Kernel Methods in Machine Learning - Homework

Tamim EL AHMAD - MVA - elahmad.tamim@gmail.com

2020/02/26

Exercise 1. Kernels

1. Let $\mathcal{X}$ be a set, $f, g : \mathcal{X} \to \mathbb{R}$ two non-negative functions :

$$\forall x, y \in \mathcal{X}, K(x, y) = min(f(x)g(y), f(y)g(x))$$

First, K is clearly symmetric.
Then, $\forall x, y \in \mathcal{X}$, such that $g(x) \neq 0$, $g(y) \neq 0$:

$$\begin{aligned} K(x, y) &= g(x)g(y)min(\frac{f(x)}{g(x)}, \frac{f(y)}{g(y)}) \quad \text{because} \quad g(x) > 0 \quad \text{and} \quad g(y) > 0 \\ &= K_1(x, y)K_2(x, y) \end{aligned}$$

With $K_1(x, y) = g(x)g(y)$ and $K_2(x, y) = min(\frac{f(x)}{g(x)}, \frac{f(y)}{g(y)})$ two kernels.

First, K_1 is p.d. because $(x, y) \mapsto xy$ is p.d. on $\mathbb{R}_+$ and $g \geq 0$.

Then, let's show that $(x, y) \mapsto min(x, y)$ is a p.d. kernel on $\mathbb{R}_+$:
$(x, y) \mapsto min(x, y)$ is clearly symmetric.
Let note that $\forall x, y \in \mathbb{R}_+, min(x, y) = \int_{\mathbb{R}_+} \mathbb{1}_{t \leq x} \mathbb{1}_{t \leq y} dt$.
Let $x_1, ..., x_N \in \mathbb{R}_+$ and $a_1, ..., a_N \in \mathbb{R}$:

$$\begin{aligned} \sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j min(x_i, x_j) &= \sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j \int_{\mathbb{R}_+} \mathbb{1}_{t \leq x_i} \mathbb{1}_{t \leq x_j} dt \\ &= \int_{\mathbb{R}_+} (\sum_{i=1}^{N} a_i \mathbb{1}_{t \leq x_i})^2 dt \geq 0 \end{aligned}$$

Back to the problem, we then have that K_2 is p.d. because $(x, y) \mapsto min(x, y)$ is p.d. on $\mathbb{R}_+$ and $\frac{f}{g} \geq 0$.

As $K = K_1 K_2$, K is p.d. on $\mathcal{X} \setminus \{x \in \mathcal{X}, g(x) = 0\}$.

Let's generalize it to $\mathcal{X}$:
Let $x_1, ..., x_N \in \mathcal{X}$ and $a_1, ..., a_N \in \mathbb{R}$. We assume there are some $i \in \{1, ..., N\}$ such that $g(x_i) = 0$.

We sort the x_i such that, for a $n \in \{1, ..., N\}$, $g(x_1) \neq 0, ..., g(x_n) \neq 0$ and $g(x_{n+1}) = ... = g(x_N) = 0$, and we note that if $g(x) = 0$ or $g(y) = 0$, $K(x, y) = 0$ because $f \geq 0$ and $g \geq 0$.

$$\sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j K(x_i, x_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j K(x_i, x_j) \geq 0 \quad \text{because K is p.d. on} \quad \mathcal{X} \setminus \{x \in \mathcal{X}, g(x) = 0\}$$

Finally, K is p.d. on $\mathcal{X}$.

2. Given a non-empty finite set E, on $\mathcal{X} = \mathcal{P}(E) = A : A \subset E$:

$$\forall A, B \subset E, K(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

We note that if A or B is empty, $K(A, B) = 0$. So, similarly to the previous question, if K is p.d. on $\mathcal{X} = \mathcal{P}(E) \setminus \{\emptyset\}$, then K is p.d. on $\mathcal{X} = \mathcal{P}$.

So let $\mathcal{X} = \mathcal{P}(E) \setminus \{\emptyset\}$. Let μ denote the counting measure on E and consider the space of measurable functions from $(E, \mathcal{P}(E))$ to $([0, 1], \mathcal{B}([0, 1]))$. We consider, on this space, the bilinear form $\langle f, g\rangle = \int f g d\mu$. It is non-negative : $\langle f, f\rangle = \int f^2 d\mu \geq 0$. We note that :

$$|A \cap B| = \mu(A \cap B) = \int \mathbb{1}_A \mathbb{1}_B d\mu = \langle \mathbb{1}_A, \mathbb{1}_B\rangle$$

Then $(A, B) \mapsto |A \cap B|$ is a p.d. kernel on $\mathcal{X}$.
Then :

$$\frac{1}{|A \cup B|} = \frac{1}{|E| - |A^c \cap B^c|} = \frac{1}{|E|}\frac{1}{1 - \frac{|A^c \cap B^c|}{|E|}}$$

As A and B are non-empty, $A^c \subsetneq E$ and $B^c \subsetneq E$, and so $0 < \frac{|A^c \cap B^c|}{|E|} < 1$.
Thus :

$$\frac{1}{|A \cup B|} = \frac{1}{|E|}\sum_{k=0}^{\infty}(\frac{|A^c \cap B^c|}{|E|})^k$$

Since $|A^c \cap B^c| = \langle \mathbb{1}_{A^c}, \mathbb{1}_{B^c}\rangle$, $(A, B) \mapsto |A^c \cap B^c|$ is a p.d. kernel.
So $(A, B) \mapsto \frac{|A \cap B|}{|E|}$ since $\frac{1}{|E|} > 0$.
Thus, $(A, B) \mapsto (\frac{|A \cap B|}{|E|})^k$ is a p.d. kernel as product of p.d. kernels.
Thus, $(A, B) \mapsto \sum_{k=0}^{K}(\frac{|A \cap B|}{|E|})^k$, for $K \geq 0$, is a p.d. kernel as a sum of p.d. kernels.
Thus, $(A, B) \mapsto \sum_{k=0}^{\infty}(\frac{|A \cap B|}{|E|})^k$ is a p.d. kernel as a limit of a sequence of p.d. kernels.
Finally, $(A, B) \mapsto \frac{1}{|A \cup B|}$ is a p.d. kernel on $\mathcal{X}$.

