Kernels Methods homework1

April 3, 2016

Exercise 1

Question 1

k_1 and k_2 are p.d.s so their gram matrix (associated to any data examples $\{x_1, ..., x_n\}$) are positive semi-definite: $\forall A \in \mathbb{R}^n$

$$A^T K_1 A \geq 0 \ and \ A^T K_2 A \geq 0$$

Or

$$A^T(\alpha K_1 + \beta K_2)A = \alpha A^T K_1 A + \beta A^T K_2 A \ \ \geq 0$$

The gram matrix associated is positive semi-definite. Let's ϕ_1 and ϕ_2 the features maps associated to k_1 and k_2, $k = \alpha k_1 + \beta k_2$ and its associated feature map ϕ.

$$k(x,y) = <\phi(x), \phi(y)> = \alpha\phi_1(x)\phi_1(y) + \beta\phi_2(x)\phi_2(y)$$

$$k(y,x) = \alpha\phi_1(y)\phi_1(x) + \beta\phi_2(y)\phi_2(x) = k(x,y)$$

k is then symmetric.
We can conclude that $\alpha K_1 + \beta K_2$ is p.d.

Question 2

(Schur Product wiki)
Again, we will use the fact that the gramm matrix associated to k_1 and k_2 are positive semi-definite.
Let's $k(x,y) = k_1(x,y).k_2(x,y)$, we can easily check that k_1 and k_2 being symmetric, k is also symmetric.
we want to prove that the gram matrix K is positive semi definite.
$\forall A \in \mathbb{R}^n$:

$$A^T K A = \sum_{i,j} A_i A_j K(x_i, x_j) = \sum_{i,j} A_i A_j K_1(x_i, x_j) * K_2(x_i, x_j)$$

Or K_1 (same goes for K_2) is a positive semi-definite matrix so it's eigenvalues decomposition follows $K_1 = U\Sigma U^T = U(\Sigma^{1/2})^T \Sigma^{1/2} U^T = M_1^T M_1$ where $M_1 = \Sigma^{1/2} U^T$

$$K_1 = M_1^T M_1 \to [K_1]_{i,j} = \sum_k (M_1)_{ik}(M_1)_{jk}$$

$$K_2 = M_2^T M_2 \to [K_2]_{i,j} = \sum_l (M_2)_{il}(M_2)_{jl}$$

If we plug it in the previous equation:

$$A^TKA = \sum_{i,j} A_iA_j \sum_k (M_1)_{ik}(M_1)_{jk} \sum_l (M_2)_{il}(M_2)_{jl}$$

$$A^TKA = \sum_{k,l} \sum_{i,j} A_iA_j(M_1)_{ik}(M_1)_{jk}(M_2)_{il}(M_2)_{jl}$$

$$A^TKA = \sum_{k,l} \sum_i A_i(M_1)_{ik}(M_2)_{il} \sum_j A_j(M_1)_{jk}(M_2)_{jl}$$

$$A^TKA = \sum_{k,l} (\sum_i A_i(M_1)_{ik}(M_2)_{il})^2 \geq 0$$

This prove that the matrix K is positive semi definite.

Question 3

Using notations in the exercise, for a given k_n being a p.d kernel implies that it's associated gram matrix is positive semi definite: $\forall A \in \mathbb{R}^n$:

$$A^TK_nA = \sum_{i,j} A_iA_jK_n(x_i, x_j) \geq 0$$

$$lim_{n\to\infty}A^TK_nA = A^TKA \geq 0$$

So k is a p.d. kernel

Question 4

The Taylor decomposition of exponential gives us

$$e^{K_1}(x, y) = lim_{N\to\infty} \sum_{n=0}^{N} \frac{1}{n!} k_1(x, y)^n$$

- $k_1(x, y)^n$ are p.d. kernels as they are products of positive definite kernels (c.f. Question 2)
- $\frac{1}{n!}k_1(x, y)^n$, we have a positive definite kernel multiplied by a positive constant. The resulting kernel will also be p.d. (it is easy to see that the associated gram matrix is positive semi definite).
- $\sum_{n=0}^{N} \frac{1}{n!}k_1(x, y)^n$ are p.d. kernels as they are sums of positive definite kernels (c.f. Question 1)
- e^{K_1} finally is p.d. because it is limit a p.d kernel sequence.

