MVA "Kernel methods in machine learning" Homework

Antoine Moulin

Upload your answers (in PDF) to:
`http://tiny.cc/dpxnjz`
before February 26, 2020, 1pm (Paris time).

Exercice 1. Kernels

Show that the following kernels are positive definite:

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

$$\forall x, y \in \mathcal{X} \quad K_4(x, y) = \min(f(x)g(y), f(y)g(x))$$

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

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

where $|\,F\,|$ denotes the cardinality of F, and with the convention $\frac{0}{0} = 0$.

Solution 1.

1. First, let's show that $(x, y) \mapsto \min(x, y)$ is a positive definite kernel on $\mathbb{R}_+$. Let $x, y \in \mathbb{R}_+$. We have:

$$\min(x, y) = \int_{\mathbb{R}_+} \mathbb{1}_{t \leq x} \mathbb{1}_{t \leq y} dt$$

Let $n \in \mathbb{N}, (a_1, \ldots, a_n) \in \mathbb{R}^n, (x_1, \ldots, x_n) \in \mathbb{R}_+^n$. We have:

$$\sum_{1 \leq i,j \leq n} a_i a_j \min(x_i, x_j) = \int_{\mathbb{R}_+} \left(\sum_{1 \leq i \leq n} a_i \mathbb{1}_{t \leq x_i} \right)^2 dt \geq 0$$

Hence, $(x, y) \mapsto \min(x, y)$ is a p.d. kernel.

Now, let $x, y \in \mathcal{X}$ such that $g(x), g(y) > 0$. We have:

$$\begin{aligned} K\,(x, y) &= \min\left(f(x)g(y), f(y)g(x)\right) \\ &= g(x)g(y) \min\left(\frac{f(x)}{g(x)}, \frac{f(y)}{g(y)}\right) \end{aligned}$$

Let $n \in \mathbb{N}, (a_1, \ldots, a_n) \in \mathbb{R}^n, (x_1, \ldots, x_n) \in \mathcal{X}^n$. g and f are nonnegative so:

$$\sum_{1 \leq i,j \leq n} a_i a_j K(x_i, x_j) = 0 + \sum_{i,j:g(x_i),g(x_j)>0} (a_i g(x_i))(a_j g(x_j)) \min\left(\frac{f(x_i)}{g(x_i)}, \frac{f(x_j)}{g(x_j)}\right)$$

As $(x, y) \mapsto \min(x, y)$ is a p.d. kernel, for all $m \in \mathbb{N}, (b_1, \ldots, b_m) \in \mathbb{R}^m, (y_1, \ldots, y_m) \in \mathcal{X}^m$ we have $\sum_{1 \leq i,j \leq m} b_i b_j \min(y_i, y_j) \geq 0$. In particular, it is true for $m = n$, $b_i = a_i g(x_i)$ and $y_i = \frac{f(x_i)}{g(x_i)}$ if $g(x_i) \neq 0$ and 0 otherwise.

Hence, $\boxed{K : (x, y) \mapsto \min(f(x)g(y), f(y)g(x)) \text{ is a p.d. kernel}}$.

2. First, let us show that for $n \in \mathbb{N}$, $(x, y) \mapsto \frac{1}{1-x^\top y}$ is a p.d. kernel. Indeed, we have for $x, y \in \mathbb{R}^n$ such that $\left|x^\top y\right| < 1$:

$$\frac{1}{1 - x^\top y} = \sum_{i=0}^{+\infty} \left(x^\top y\right)^i = \lim_{p \to +\infty} \sum_{i=0}^{p} \left(x^\top y\right)^i$$

The polynomial kernel is a p.d. kernel. A sum of p.d. kernels is a p.d. kernel, and the limit of a sequence of p.d. kernels is a p.d. kernel. Hence, the previous function is a p.d. kernel (this is actually also valid with any positive kernel which is strictly bounded by 1 in absolute value).

E is a non-empty finite set, which we denote by $E = \{x_1, \ldots, x_n\}$. For $A \subset E$, we define $x_A \in \mathbb{R}^n$ the vector whose i-th component is 1 if $x_i \in A$, 0 otherwise.

Hence, for $A, B \subset E$, we have $|A \cap B| = x_A^\top x_B$ which means that $(A, B) \mapsto |A \cap B|$ is a p.d. kernel. By Morgan's law, $|A \cup B| = \left|\overline{\overline{A} \cap \overline{B}}\right| = n - \left|\overline{A} \cap \overline{B}\right|$. Hence,

$$\frac{1}{|A \cup B|} = \frac{1}{n} \frac{1}{1 - \frac{1}{n}\left|\overline{A} \cap \overline{B}\right|}$$

$\frac{1}{n} \geq 0$ so $(A, B) \mapsto \frac{1}{n}\left|\overline{A} \cap \overline{B}\right|$ is a p.d. kernel. Thanks to the first remark, we know that $(A, B) \mapsto \frac{1}{1-\frac{1}{n}\left|\overline{A} \cap \overline{B}\right|}$ is a p.d. kernel and so is $\frac{1}{|A \cup B|}$. The product of two p.d. kernels is a p.d. kernel, so we can conclude that $\boxed{K : (A, B) \mapsto \frac{|A \cap B|}{|A \cup B|} \text{ is a p.d. kernel}}$.