So K is a p.d. kernel on $\mathcal{X} = \mathcal{P}(E) \setminus \{\emptyset\}$.

And finally, K is a p.d. kernel on $\mathcal{X} = \mathcal{P}(E)$.

Exercise 2. Kernels encoding equivalence classes.

$\Longrightarrow$: Let K be p.d.
Let $x, x', x'' \in \mathcal{X}$:

— K is p.d. $\Longrightarrow K(x,x') = K(x',x) \Longrightarrow (K(x,x') = 1 \iff K(x',x) = 1)$

— We assume $K(x,x') = K(x',x'') = 1$. For all $a, a', a'' \in \mathbb{R}$:

$$\begin{aligned} C &= a^2K(x,x) + aa'K(x,x') + aa''K(x,x'') \\ &\quad + a'aK(x',x) + a'^2K(x',x') + a'a''K(x',x'') \\ &\quad + a''aK(x'',x) + a''a'K(x'',x') + a''^2K(x'',x'') \\ &\quad = a^2 + a'^2 + a''^2 + 2aa' + 2a'a'' + 2aa''K(x,x'') \\ C &= (a+a')^2 + (a'+a'')^2 - a'^2 + 2aa''K(x,x'') \geq 0 \quad \text{since K is p.d.} \end{aligned}$$

We assume by contradiction that $K(x,x'') = 0$. Then, with $a' = 2$, $a = -2$ and $a'' = -1$:

$$C = -3 < 0 \Longrightarrow \text{contradiction} \Longrightarrow K(x,x'') = 1$$

Moreover, we note that with $K(x,x'') = 1$:

$$C = (a + a' + a'')^2 \geq 0$$

Thus $K(x,x'') = 1$.

$\Longleftarrow$:
Let $x, x' \in \mathcal{X}$:

— $K(x,x') = 1 \iff K(x',x) = 1$, so $K(x,x') = 0 \iff K(x',x) = 0$ too. And so $K(x,x') = K(x',x)$

— Let $x_1, ..., x_N \in \mathcal{X}$ and $a_1, ..., a_N \in \mathbb{R}$. K define an equivalence relation : $x \sim x' \iff K(x,x') = 1$. So let sort the x_i by equivalence classes : if x_i and x_j are in the same equivalence class, then $K(x_i,x_j) = 1$. On the contrary, if x_i and x_j are in 2 different equivalence classes, then $K(x_i,x_j) = 0$. We suppose there are r equivalence classes in $\{x_1, ..., x_N\}$, and we note $c_1, ..., c_r$ subsets of $\{1, ..., N\}$ partitioning it and designing the indices of each equivalence class.

$$\begin{aligned} \sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j K(x_i,x_j) &= \sum_{i=1}^{N} a_i^2 + 2\sum_{1\leq i<j\leq N} a_i a_j K(x_i,x_j) \\ &= \sum_{k=1}^{r}\sum_{i\in c_k} a_i^2 + 2\sum_{k=1}^{r}\sum_{i,j\in c_k, i<j} a_i a_j \\ \sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j K(x_i,x_j) &= \sum_{k=1}^{r}(\sum_{i\in c_k} a_i)^2 \geq 0 \end{aligned}$$

Thus K is p.d.

Exercise 3. COCO

1. Let $K(a,b) = ab$ be the linear kernel. Then , its RKHS is $\mathcal{H} = \{f_y(x) = xy : y \in \mathbb{R}\}$ and $\forall y$, $\|f_y\|_{\mathcal{H}} = |y|$.
Let $f, g \in \mathcal{H}$, $\exists u, v \in \mathbb{R}$ such that $f : x \in \mathbb{R} \mapsto ux$ and $g : y \in \mathbb{R} \mapsto vy$. So :

$$cov_n(f(X), g(Y)) = \frac{\sum_{i=1}^n ux_i vy_i}{n} - \frac{(\sum_{i=1}^n ux_i)(\sum_{j=1}^n vy_j)}{n^2}$$

So :

$$\begin{aligned}
C_n^K(X,Y) &= \max_{u,v\in\mathbb{R},|u|\leq 1,|v|\leq 1}\left[\frac{uv\sum_{i=1}^n x_i y_i}{n} - \frac{uv(\sum_{i=1}^n x_i)(\sum_{j=1}^n y_j)}{n^2}\right] \\
&= \max_{u,v\in\mathbb{R},|u|\leq 1,|v|\leq 1}\left[\frac{uv}{n}\left(\sum_{i=1}^n x_i y_i - \frac{(\sum_{i=1}^n x_i)(\sum_{j=1}^n y_j)}{n}\right)\right] \\
&= \frac{1}{n}\left|\sum_{i=1}^n x_i y_i - \frac{1}{n}\left(\sum_{i=1}^n x_i\right)\left(\sum_{j=1}^n y_j\right)\right| \\
&= \frac{1}{n}|X^T Y - \frac{1}{n}X^T \mathbb{1}_n \mathbb{1}_n^T Y| \quad \text{with} \quad \mathbb{1}_n = \underbrace{(1, ..., 1)^T}_{n \text{ times}} \\
C_n^K(X,Y) &= \frac{1}{n}|X^T(I_n - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T)Y|
\end{aligned}$$