Exercise 2

Lemme 1. *The largest entry of a symmetric positive semi-definite matrix A is on its diagonal. If an diagonal coefficient is 0 then all coefficients of the corresponding row and column are equal to 0 too.*

Proof: Let us suppose that the largest entry (strictly) is $A_{i,j}$ not on the diagonal, so $i \neq j$. Let $x = e_i - e_j$.

$$x^T Ax = A_{i,i} - 2A_{i,j} + A_{j,j} = (A_{i,i} - A_{i,j}) + (A_{j,j} - A_{j,i})$$

Since $A_{i,j}$ is the largest element, $A_{i,i} - A_{i,j} < 0$ and $A_{j,j} - A_{j,i} < 0$ hence $x^T Ax < 0$. This is not possible since A is positive.
Let $u = se_i - e_j$, $u^T Au = s^2 A_{i,i} - 2sA_{i,j} + A_{j,j}$. If $A_{i,i} = 0$ then $u^T Au = -2sA_{i,j} + A_{j,j} < 0$ for large enough values of s.

Lemme 2. *If k is a p.d. kernel an $f : \mathbb{X} \to \mathbb{R}$, $k'(x, y) = f(x)k(x, y)f(y)$ is also a p.d. kernel.*

Proof: Let's Φ be the feature associated with k. We can easily check that $k'(x, y) = < f(x)\Phi(x), f(y)\Phi(y) >$ and use the fact that $\Phi'(x) = f(x).\Phi(x)$ is the feature map associated with k' to complete the proof.

Question 1

$K(x, y) = \frac{1}{1-xy}$ is p.d. since the diagonal coefficients are infinite. We can then apply lemma 1.

Question 2

$K(x, y) = 2^{xy} = e^{xy\ln(2)}$: $K_0(x, y) = xy\ln(2)$ is a psd kernel (using exercise 1.1 and the fact that $\ln(2) > 0$), so $e^{K_0(x,y)}$ is psd.

Question 3

$K(x, y) = log(1 + xy)$. Let's take two points $x = 1$ and $y = \epsilon$, the resulting gram matrix will be:

$$K = \begin{bmatrix} log(2) & log(1+\epsilon) \\ log(1+\epsilon) & log(1+\epsilon^2) \end{bmatrix}$$

$$det(K) = log(2)log(1 + x^2) - log(1 + x)^2$$

By taking the first order approximation of log around 0 we have,
$det(K) = log(2)x^2 - x^2 < 0$ which is true as $log(2) < 1$
Or K being a $2 * 2$ matrix $det(K)$ is equal to the product of his 2 eigen values. so for ϵ small enough K is not semi positive definite which means that the kernel k is not p.d.
For example we can take $\epsilon = 0.01$, we will have $det(K) = -21$

Question 4

$K : \mathbb{R}_+^2 \rightarrow \mathbb{R}$ with $K(x,y) = \mathrm{e}^{-(x-y)^2} = \mathrm{e}^{-x^2}\,\mathrm{e}^{2xy}\,\mathrm{e}^{-y^2}$

By giving $f(x) = \mathrm{e}^{-x^2}$ and $K_2 = \mathrm{e}^{2xy}$, we see that $K(x,y) = f(x)K_2(x,y)f(y)$ Thanks to the exercice 1, we can see that K_2 is a p.d. kernel. Using lemma 2 we complete the proof.

Finally, we can see that K is a p.d. kernel.

Question 5

$K : \mathbb{R}^2 \rightarrow \mathbb{R}$ with $K(x,y) = cos(x+y)$
Let's take 2 points with $x = \frac{\pi}{4}$ and $y = 2\pi$. The gram matrix will be:

$$K = \begin{bmatrix} 0 & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & 1 \end{bmatrix}$$

We can either apply lemma 1 or see that $det(k) = -\frac{1}{2}$ to prove that K is not positive semi definite.
Therefore we can conclude that k is not p.d.