Exercice 2. Kernels encoding equivalence classes.

Consider a similarity measure $K : \mathcal{X} \times \mathcal{X} \to \{0, 1\}$ with $K(x, x) = 1$ for all x in $\mathcal{X}$. Prove that K is p.d. if and only if, for all x, x', x'' in $\mathcal{X}$,

- $K(x, x') = 1 \Leftrightarrow K(x', x) = 1$, and
- $K(x, x') = K(x', x'') = 1 \Rightarrow K(x, x'') = 1$.

Solution 2.

- **Suppose K is p.d.** By definition, K is symmetric, hence the first property is verified:

  $$\forall x, x' \in \mathcal{X}, K\left(x, x'\right) = 1 \Leftrightarrow K\left(x', x\right) = 1$$

  Let $x, x', x'' \in \mathcal{X}, a_1, a_2, a_3 \in \mathbb{R}$. As K is p.d., we have:

  $$a_1^2 K\left(x, x\right) + a_2^2 K\left(x', x'\right) + a_3^2 K\left(x'', x''\right) + 2a_1a_2K\left(x, x'\right) + 2a_1a_3K\left(x, x''\right) + 2a_2a_3K\left(x', x''\right) \geq 0$$

  By definition of K, this can be written as:

  $$a_1^2 + a_2^2 + a_3^2 + 2a_1a_2K\left(x, x'\right) + 2a_1a_3K\left(x, x''\right) + 2a_2a_3K\left(x', x''\right) \geq 0$$

  We suppose $K\left(x, x'\right) = K\left(x', x''\right) = 1$:

  $$a_1^2 + a_2^2 + a_3^2 + 2a_1a_2 + 2a_1a_3K\left(x, x''\right) + 2a_2a_3 \geq 0$$

  We take $a_2 = -a_1$ and $a_1 = a_3$:

  $$a_1^2 + a_1^2 + a_1^2 - 2a_1^2 + 2a_1^2K\left(x, x''\right) - 2a_1^2 \geq 0$$

  i.e.

  $$a_1^2\left(2K\left(x, x''\right) - 1\right) \geq 0$$

  and

  $$K\left(x, x''\right) \geq \frac{1}{2}$$

  As K only takes values in $\{0, 1\}$, we conclude that $K\left(x, x''\right) = 1$. Hence,

  $$\forall x, x', x'' \in \mathcal{X}, K\left(x, x'\right) = K\left(x', x''\right) = 1 \Rightarrow K\left(x, x''\right) = 1$$

- **Suppose the properties are verified.** Let $n \in \mathbb{N}, (a_1, \ldots, a_n) \in \mathbb{R}^n, (x_1, \ldots, x_n) \in \mathcal{X}^n$. For $x, y \in \mathcal{X}$, we define $x \sim y$ if $K\left(x, y\right) = 1$. $\sim$ is an equivalence relation: (i) the reflexive property is verified by construction of K, (ii) the symmetric property corresponds to the first assumption and (iii) the transitive property corresponds to the second assumption. By definition of this relation, we have:

  $$\sum_{1 \leq i,j \leq n} a_i a_j K(x_i, x_j) = \sum_{i=1}^{n} a_i^2 + 2 \sum_{i,j : x_i \sim x_j} a_i a_j$$

  Let us denote K the number of equivalence classes and C_k the k-th class. We can group the terms and see that the previous term can be written as:

  $$\sum_{1 \leq i,j \leq n} a_i a_j K\left(x_i, x_j\right) = \sum_{k=1}^{K} \left(\sum_{x \in C_k} x\right)^2 \geq 0$$

  Hence, K is p.d.

Exercice 3. COCO

Given two sets of real numbers $X = (x_1, \ldots, x_n) \in \mathbb{R}^n$ and $Y = (y_1, \ldots, y_n) \in \mathbb{R}^n$, the covariance between X and Y is defined as

$$\mathrm{cov}_n(X, Y) = \mathbf{E}_n(XY) - \mathbf{E}_n(X)\mathbf{E}_n(Y)\,,$$

where $\mathbf{E}_n(U) = (\sum_{i=1}^n u_i)/n$. The covariance is useful to detect linear relationships between X and Y. In order to extend this measure to potential nonlinear relationships between X and Y, we consider the following criterion:

$$C_n^K(X, Y) = \max_{f,g \in \mathcal{B}_K} \mathrm{cov}_n(f(X), g(Y))\,,$$

where K is a positive definite kernel on $\mathbb{R}$, $\mathcal{B}_K$ is the unit ball of the RKHS of K, and $f(U) = (f(u_1), \ldots, f(u_n))$ for a vector $U = (u_1, \ldots, u_n)$.

1. Express simply $C_n^K(X, Y)$ for the linear kernel $K(a, b) = ab$.
2. For a general kernel K, express $C_n^K(X, Y)$ in terms of the Gram matrices of X and Y.

