric Matrix, $t^T x = x^T t$.
   If not, we can not complete the step of $(t^T x)(t^T x) = x^T \cdot (tt^T) \cdot x$ in (b).

3. Define the RKHS over $\mathbb{R}^d$ $K(x,y) = (x^T y + c)^2$ where $c > 0$.

   (a) What is the RKHS associated with the kernel K? no proof is required.

   $$\{f_{S,s,s_0} : f_{S,s,s_0}(x) = x^T S x + 2 s_0 s^T x + s_0^2; \quad S \in \mathbb{R}^{d\times d}, s \in \mathbb{R}^d, s_0 \in \mathbb{R}\}$$

   (b) What is the inner product in this RKHS? no proof required.

   $$\langle f_{S_1,s_1,s_{10}}, f_{S_2,s_2,s_{20}}\rangle_{\mathcal{H}} = \langle S_1, S_2\rangle_{\mathcal{F}} + \frac{2 s_{10} s_{20}}{c}(s_1^T s_2 + \frac{s_{10} s_{20}}{c})$$

   (c) Verify the reproducing property
   $\mathcal{H}$ contains all the functions $k(\cdot, x_i) : t \mapsto k(t,x) = (t^T x + c)^2 = x^T \cdot (tt^T) \cdot x + 2ct^T x + c^2 = f_t(x)$

   $$\langle f_{S,s,s_0}, k(\cdot, x)\rangle_{\mathcal{H}} = \langle f_{S,s,s_0}, f_{xx^T,x,c}\rangle_{\mathcal{H}} = \langle S, xx^T\rangle_{\mathcal{F}} + \frac{2 s_0 c}{c}(s^T x + \frac{s_0 c}{c})$$
   $$= x^T S x + 2 s_0 s^T x + s_0^2 = f_{S,s,s_0}(x)$$

   $\therefore \langle f_{S,s,s_0}, k(\cdot, x)\rangle_{\mathcal{H}} = f_{S,s,s_0}(x)$ for each $f \in \mathcal{H}$, $x \in \mathcal{X}$

3 Fisher kernel

Let $\theta \in \mathbb{R}$ be a parameter and let p_θ be a probabilistic model (i.e a point mass function or a density) over a set $\mathcal{X}$ indexed by θ. Let $\theta_0 \in \mathbb{R}$ be a specific value for θ. Let us define the Fisher score at $x \in \mathcal{X}$ as $\phi(x,\theta_0) = \frac{\delta}{\delta\theta}\ln p_\theta(x)$ evaluated at $\theta = \theta_0$ assuming that this quantity exists. Define $I(\theta)$, the Fisher information associated with the parameter θ, i.e., $I(\theta) = E[\phi^2(X,\theta)]$ where E stands for expectation and X is a random variable with distribution p_θ. The Fisher kernel is then $k(x,x') = \frac{\phi(x,\theta_0)\phi(x',\theta_0)}{I(\theta_0)}$ where

1. Verify that $k(.,.)$ is a positive definite kernel over $\mathcal{X}$

   $k(x,x') = \frac{\phi(x,\theta_0)\phi(x',\theta_0)}{I(\theta_0)} = \frac{\phi(x',\theta_0)\phi(x,\theta_0)}{I(\theta_0)} = k(x',x)$ symmetric.

   For $I(\theta) = E[\phi^2(X,\theta)] \geq 0$,

   $k(x,x') = \frac{1}{I(\theta_0)}\sum_{i=1}^n \alpha_i\phi(x_i,\theta_0)\sum_{j=1}^n \alpha_j\phi(x_j,\theta_0) = \frac{1}{I(\theta_0)}[\sum_{i=1}^n \alpha_i\phi(x_i,\theta_0)]^2 \geq 0$

   $\therefore k(.,.)$ is a positive definite kernel over $\mathcal{X}$

2. Consider the following model: $x \in \{0,1\}$, $X \sim Bernoulli(\theta)$, $0 < \theta < 1$, that is $p_\theta(x) = \theta^x(1-\theta)^{(1-x)}$ We recall that in this case $E[X] = \theta$ and $Var[X] = E[(X-\theta)^2] = \theta(1-\theta)$ Compute $k(x,x')$

$$\begin{aligned} p_\theta(x) &= \theta^x(1-\theta)^{(1-x)} \\ \ln p_\theta(x) &= x\ln\theta + (1-x)\ln(1-\theta) \\ \frac{d}{d\theta}\ln p_\theta(x) &= \frac{x}{\theta} + \frac{1-x}{1-\theta} = \frac{x-\theta}{\theta(1-\theta)} \end{aligned}$$

$$I(\theta) = E[\phi^2(X,\theta)] = E[(\frac{X-\theta}{\theta(1-\theta)})^2] = \frac{E[(X-\theta)^2]}{\theta^2(1-\theta)^2} = \frac{V[X]}{\theta^2(1-\theta)^2} = \frac{\theta(1-\theta)}{\theta^2(1-\theta)^2} = \frac{1}{\theta(1-\theta)}$$

$$k(x,x') = \frac{\phi(x,\theta_0)\phi(x',\theta_0)}{I(\theta_0)} = \frac{(x-\theta_0)(x'-\theta_0)}{\theta_0^2(1-\theta_0)^2}\theta_0(1-\theta_0) = \frac{(x-\theta_0)(x'-\theta_0)}{\theta_0(1-\theta_0)} \qquad \square$$

3. Assume now $x = (x_1,x_2)$ with $x_1 \in \{0,1\}$ and $x_2 \in \{0,1\}$. We consider the following model where $X = (X_1,X_2)$, X_1 and X_2 are independent with the same $Bernoulli(\theta)$ distribution. Compute $k(x,x')$.

$$\begin{aligned} p_\theta(\vec{x}) &\underset{x_1\perp x_2}{=} p_\theta(x_1)p_\theta(x_2) = \theta^{x_1+x_2}(1-\theta)^{2-x_1-x_2} \\ \ln p_\theta(x) &= (x_1+x_2)\ln\theta + (2-x_1-x_2)\ln(1-\theta) \\ \phi(\vec{x},\theta) &= \frac{d}{d\theta}\ln p_\theta(x) = \frac{x_1+x_2}{\theta} + \frac{2-x_1-x_2}{1-\theta} = \frac{x_1+x_2-2\theta}{\theta(1-\theta)} \end{aligned}$$

$$\begin{aligned} I(\theta) &= E[\phi^2(\vec{X},\theta)] = \frac{E[(X_1+X_2-2\theta)^2]}{\theta^2(1-\theta)^2} \\ &\underset{x_1\perp x_2}{=} \frac{E[(X_1-\theta)^2] + E[(X_2-\theta)^2] + 2(E[X_1]-\theta)(E[X_2]-\theta)}{\theta^2(1-\theta)^2} \\ &= \frac{V[X_1]+V[X_2]-0}{\theta^2(1-\theta)^2} = \frac{2\theta(1-\theta)}{\theta^2(1-\theta)^2} = \frac{2}{\theta(1-\theta)} \end{aligned}$$

$$\begin{aligned} k(x,x') &= \frac{\phi(\vec{x},\theta_0)\phi(\vec{x}',\theta_0)}{I(\theta_0)} = \frac{(x_1+x_2-2\theta_0)(x_1'+x_2'-2\theta_0)}{\theta_0^2(1-\theta_0)^2}\frac{\theta_0(1-\theta_0)}{2} \\ &= \frac{(x_1+x_2-2\theta_0)(x_1'+x_2'-2\theta_0)}{2\theta_0(1-\theta_0)} \qquad \square \end{aligned}$$