Question 6:

We have ,

$$\begin{aligned} cos(x-y) &= cos(x)cos(y) + sin(x)sin(y) \\ &= k_1(x,y) + k_2(x,y) \end{aligned}$$

Using lemma 2, $k_1(x,y) = cos(x)cos(y)$, we can check the k_1 is p.d.
The same way, we can check that $k_2(x,y) = sin(x)sin(y)$
We finally use the statement in exercise 1.1 to complete the proof that $cos(x,y)$ is a p.d. kernel.

Question 7:

$K : \mathbb{R}_+^2 \rightarrow \mathbb{R}$ with $K(x,y) = min(x,y)$

For more simplicity, we will order our x_i

$0 = x_0 \leq x_1 < ... < x_n$

We have

$$\begin{aligned} \sum_{i,j=1}^{n} a_i a_j K(x_i, x_j) &= \sum_{i=j} a_i^2 x_i + \sum_{i \neq j} a_i a_j K(x_i, x_j) \\ &= \sum_{j=1}^{n} \lambda_i^2 (x_j - x_{j-1}),\ with\ \lambda_i = \sum_{i=j}^{n} a_i \end{aligned}$$

This sum is positive, because $\forall j,\ x_j - x_{j-1} > 0$.
So this kernel is p.d.

Question 8:

Let's take 2 points with $x = 0$ and $y = 1$. The gram matrix will be:

$$K = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$$

We can either apply lemma 1 or see that $det(k) = -1$ to prove that K is not positive semi definite.
Therefore we can conclude that k is not p.d.

Question 9:

We have,

$$\frac{1}{max(x,y)} = min(\frac{1}{x}, \frac{1}{y})$$

so,

$$\begin{aligned} k(x,y) &= \frac{min(x,y)}{max(x,y)} \\ &= min(x,y) * \frac{1}{max(x,y)} \end{aligned}$$

In question 7 we have proven that $K : \mathbb{R}^2 \to \mathbb{R}$ with $K(x,y) = min(x,y)$ is a p.d. kernel. Using the statement in exercise 1.2, we can prove the k is s p.d kernel.

Question 10:

Let's have the integer factorization of x and y: $x = 2^{a_1}.3^{a_2}.5^{a_3}\dots$ and $y = 2^{b_1}.3^{b_2}.5^{b_3}\dots$.
So we have

$$\begin{aligned} k(x,y) &= GCD(x,y) \\ &= 2^{min(a_1,b_1)}.3^{min(a_2,b_2)}.5^{min(a_3,b_3)}\dots \\ &= \prod_i Prime_i^{min(a_i,b_i)} \end{aligned}$$

In question 7 we have proven that $K : \mathbb{R}^2 \to \mathbb{R}$ with $K_0(a_i, b_i) = min(a_i, b_i)$ is a p.d. kernel.
Following the proof in question 2, we can prove that $K : \mathbb{R}^2 \to \mathbb{R}$ with $K_1(x,x) = Prime_i^{k_0(a_i,b_i)}$ is a p.d. kernel.
Using the statement in exercise 1.2, we can prove the k is s p.d kernel.

Question 11:

$K : \mathbb{N}^2 \to \mathbb{R}$ with $K(x,y) = LCM(x,y)$

Let's take $x_1 = 2, x_2 = 1, a_1 = 1$ and $a_2 = -2$.
$\sum\limits_{i,j=1}^{2} a_i a_j K(x_1, x_2) = 1 \times 2 + 2 \times (-2) \times 2 + 4 \times 1 = -2 < 0.$

Question 12:

Using the notation in question 10 and the integer factorisation of x and y, we have

$$\begin{aligned} k(x,y) &= GCD(x,y)/LCM(x,y) \\ &= \frac{\prod_i Prime_i^{min(a_i,b_i)}}{\prod_i Prime_i^{max(a_i,b_i)}} \\ &= \prod_i Prime_i^{min(a_i,b_i)-max(a_i,b_i)} \end{aligned}$$

So K is not a p.d. kernel.