2. Let $f, g \in \mathcal{H}$:

$$\begin{aligned}
C_n^K(X,Y) &= \max_{f,g\in\mathcal{B}_K} cov_n(f(X), g(Y)) \\
&= \max_{f,g\in\mathcal{B}_K} \frac{\sum_{i=1}^n f(x_i)g(y_i)}{n} - \frac{(\sum_{i=1}^n f(x_i))(\sum_{j=1}^n g(y_j))}{n^2} \\
C_n^K(X,Y) &= \max_{g\in\mathcal{B}_K}\left(\max_{f\in\mathcal{B}_K}\frac{\sum_{i=1}^n f(x_i)g(y_i)}{n}\right) - \frac{(\sum_{i=1}^n f(x_i))(\sum_{j=1}^n g(y_j))}{n^2}
\end{aligned}$$

For all $g \in \mathcal{B}_K$, let first show that we can restrict this problem to $\mathcal{S} = Span(K_{X_{x_1}}, ..., K_{X_{x_n}})$ for f. First, let's note that, $\forall f \in \mathcal{H}$, we can write f as : $f = f_{\mathcal{S}} + f_{\perp}$ with $f_{\mathcal{S}} \in \mathcal{S}$ and $f_{\perp} \in \mathcal{S}^{\perp}$ (so $\forall i \in \{1, ..., n\}, \langle f_{\perp}, K_{X_{x_i}}\rangle = 0$). Thus, $\forall g \in \mathcal{B}_K$:

$$\begin{aligned}
cov_n(f(X), g(Y)) &= \frac{\sum_{i=1}^n f(x_i)g(y_i)}{n} - \frac{(\sum_{i=1}^n f(x_i))(\sum_{j=1}^n g(y_j))}{n^2} \\
&= \frac{\sum_{i=1}^n \langle f, K_{X_{x_i}}\rangle g(y_i)}{n} - \frac{(\sum_{i=1}^n \langle f, K_{X_{x_i}}\rangle)(\sum_{j=1}^n g(y_j))}{n^2} \\
&= \frac{\sum_{i=1}^n \langle f_{\mathcal{S}} + f_{\perp}, K_{X_{x_i}}\rangle g(y_i)}{n} - \frac{(\sum_{i=1}^n \langle f_{\mathcal{S}} + f_{\perp}, K_{X_{x_i}}\rangle)(\sum_{j=1}^n g(y_j))}{n^2} \\
&= \frac{\sum_{i=1}^n \langle f_{\mathcal{S}}, K_{X_{x_i}}\rangle g(y_i)}{n} - \frac{(\sum_{i=1}^n \langle f_{\mathcal{S}}, K_{X_{x_i}}\rangle)(\sum_{j=1}^n g(y_j))}{n^2} \\
cov_n(f(X), g(Y)) &= cov_n(f_{\mathcal{S}}(X), g(Y))
\end{aligned}$$

Then, since, by Pytagora's theorem, $\forall f \in \mathcal{H}, \|f\|^2_{\mathcal{H}} = \|f_{\mathcal{S}}\|^2_{\mathcal{H}} + \|f_{\perp}\|^2_{\mathcal{H}}$, and so $\|f\|_{\mathcal{H}} \geq \|f_{\mathcal{S}}\|_{\mathcal{H}}$, looking for a *max* of $cov_n(f(X), g(Y))$ such that $\|f\|_{\mathcal{H}} \leq 1$ is the same as looking for it on $Span(K_{X_{x_1}}, ..., K_{X_{x_n}})$. So, for all $g \in \mathcal{B}_K$, it admits a solution of the form :

$$\forall x \in \mathbb{R}, \hat{f}(x) = \sum_{i=1}^{n} \alpha_i K_X(x_i, x) \quad \text{with } \alpha \in \mathbb{R}^n,\ K_X \text{ the gram matrix of } X\text{, such that } \|\hat{f}\|_{\mathcal{H}} = \alpha^T K_X \alpha \leq 1$$

Thus, similarly, the problem $\underset{g \in \mathcal{B}_K}{min} \frac{(\sum_{i=1}^n \hat{f}(x_i))(\sum_{j=1}^n g(y_j))}{n^2} - \frac{\sum_{i=1}^n \hat{f}(x_i) g(y_i)}{n}$ admits a solution of the form :

$$\forall y \in \mathbb{R}, \hat{g}(y) = \sum_{j=1}^{n} \beta_j K_Y(y_j, y) \quad \text{with } \beta \in \mathbb{R}^n,\ K_Y \text{ the gram matrix of } Y\text{, such that } \|\hat{g}\|_{\mathcal{H}} = \beta^T K_Y \beta \leq 1$$

So finally :

$$\begin{aligned}
C_n^K(X,Y) &= \underset{\alpha^T K_X \alpha \leq 1, \beta^T K_Y \beta \leq 1}{max} \frac{1}{n} \sum_{i=1}^{n} [\alpha^T K_X]_i [K_Y \beta]_i - \frac{1}{n^2} \alpha^T K_X \mathbb{1}_n \mathbb{1}_n^T K_Y \beta \\
&= \underset{\alpha^T K_X \alpha \leq 1, \beta^T K_Y \beta \leq 1}{max} \frac{1}{n} \alpha^T K_X K_Y \beta - \frac{1}{n^2} \alpha^T K_X \mathbb{1}_n \mathbb{1}_n^T K_Y \beta \\
&= \underset{\alpha^T K_X \alpha \leq 1, \beta^T K_Y \beta \leq 1}{max} \frac{1}{n} \alpha^T K_X^{\frac{1}{2}} K_X^{\frac{1}{2}} K_Y^{\frac{1}{2}} K_Y^{\frac{1}{2}} \beta - \frac{1}{n^2} \alpha^T K_X^{\frac{1}{2}} K_X^{\frac{1}{2}} \mathbb{1}_n \mathbb{1}_n^T K_Y^{\frac{1}{2}} K_Y^{\frac{1}{2}} \beta \\
C_n^K(X,Y) &= \underset{\|K_X^{\frac{1}{2}} \alpha\|_2 \leq 1, \|K_Y^{\frac{1}{2}} \beta\|_2 \leq 1}{max} \frac{1}{n} (K_X^{\frac{1}{2}} \alpha)^T K_X^{\frac{1}{2}} (I_n - \frac{1}{n} \mathbb{1}_n \mathbb{1}_n^T) K_Y^{\frac{1}{2}} K_Y^{\frac{1}{2}} \beta
\end{aligned}$$

