STAT 671

Statistical Learning I

Fall 2019
Homework 2
Due October 28^{th} at the beginning of class

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 $\mathbb{I}_A = \begin{cases} 1 & \text{if A is true} \\ 0 & \text{otherwise} \end{cases}$

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

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

   $\max(x, y) = \int_0^\infty \mathbb{I}_{t\geq x}\mathbb{I}_{t\geq y}dt$

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

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

   $k_1(x, y) = \sum_{i=1}^n \alpha_i E[\mathbb{I}_A] \sum_{j=1}^n \alpha_j E[\mathbb{I}_A] = \| \sum_{i=1}^n \alpha_i E[\mathbb{I}_A]\|^2 \geq 0$

3. 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.
   $K_2(A, B) = P(A \cap B) - P(A)P(B) = E[\mathbb{I}_A\mathbb{I}_B] - E[\mathbb{I}_A]E[\mathbb{I}_B] = Cov[\mathbb{I}_A, \mathbb{I}_B]$
   $K_2(x, y) = \sum_{i=1}^n \sum_{j=1}^n \alpha_i\alpha_j Cov[\mathbb{I}_A, \mathbb{I}_A] = \sum_{i=1}^n \sum_{j=1}^n \alpha_i\alpha_j Var[\mathbb{I}_A] \geq 0$

2 Kernels and RKHS

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

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

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

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

   $$\langle f_{v,v_0}, f_{w,w_0}\rangle_\mathcal{H} = v^Tw + \frac{1}{c}v_0w_0 \Rightarrow \langle f_{v,v_0}, f_{v,v_0}\rangle = \|f_{v,v_0}\|^2_\mathcal{H} = \|v\|^2 + \frac{v_0^2}{c}$$

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

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

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

$$\mathcal{H} = \{f_S : f_S(x) = x^T S_x;\}$$

   $\mathcal{H}$ contains all the functions $k(\cdot, x_i) : t \mapsto k(x,t) = x x^T t t^T$

$$\langle f_{S_1}, f_{S_2}\rangle_{\mathcal{H}} = \langle S_1, S_2\rangle_{\mathcal{F}} = \sum_{i,j=1}^{n} [S_1]_{ij}[S_2]_{ij}$$

$$[S_1]_{ij}[S_2]_{ij} = \text{trace}[(x_i^T x_j)(y_j^T y_i)] = \text{trace}[(y_i x_i^T)(x_j y_j^T)] = \langle x_i y_i^T, x_j y_j^T\rangle_{\mathcal{F}} = \langle z_i, z_j\rangle_{\mathbb{R}^{n^2}}$$

$$\langle f_{S_1}, k(\cdot, x_i)\rangle_{\mathcal{H}} = \langle f_{S_1}, f_{xx^T}\rangle_{\mathcal{H}} = \langle S_1, x x^T\rangle_{\mathcal{F}} = f_{S_1}(x)$$

   (b) Why do we require that S is symmetric?
   $\underset{(d,d)}{S}$ is a symmetric Matrix, $y^T x = x^T y$

$$k(y,x) = (y^T x)(y^T x) = y^T \cdot x x^T \cdot y$$

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.

$$\{f_{S,s_0}; f_S(x) = x^T S_x + s_0\}$$

   where S is a symmetric (d,d) matrix

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

$$\langle f_{S_1}, f_{S_2}\rangle_{\mathcal{H}} = \langle S_1, S_2\rangle_{\mathcal{F}} + \frac{s_0^2}{c}$$

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

$$\langle f_{S_1}, k(\cdot, x_i)\rangle_{\mathcal{H}} = \langle f_{S_1}, f_{xx^T}\rangle_{\mathcal{H}} = \langle S_1, x x^T\rangle_{\mathcal{F}} = x^T S_1 + s_0 = f_{S_1}(x)$$

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

$$
\begin{aligned}
p_\theta(x) &= \theta^x (1-\theta)^{(1-x)} \\
\ln p_\theta(x) &= x \ln \theta + (1-x) \ln(1-\theta) \\
\phi(x, \theta_0) = \frac{d}{d\theta} \ln p_\theta(x) &= \frac{x}{\theta} + \frac{1-x}{1-\theta} = \frac{x-\theta}{\theta(1-\theta)}
\end{aligned}
$$

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

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 $k(x, x')$

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

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

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 $k(x, x')$.

$$
k(x, x') = \frac{(x-\theta_0)(x'-\theta_0)}{\theta_0(1-\theta_0)} = \frac{xx' - (x+x')\theta_0 + \theta_0^2}{\theta_0(1-\theta_0)} = \frac{x_1x_1' + x_2x_2' - (x_1+x_2, x_1'+x_2')^T\theta_0 + \theta_0^2}{\theta_0(1-\theta_0)}
$$