Exercise 3

Question 1 :

$$K(a,b) = a.b \to f(x) = \sum_i \lambda_i x_i.x \ with \ x \in \mathbb{R} \ , \ f \in \mathbb{H}_k$$

$$f(x) = \lambda x \ with \ \lambda = \sum_i \lambda_i x_i \ and \ ||f|| = |\lambda|$$

The same way apply for g: $g(y) = \beta y \ with \ \beta = \sum_i \beta_i y_i$
The criterion can then be written as :

$$C_n^K(X,Y) = \max_{\lambda,\beta \in [-1,1]} Cov_n(\lambda X, \beta Y)$$

$$C_n^K(X,Y) = \max_{\lambda,\beta \in [-1,1]} \mathbb{E}_n(\lambda X \beta Y) - \mathbb{E}_n(\lambda X).\mathbb{E}_n(\beta Y)$$

By linearity we have

$$C_n^K(X,Y) = \max_{\lambda,\beta \in [-1,1]} \lambda\beta(\mathbb{E}_n(XY) - \mathbb{E}_n(X).\mathbb{E}_n(Y))$$

$$C_n^K(X,Y) = \max_{\lambda,\beta \in [-1,1]} \lambda\beta Cov_n(X,Y)$$

giving the constraints on $\lambda \ and \ \beta$, the criterion above is maximized when $\lambda.\beta = sign(Cov_n(X,Y))$ which means :

$$f(x) = x \ or \ f(x) = -x \to f \ is \ Id \ or f \ is \ -Id$$

$$g(y) = y \ or \ g(y) = -y \to g \ is \ Id \ or g \ is \ -Id$$

Finally we will have:

$$C_n^K(X,Y) = |Cov_n(X,Y)|$$

Question 2 :

Let us suppose that the centering term disappears. We have to solve a maximization problem :

$$\max_{f,g \in B^K} \operatorname{cov}(f(X), g(Y))$$

This problem can be rewritten as a maximization problem on the lagrangian :

$$\max \operatorname{cov}(f(X), g(Y)) - c_1(\|f\|_{B^K} - 1) - c_2(\|g\|_{B^K} - 1)$$

A representer theorem can be applied since we want :

$$\max_{f,g \in B^K} (f(x_1), ..., f(x_n), g(y_1), ...g(y_n), \|f\|_{B^K}, \|g\|_{B^K})$$

(we can consider $f \otimes g \in B^K \otimes B^K$ to do a proper use of the representer theorem). This gives us :

$$f(x) = \sum_i \alpha_i K(x_i, x)$$

$$g(y) = \sum_j \lambda_j K(y_j, y)$$

Then, using the linearity of *cov* shown in question 1,

$$\begin{aligned} \operatorname{cov}(f(X), g(Y)) &= \operatorname{cov}\left(\sum_i \alpha_i K(x_i, X), \sum_j \lambda_j K(y_j, Y)\right) \\ &= \sum_i \sum_j \alpha_i \lambda_j \operatorname{cov}(K(x_i, X), K(y_j, Y)) \end{aligned}$$

Hence,

$$\begin{aligned} \operatorname{cov}(K(x_i, X), K(y_j, Y)) &= \frac{1}{n} \sum_k K(x_i, x_k) K(y_j, y_k) - \frac{1}{n} \sum_k K(x_i, x_k) \frac{1}{n} \sum_k K(y_j, y_k) \\ &= \frac{1}{n} \sum_k G^X_{i,k} G^Y_{k,j} - \overline{K_X}.\overline{K_Y} \\ &= \frac{1}{n} [G^X G^Y]_{i,j} - \overline{K_X}.\overline{K_Y} \end{aligned}$$

since $K(y_j, y_k) = K(y_k, y_j)$ and where G^X and G^Y are the Gram matrices of X and Y in that order. So,if we supposed that the centering term dissapear

$$\operatorname{cov}(f(X), g(Y)) = \frac{1}{n} \sum_i \sum_j \alpha_i \lambda_j [G^X G^Y]_{i,j} \qquad = \frac{1}{n} \alpha^T G^X G^Y \lambda$$