Since K_X and K_Y are symmetric positive semi-definite, they both admit a square root symmetric positive semi-definite too, and $\alpha^T K_X \alpha = \alpha^T K_X^{\frac{1}{2}T} K_X^{\frac{1}{2}} \alpha = \|K_X^{\frac{1}{2}} \alpha\|_2^2$ and $\beta^T K_Y \beta = \beta^T K_Y^{\frac{1}{2}T} K_Y^{\frac{1}{2}} \beta = \|K_Y^{\frac{1}{2}} \beta\|_2^2$.

Let show that this is equivalent to $\underset{\alpha \in \mathbb{R}^n, \beta \in \mathbb{R}^n}{max} \frac{1}{n} \alpha^T K_X^{\frac{1}{2}} (I_n - \frac{1}{n} \mathbb{1}_n \mathbb{1}_n^T) K_Y^{\frac{1}{2}} \beta$ subject to $\|\alpha\|_2 \leq 1$ and $\|\beta\|_2 \leq 1$:

- $\Longrightarrow$: for any $\alpha, \beta \in \mathbb{R}^n$ such that $\|K_X^{\frac{1}{2}} \alpha\|_2 \leq 1$ and $\|K_Y^{\frac{1}{2}} \beta\|_2 \leq 1$, with $\bar{\alpha} = K_X^{\frac{1}{2}} \alpha$ and $\bar{\beta} = K_Y^{\frac{1}{2}} \beta$, we have $\|\bar{\alpha}\|_2 \leq 1$, $\|\bar{\beta}\|_2 \leq 1$ and $\frac{1}{n} (K_X^{\frac{1}{2}} \alpha)^T K_X^{\frac{1}{2}} (I_n - \frac{1}{n} \mathbb{1}_n \mathbb{1}_n^T) K_Y^{\frac{1}{2}} K_Y^{\frac{1}{2}} \beta = \frac{1}{n} \bar{\alpha}^T K_X^{\frac{1}{2}} (I_n - \frac{1}{n} \mathbb{1}_n \mathbb{1}_n^T) K_Y^{\frac{1}{2}} \bar{\beta}$
- $\Longleftarrow$: $K_X^{\frac{1}{2}}$ and $K_Y^{\frac{1}{2}}$ are symmetric, so they are diagonalizable in an orthogonal basis. So, for any $\bar{\alpha}, \bar{\beta} \in \mathbb{R}^n$ such that $\|\bar{\alpha}\|_2 \leq 1$ and $\|\bar{\beta}\|_2 \leq 1$, we can write $\bar{\alpha} = K_X^{\frac{1}{2}} \alpha + k_X$ and $\bar{\beta} = K_Y^{\frac{1}{2}} \beta + k_Y$ for some vectors α, k_X, β, k_Y such that $K_X^{\frac{1}{2}} k_X = K_Y^{\frac{1}{2}} k_Y = 0$ and so $\langle k_X, K_X^{\frac{1}{2}} \alpha \rangle = \alpha^T K_X^{\frac{1}{2}T} k_X = \alpha^T K_X^{\frac{1}{2}} k_X = 0$ and similarly $\langle k_Y, K_Y^{\frac{1}{2}} \beta \rangle = 0$. So, by orthogonality, $\|K_X^{\frac{1}{2}} \alpha\|_2^2 = \|\bar{\alpha}\|_2^2 - \|k_X\|_2^2 \leq 1 - \|k_X\|_2^2 \leq 1$, so $\|K_X^{\frac{1}{2}} \alpha\|_2 \leq 1$ and similarly $\|K_Y^{\frac{1}{2}} \beta\|_2 \leq 1$. Finally, $\frac{1}{n} \bar{\alpha}^T K_X^{\frac{1}{2}} (I_n - \frac{1}{n} \mathbb{1}_n \mathbb{1}_n^T) K_Y^{\frac{1}{2}} \bar{\beta} = \frac{1}{n} (K_X^{\frac{1}{2}} \alpha + k_X)^T K_X^{\frac{1}{2}} (I_n - \frac{1}{n} \mathbb{1}_n \mathbb{1}_n^T) K_Y^{\frac{1}{2}} (K_Y^{\frac{1}{2}} \beta + k_Y) = \frac{1}{n} (K_X^{\frac{1}{2}} \alpha)^T K_X^{\frac{1}{2}} (I_n - \frac{1}{n} \mathbb{1}_n \mathbb{1}_n^T) K_Y^{\frac{1}{2}} K_Y^{\frac{1}{2}} \beta$

Finally, we have :

$$\begin{aligned}
C_n^K(X,Y) &= \underset{\|K_X^{\frac{1}{2}}\alpha\|_2\leq 1, \|K_Y^{\frac{1}{2}}\beta\|_2\leq 1}{max} \frac{1}{n}(K_X^{\frac{1}{2}}\alpha)^T K_X^{\frac{1}{2}}(I_n - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T)K_Y^{\frac{1}{2}}K_Y^{\frac{1}{2}}\beta \\
&= \underset{\|\alpha\|_2\leq 1, \|\beta\|_2\leq 1}{max} \frac{1}{n}\alpha^T K_X^{\frac{1}{2}}(I_n - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T)K_Y^{\frac{1}{2}}\beta \\
&= \underset{\|\beta\|_2\leq 1}{max}\underset{\|\alpha\|_2\leq 1}{max} \alpha^T \frac{1}{n}K_X^{\frac{1}{2}}(I_n - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T)K_Y^{\frac{1}{2}}\beta \\
C_n^K(X,Y) &= \underset{\|\beta\|_2\leq 1}{max} \|\frac{1}{n}K_X^{\frac{1}{2}}(I_n - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T)K_Y^{\frac{1}{2}}\beta\|_2 \quad \text{by Cauchy-Schwartz}
\end{aligned}$$

Finally, we recognize the spectral norm $\|.\|_2$:

$$C_n^K(X,Y) = \frac{1}{n}\|K_X^{\frac{1}{2}}(I_n - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T)K_Y^{\frac{1}{2}}\|_2$$

Exercise 4. Dual coordinate ascent algorithms for SVMs

1. For all $j \in \{1, ..., n\}$, $g : \delta \in \mathbb{R} \mapsto 2(\alpha + \delta e_j)^T y - (\alpha + \delta e_j)^T K(\alpha + \delta e_j) = 2\delta(y_j - [K\alpha]_j) - \delta^2 K_{jj} + cst$ (with e_js vectors of the usual basis of $\mathbb{R}^n$) is differentiable and concave (as $K_{jj} \geq 0$ since K is a p.d. kernel). So, we can find an eventual optimal δ^*, such that $0 \leq y_j(\alpha_j + \delta^*) \leq \frac{1}{2\lambda n}$, maximizing g as follow :

$$g'(\delta^*) = 0 \iff 2y_j - 2[K\alpha]_j - 2\delta^* K_{jj} = 0 \iff \delta^* = \frac{y_j - [K\alpha]_j}{K_{jj}}$$

Let's look at the constraints :

— If $y = -1$:

$$0 \leq -(\alpha_j + \delta) \leq \frac{1}{2\lambda n} \iff -\frac{1}{2\lambda n} - \alpha_j \leq \delta \leq -\alpha_j$$

— If $y = 1$:

$$0 \leq \alpha_j + \delta \leq \frac{1}{2\lambda n} \iff -\alpha_j \leq \delta \leq \frac{1}{2\lambda n} - \alpha_j$$

Finally $-\frac{1}{2\lambda n} - \alpha_j \leq \delta^* \leq \frac{1}{2\lambda n} - \alpha_j$ and $\delta^* = min(max(-\frac{1}{2\lambda n} - \alpha_j, \frac{y_j - [K\alpha]_j}{K_{jj}}), \frac{1}{2\lambda n} - \alpha_j)$.
So the update rule is :

$$\alpha_j^{t+1} = \alpha_j^t + \delta^*$$

2. We now consider the primal formulation of SVMs with intercept :

$$\underset{f\in\mathcal{H}, b\in\mathbb{R}}{min} \frac{1}{n}\sum_{i=1}^n max(0, 1 - y_i(f(x_i) + b)) + \lambda\|f\|_{\mathcal{H}}^2$$

$\Psi : (z_1, ..., z_n, z_{n+1}) \mapsto \underset{b\in\mathbb{R}}{min}\{\frac{1}{n}\sum_{i=1}^n max(0, 1 - y_i(z_i + b))\} + \lambda z_{n+1}^2$ is strictly increasing with respect to z_{n+1} on $\mathbb{R}_+$ with $\lambda > 0$, so we can use the representer theorem. Thus, the solution of the above problem satisfies $\hat{f}(x) = \sum_{i+1}^{n} \hat{\alpha}_i K(x_i, x)$ where $\hat{\alpha}$ solves :

$$\underset{\alpha\in\mathbb{R}^n, b\in\mathbb{R}}{min} \frac{1}{n}\sum_{i=1}^{n} max(0, 1 - y_i([K\alpha]_i + b)) + \lambda\alpha^T K\alpha$$

Introducing additional slack variables $\xi_1, ..., \xi_n \in \mathbb{R}$, the problem is equivalent to :

$$\begin{aligned}\underset{\alpha\in\mathbb{R}^n, \in\mathbb{R}^n, b\in\mathbb{R}}{min} \frac{1}{n}\sum_{i=1}^{n} \xi_i + \lambda\alpha^T K\alpha \\ \text{s. t. } \xi_i \geq 1 - y_i([K\alpha]_i + b) \\ \xi_i \geq 0\end{aligned}$$

Let's compute the Lagrangian, for $\mu, \nu \in \mathbb{R}^n$:

$$\begin{aligned}\mathcal{L}(\alpha, \xi, b, \mu, \nu) &= \frac{1}{n}\sum_{i=1}^{n} \xi_i + \lambda\alpha^T K\alpha - \sum_{i=1}^{n} \mu_i(y_i[K\alpha]_i + y_i b + \xi_i - 1) - \sum_{i=1}^{n} \nu_i \xi_i \\ &= \frac{1}{n}\xi^T \mathbb{1}_n + \lambda\alpha^T K\alpha - (diag(y)\mu)^T K\alpha - b\mu^T y - (\mu + \nu)^T \xi + \mu^T \mathbb{1}_n\end{aligned}$$

$\mathcal{L}$ is a convex quadratic function in α. It is minimized whenever its gradient is null :

$$\nabla_\alpha \mathcal{L} = 2\lambda K\alpha - K diag(y)\mu = K(2\lambda\alpha - diag(y)\mu)$$

$$\nabla_\alpha \mathcal{L} = 0 \iff \alpha = \frac{1}{2\lambda} diag(y)\mu$$

$\mathcal{L}$ is linear in ξ, then its minimum is $-\infty$ except when $\mu + \nu = \frac{1}{n}\mathbb{1}_n$.
$\mathcal{L}$ is linear in b, then its minimum is $-\infty$ except when $\mu^T y = 0$.
We therefore obtain the Lagrange dual function :

$$\begin{aligned}q(\mu, \nu) &= \underset{\alpha,\xi,b}{inf}\, \mathcal{L}(\alpha, \xi, b, \mu, \nu) \\ &= \begin{cases} \mu^T \mathbb{1}_n - \frac{1}{4\lambda}\mu^T diag(y) K diag(y)\mu & \text{if } \mu + \nu = \frac{1}{n}\mathbb{1}_n \quad \text{and} \quad \mu^T y = 0 \\ -\infty & \text{otherwise.} \end{cases}\end{aligned}$$