We have $\|f\|_{B^K} = \alpha^T G^X \alpha$ and $\|g\|_{B^K} = \lambda^T G^Y \lambda$ Let w be an eigen vector of $G^X G^Y$ associated to its largest eigenvalue. Then $\alpha = \lambda = w$ maximizes the covariance under constraints $w^T G^X w = 1$ and $w^T G^Y w = 1$ where w has been re-normalized to satisfy these constraints.

$$\operatorname{cov}(f(X), g(Y)) = \frac{1}{n} \|w\|^2 \left\|G^X G^Y\right\|_{\text{spec}}$$

where $\|A\|_{\text{spec}}$ is the largest absolute value of the eigenvalues of A.

Exercise 4

Question 1

Let us assume that ϕ is a Lipschiz function.

$\mid R_\phi(f,x) - R_\phi(g,x) \mid = \mid \phi(f(x)) - \phi(g(x)) + \lambda(\|f\|^2_{\mathcal{H}_K} - \|g\|^2_{\mathcal{H}_K}) \mid$

With the triangular inequality on the norm $\|.\|_{H_K}$
$\mid R_\phi(f,x) - R_\phi(g,x) \mid \leq \mid \phi(f(x)) - \phi(g(x)) \mid + \lambda(\|f\|_{\mathcal{H}_K} - \|g\|_{\mathcal{H}_K})(\|f\|_{\mathcal{H}_K} + \|g\|_{\mathcal{H}_K})$

With this inequality: $\|f\|_{\mathcal{H}_K} - \|g\|_{\mathcal{H}_K} \leq \|f-g\|_{\mathcal{H}_K}$
We can summaries with the fact that ϕ is a Lipschiz function.

So $\mid R_\phi(f,x) - R_\phi(g,x) \mid \leq (2\lambda R + L)\|f-g\|_{\mathcal{H}_K}$.
So we find the inequality with $C_1 = 2\lambda R + L$.

Question 2

We know that ϕ is a convex function. And f_x the minimizer of R at the point x.

Let us suppose that $\forall C_2 > 0$, $\psi(f,x) < C_2\|f - f_x\|^2_{\mathcal{H}_K}$.

This supposition is equivalent to: $\psi(f,x) < 0$

But it means that $R_\phi(f,x) - R_\phi(f_x,x) < 0$

It is impossible, because f_x minimize $R_\phi(.,x)$.

So $\exists C_2 > 0$ such that $\psi(f,x) \geq C_2\|f - f_x\|^2_{\mathcal{H}_K}$.

Question 3

We will use the 2 previous answers.
ϕ is L-lipschitz, so $\mid \psi(f,x) \mid^2 \leq C_1^2\|f - f_x\|^2_{\mathcal{H}_K}$.

It means that $\mathbb{E}(\psi(f,x)^2) \times C_1^{-2} \leq \mathbb{E}(\|f - f_x\|^2_{\mathcal{H}_K})$.

Or, thanks to the Jensen's formula, we have $E(\psi(f,x)^2) \geq E(\psi(f,x))^2$.

Then, $\mathbb{E}(\psi(f,x))^2 \times C_1^{-2} \leq \mathbb{E}(\|f - f_x\|^2_{\mathcal{H}_K})$.

ϕ is convex, so $\mathbb{E}(\psi(f,x)) \times C_2^{-1} \geq \mathbb{E}(\|f - f_x\|^2_{\mathcal{H}_K})$.

By computation, we have $\mathbb{E}(\psi(f,x))^2 \leq \frac{C_1^2}{C_2}\mathbb{E}(\psi(f,x))$.

So we have $\mathbb{E}(\psi(f,x))^2 \leq C\mathbb{E}(\psi(f,x))$ with $C = \frac{C_1^2}{C_2}$.