Thus, the dual problem is :

$$\begin{aligned}\underset{\mu\in\mathbb{R}^n, \nu\in\mathbb{R}^n}{max} \mu^T \mathbb{1}_n - \frac{1}{4\lambda}\mu^T diag(y) K diag(y)\mu \\ \text{s.t. } \mu \geq 0, \nu \geq 0 \\ \mu + \nu = \frac{1}{n}\mathbb{1}_n \\ \mu^T y = 0\end{aligned}$$

Which is equivalent to :

$$\begin{aligned}\underset{\mu\in\mathbb{R}^n}{max} \mu^T \mathbb{1}_n - \frac{1}{4\lambda}\mu^T diag(y) K diag(y)\mu \\ \text{s.t. } 0 \leq \mu \leq \frac{1}{n}\mathbb{1}_n \\ \mu^T y = 0\end{aligned}$$

And by $\alpha = \frac{1}{2\lambda} diag(y)\mu$ (which is possible since $y_i \in \{-1, 1\}$ and so $diag(y)^{-1} = diag(y)$ is inversible), this problem is equivalent to :

$$\begin{aligned} &\max_{\alpha \in \mathbb{R}^n} 2\lambda\alpha^T y - \lambda\alpha^T K\alpha \\ \text{s.t. } \forall i \in \{1, ..., n\}, 0 \leq \alpha_i y_i \leq &\frac{1}{2\lambda n} \\ \alpha^T \mathbb{1}_n &= 0 \end{aligned}$$

We cannot apply the coordinate ascent method to this dual. We denote $\alpha^{t+1} = \alpha^t + \delta e_j$ for all $j \in \{1, ..., n\}$. The constraint $\alpha^T \mathbb{1}_n = 0$ gives :

$$\delta = \sum_{i=1}^n \alpha_i^t + \delta = \sum_{i=1}^n ([\alpha^t + \delta e_j]_i = \sum_{i=1}^n \alpha_i^{t+1} = 0$$

3. Let find the update rule of two variables (α_i, α_j) while fixing the others. The constraint $\alpha^T \mathbb{1}_n = 0$ gives us, by fixing other variables :

$$\alpha^{t+1^T} \mathbb{1}_n = 0 = \alpha^{t^T} \mathbb{1}_n \iff \alpha_i^{t+1} + \alpha_j^{t+1} = \alpha_i^t + \alpha_j^t \iff \alpha_i^{t+1} + \alpha_j^{t+1} = \alpha_i^t + \delta + \alpha_j^t - \delta$$

So, similarly to question 1, we want to maximize the following quantity :

$$\begin{aligned} 2(\alpha + \delta e_i - \delta e_j)^T - (\alpha + \delta e_i - \delta e_j)^T K(\alpha + \delta e_i - \delta e_j) &= 2\alpha^T y + 2\delta e_i^T y - 2\delta e_j^T y - \alpha^T K\alpha - \alpha^T K\delta e_i + \alpha^T K\delta e_j \\ &\quad - \delta e_i^T K\alpha + \delta e_j^T K\alpha - \delta^2 e_i^T K e_i + \delta^2 e_i^T K e_j + \delta^2 e_j^T K e_i - \delta^2 e_j^T K e_j \\ &= f(\alpha) + g(\alpha, \delta) \end{aligned}$$

We note that $g(\alpha, \delta) = \delta(2y_i - 2y_j - 2e_i K\alpha + 2e_j K\alpha) - \delta^2(K_{ii} + K_{jj} - 2K_{ij})$ is a concave function with respect to δ. Indeed, $K_{ii} + K_{jj} - 2K_{ij} \geq 0$ since K is a p.d. kernel, so K is a semi-definite positive matrix, so $\forall x \in \mathbb{R}^n, x^T K x \geq 0$, and so by taking the vector full of 0 but with 1 as its i^{th} variable and -1 as its j^{th} variable, we get $K_{ii} + K_{jj} - 2K_{ij} \geq 0$.
Then, by putting the gradient to 0, we compute the maximum :

$$\begin{aligned} \nabla_\delta g(\alpha, \delta) = 0 &\iff 2y_i - 2y_j - 2e_i K\alpha + 2e_j K\alpha - 2\delta(K_{ii} + K_{jj} - 2K_{ij}) = 0 \\ &\iff \delta^* = \frac{y_i - y_j - e_i K\alpha + e_j K\alpha}{K_{ii} + K_{jj} - 2K_{ij}} \end{aligned}$$

Finally, the update rule is :

$$\begin{aligned} \alpha_i^{t+1} &= \alpha_i^t + \frac{y_i - y_j - e_i K\alpha + e_j K\alpha}{K_{ii} + K_{jj} - 2K_{ij}} \\ \alpha_j^{t+1} &= \alpha_j^t - \frac{y_i - y_j - e_i K\alpha + e_j K\alpha}{K_{ii} + K_{jj} - 2K_{ij}} \end{aligned}$$

With $\forall k \in \{i, j\}, 0 \leq y_k \alpha_k \leq \frac{1}{2\lambda n}$

Exercise 5. Duality

1. For all $f \in \mathcal{H}_K$ and $\lambda \in \mathbb{R}$, the Lagrangian of the problem is :

$$\mathcal{L}(f, \lambda) = \frac{1}{n} \sum_{i=1}^{n} l_{y_i}(f(x_i)) + \lambda \|f\|_{\mathcal{H}_K} - \lambda B$$

Let's show that this problem is a convex problem :

— $\forall y \in \{-1, 1\}, \forall x \in \mathcal{X}, l_y(f(x)) = l_y(\langle f, K_x \rangle)$ and since l_y is convex for $y \in \{-1, 1\}$ and $\forall g \in \mathcal{H}_K, f \in \mathcal{H}_K \mapsto \langle f, g \rangle$ is linear in f, then $\forall x \in \mathcal{X}, f \in \mathcal{H}_K \mapsto l_y(f(x))$ is convex.

— $f \in \mathcal{H}_K \mapsto \|f\|_{\mathcal{H}_K} - \lambda$ is clearly convex.

Moreover, with $f = 0$ (0 function), we do have $\|f\|_{\mathcal{H}_K} = 0 < B$. So the problem respect Slater's constraints. Thus :

$$\begin{aligned}
& \underset{f \in \mathcal{H}_K}{min} \frac{1}{n} \sum_{i=1}^{n} l_{y_i}(f(x_i)) \quad \text{s.t.} \quad \|f\|_{\mathcal{H}_K} \leq B \qquad (1) \\
&= \underset{\lambda \in \mathbb{R}}{max} \underset{f \in \mathcal{H}_K}{min} \mathcal{L}(f, \lambda) \quad \text{s.t.} \quad \lambda \geq 0 \\
&= \underset{f \in \mathcal{H}_K}{min} \mathcal{L}(f, \lambda^*) \quad \text{for some } \lambda^* \geq 0 \\
&= \underset{f \in \mathcal{H}_K}{min} \frac{1}{n} \sum_{i=1}^{n} l_{y_i}(f(x_i)) + \lambda^* \|f\|_{\mathcal{H}_K} - \lambda^* B
\end{aligned}$$

Removing the last term not depending on f, we find that the solution to problem (1) can be found by solving :

$$\underset{f \in \mathcal{H}_K}{min} \frac{1}{n} \sum_{i=1}^{n} l_{y_i}(f(x_i)) + \lambda \|f\|_{\mathcal{H}_K} \quad \text{for some } \lambda \geq 0$$

This problem obviously respects the conditions of the representer theorem (since $\lambda \geq 0$), so any of its solution admits a representation of the form :

$$\forall x \in \mathcal{X}, f(x) = \sum_{j=1}^{n} \alpha_j K(x_j, x) \quad \text{for some } \alpha \in \mathbb{R}^n$$

Finally, there exists $\lambda \geq 0$ such that the solution to problem (1) can be found by solving the following problem :

$$\underset{\alpha \in \mathbb{R}^n}{min} \frac{1}{n} \sum_{i=1}^{n} l_{y_i}([K\alpha]_i) + \lambda \alpha^T K \alpha = \underset{\alpha \in \mathbb{R}^n}{min} R(K\alpha) + \lambda \alpha^T K \alpha \qquad (2)$$

With $R : u \in \mathbb{R}^n \mapsto \frac{1}{n} \sum_{i=1}^{n} l_{y_i}(u_i)$.

2. $\forall u \in \mathbb{R}^n$,

$$\begin{aligned} R^*(u) &= \underset{x\in\mathbb{R}^n}{sup}\, x^T u - R(x) \\ &= \underset{x\in\mathbb{R}^n}{sup}\, x^T u - \frac{1}{n}\sum_{i=1}^{n} l_{y_i}(x_i) \\ &= \frac{1}{n}\underset{x\in\mathbb{R}^n}{sup}\sum_{i=1}^{n} nx_iu_i - l_{y_i}(x_i) \end{aligned}$$

Moreover :

$$\begin{aligned} \forall i \in \{1,...,n\}, nx_iu_i - l_{y_i}(x_i) &\leq \underset{x_i\in\mathbb{R}}{sup} nx_iu_i - l_{y_i}(x_i) = nx_i^*u_i - l_{y_i}(x_i^*) \\ \implies \sum_{i=1}^{n} nx_iu_i - l_{y_i}(x_i) &\leq \sum_{i=1}^{n} nx_i^*u_i - l_{y_i}(x_i^*) \\ \implies \underset{x\in\mathbb{R}^n}{sup}\sum_{i=1}^{n} nx_iu_i - l_{y_i}(x_i) &\leq \sum_{i=1}^{n} nx_i^*u_i - l_{y_i}(x_i^*) \end{aligned}$$

Taking $x = (x_1^*, ..., x_n^*)^T$, we have the equality :

$$\underset{x\in\mathbb{R}^n}{sup}\sum_{i=1}^{n} nx_iu_i - l_{y_i}(x_i) = \sum_{i=1}^{n}\underset{x_i\in\mathbb{R}}{sup} nx_iu_i - l_{y_i}(x_i)$$

And finally :

$$\begin{aligned} R^*(u) &= \frac{1}{n}\sum_{i=1}^{n}\underset{x_i\in\mathbb{R}}{sup} nx_iu_i - l_{y_i}(x_i) \\ R^*(u) &= \frac{1}{n}\sum_{i=1}^{n} l_{y_i}^*(nu_i) \end{aligned}$$

3.

$$\underset{\alpha\in\mathbb{R}^n, u\in\mathbb{R}^n}{min} R(u) + \lambda\alpha^T K\alpha \quad \text{s.t.} \quad u = K\alpha \tag{3}$$

So, for all $\alpha \in \mathbb{R}^n, u \in \mathbb{R}^n, \nu \in \mathbb{R}^n$:

$$\begin{aligned} \mathcal{L}(\alpha, u, \nu) &= R(u) + \lambda\alpha^T K\alpha + \nu^T(K\alpha - u) \\ &= R(u) - \nu^T u + \lambda\alpha^T K\alpha + \nu^T K\alpha \end{aligned}$$

And so :

$$\begin{aligned} \underset{\alpha\in\mathbb{R}^n, u\in\mathbb{R}^n}{inf} \mathcal{L}(\alpha, u, \nu) &= \underset{u\in\mathbb{R}^n}{inf}\,(R(u) - \nu^T u) + \underset{\alpha\in\mathbb{R}^n}{inf}\,(\lambda\alpha^T K\alpha + (K\nu)^T\alpha) \\ &= -\underset{u\in\mathbb{R}^n}{sup}\,(\nu^T u - R(u)) + \underset{\alpha\in\mathbb{R}^n}{inf}\,(\lambda\alpha^T K\alpha + (K\nu)^T\alpha) \\ &= -R^*(\nu) + \underset{\alpha\in\mathbb{R}^n}{inf}\,(\lambda\alpha^T K\alpha + (K\nu)^T\alpha) \end{aligned}$$

Moreover, $g : \alpha \in \mathbb{R}^n \mapsto \lambda\alpha^T K\alpha + (K\nu)^T\alpha$ is convex and differentiable, and $\forall\alpha \in \mathbb{R}^n, \nabla_\alpha g = 2\lambda K\alpha + K\nu$. Thus :

$$\nabla_{\alpha^*} g = 0 \iff 2\lambda K\alpha^* + K\nu = 0 \iff K\alpha^* = -\frac{1}{2\lambda}K\nu$$

Then :

$$\lambda{\alpha^*}^T K\alpha^* = -\frac{1}{2}{\alpha^*}^T K\nu = -\frac{1}{2}\nu^T K\alpha^* = \frac{1}{4\lambda}\nu^T K\nu$$

Finally :

$$\underset{\alpha\in\mathbb{R}^n}{inf}\,(\lambda\alpha^T K\alpha + \nu^T K\alpha) = \frac{1}{4\lambda}\nu^T K\nu - \frac{1}{2\lambda}K\nu = -\frac{1}{4\lambda}\nu^T K\nu$$

Thus, the dual problem of (3) is :

$$\underset{\nu\in\mathbb{R}^n}{max} - R^*(\nu) - \frac{1}{4\lambda}\nu^T K\nu \iff \underset{\nu\in\mathbb{R}^n}{min} R^*(\nu) + \frac{1}{4\lambda}\nu^T K\nu$$

We got the condition that $K(2\lambda\alpha + \nu) = 0$, so for ν^* solution of the dual problem, a solution α^* of the problem (3) is such that $(2\lambda\alpha + \nu^*) \in Ker(K)$.

4. Let note $H_u(x) = xu - l_y(x)$, so that $l_y^*(u) = \underset{x\in\mathbb{R}}{sup} H_u(x)$.

— First, $H_u(x) = xu - log(1 + e^{-yx})$. Since $y^2 = 1$:

$$H_u(x) = (xy)(yu) - log(1 + e^{-yx}) = (yu + 1)yx - log(1 + e^{yx})$$

So $H_u(x) \longrightarrow \infty$ when $yu > 0$ and $xy \longrightarrow \infty$ or $yu < -1$ and $xy \longrightarrow -\infty$. So we can focus on the case $-1 \le yu \le 0$. H_u is twice differentiable, and for all $x \in \mathbb{R}$:

$$H'_u(x) = u + \frac{ye^{-yx}}{1 + e{-}yx}$$
$$H''_u(x) = \frac{-y^2 e^{-yx}}{(1 + e{-}yx)^2} \le 0 \Longrightarrow \text{concave}$$

So H_u has an upper bound on $-1 \le yu \le 0$ given by :

$$H'_u(x^*) = 0 \iff x^* = ylog(-1 - \frac{1}{uy})$$

And $H_u(x^*) = (uy + 1)log(uy + 1) - uylog(-uy)$. Finally :

$$l_y^*(u) = \begin{cases} (uy + 1)log(uy + 1) - uylog(-uy) & \text{if } -1 \le yu \le 0 \\ +\infty & \text{otherwise.} \end{cases}$$

So the dual problem is :

$$\underset{\nu\in\mathbb{R}^n}{min}\frac{1}{n}\sum_{i=1}^{n}[(n\nu_i y_i + 1)log(n\nu_i y_i + 1) - n\nu_i y_i log(-n\nu_i y_i)] + \frac{1}{4\lambda}\nu^T K\nu$$

Subject to $0 \le -\nu_i y_i \le \frac{1}{n}$, for all $i \in \{1, ..., n\}$.

— Second, $H_u(x) = xu - max(0, 1 - yx)^2$. Since $H_u(x) = (xy)(yu) - max(0, 1 - yx)^2$, it is the same as maximizing $A_v(z) = zv - max(0, 1-z)^2$ where $v = yu$ and $z = xy$. Since $A_v(z) \longrightarrow \infty$ when $v > 0$ and $z \longrightarrow \infty$, we can focus on the case $v \leq 0$. In this case, $A_v(z) = vz$ when $z \geq 1$ and his supremum is v (since $v \leq 0$). When $z < 1$, $A_v(z) = vz - (1-z)^2$. This quadratic function is concave and reaches its maximum when $z = 1 + \frac{v}{2}$ and its value is $v + \frac{v^2}{4} \geq v$. Thus, $\underset{z \in \mathbb{R}}{sup} A_v(z) = \infty$ if $v > 0$ and $v + \frac{v^2}{4}$ if $v \leq 0$. Finally :

$$l_y^*(u) = \begin{cases} uy + \frac{u^2}{4} & \text{if } yu \leq 0 \\ +\infty & \text{otherwise.} \end{cases}$$

So the dual problem is :

$$\underset{\nu \in \mathbb{R}^n}{min} \frac{1}{n} \sum_{i=1}^{n} [n\nu_i y_i + \frac{n^2 \nu_i^2}{4}] + \frac{1}{4\lambda} \nu^T K \nu$$

Subject to $\nu_i y_i \leq 0$, for all $i \in \{1, ..., n\}$.