Solution 3.

1. For the linear kernel, the RKHS can be represented as $\mathbb{R}$ (for $x \in \mathbb{R}$, $K_x : y \mapsto xy$ is just a linear function). Hence,

$$\begin{aligned} C_n^K(X, Y) &= \max_{-1 \leq f, g \leq 1} \mathbb{E}_n(fXgY) - \mathbb{E}_n(fX)\,\mathbb{E}_n(gY) \\ &= \max_{-1 \leq f, g \leq 1} fg\left(\frac{1}{n}\sum_{i=1}^n x_i y_i - \frac{1}{n^2}\sum_{i=1}^n x_i \sum_{i=1}^n y_i\right) \\ &= \max_{-1 \leq f, g \leq 1} \frac{fg}{n}\left(X^\top Y - X^\top \frac{\mathbf{1}\mathbf{1}^\top}{n} Y\right) \end{aligned}$$

   Finally,

$$\boxed{C_n^K(X, Y) = \frac{1}{n}\left| X^\top \left(\mathbf{I}_n - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right) Y \right|}$$

2. Let (f^*, g^*) be a solution of the maximization problem. Then, fixing the second variable, we have that f^* is a solution of the maximization problem:

$$\begin{aligned} &\max_{f \in \mathcal{H}} \mathrm{cov}_n(f(X), g^*(Y)) \\ &\text{s.t. } \|f\| \leq 1 \end{aligned}$$

   By the reproducing property, this problem is linear in f, thus convex. The Slater's condition can be verified with $f = 0$, hence strong duality holds. This means that if λ^* is a dual optimal, then f^* minimizes the Lagrangian $f \mapsto \mathcal{L}(f, \lambda^*)$ where:

$$\mathcal{L}(f, \lambda^*) = -\mathrm{cov}_n(f(X), g^*(Y)) + \lambda^*(\|f\| - 1)$$

By the representer theorem, there exists $\alpha \in \mathbb{R}^n$ such that $f^* = \sum_{1 \leq i \leq n} \alpha_i \mathbf{K}_{x_i}$. Similarly, there exists $\beta \in \mathbb{R}^n$ such that $g^* = \sum_{1 \leq i \leq n} \beta_i \mathbf{K}_{y_i}$. We denote $\mathbf{K}^X$ and $\mathbf{K}^Y$ the kernel matrices associated to X and Y, respectively. We have:

$$\begin{aligned}
\mathrm{cov}_n\left(f\left(X\right), g\left(Y\right)\right) &= \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{n}\alpha_j K\left(x_j, x_i\right)\right)\left(\sum_{j=1}^{n}\beta_j K\left(y_j, y_i\right)\right) \\
&\quad - \frac{1}{n^2}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_j K\left(x_j, x_i\right)\right)\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\beta_j K\left(y_j, y_i\right)\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{K}^X\alpha\right]_i\left[\mathbf{K}^Y\beta\right]_i - \frac{1}{n^2}\left(\sum_{i=1}^{n}\left[\mathbf{K}^X\alpha\right]_i\right)\left(\sum_{i=1}^{n}\left[\mathbf{K}^Y\beta\right]_i\right) \\
&= \frac{1}{n}\alpha^\top\mathbf{K}^X\mathbf{K}^Y\beta - \frac{1}{n^2}\alpha^\top\mathbf{K}^X\mathbf{1}\mathbf{1}^\top\mathbf{K}^Y\beta \\
&= \frac{1}{n}\alpha^\top\mathbf{K}^X\left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right)\mathbf{K}^Y\beta
\end{aligned}$$

Hence, $C_n^K\left(X, Y\right)$ can be written as

$$\begin{aligned}
\max_{\alpha,\beta\in\mathbb{R}^n} \; & \frac{1}{n}\alpha^\top\mathbf{K}^X\left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right)\mathbf{K}^Y\beta \\
\text{s.t. } \; & \alpha^\top\mathbf{K}^X\alpha \leq 1 \\
& \beta^\top\mathbf{K}^Y\beta \leq 1
\end{aligned}$$

$C_n^K\left(X, Y\right)$ looks like a singular value but with a different norm, let us write the problem differently. As $\mathbf{K}^X$ and $\mathbf{K}^Y$ are positive semi-definite matrices, we can write $\mathbf{K}^X = \mathbf{K}^{X\frac{1}{2}}\mathbf{K}^{X\frac{1}{2}}$ and $\mathbf{K}^Y = \mathbf{K}^{Y\frac{1}{2}}\mathbf{K}^{Y\frac{1}{2}}$. Thus, $C_n^K\left(X, Y\right)$ becomes

$$\begin{aligned}
\max_{\alpha,\beta\in\mathbb{R}^n} \; & \frac{1}{n}\left(\mathbf{K}^{X\frac{1}{2}}\alpha\right)^\top\mathbf{K}^{X\frac{1}{2}}\left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right)\mathbf{K}^{Y\frac{1}{2}}\left(\mathbf{K}^{Y\frac{1}{2}}\beta\right) \\
\text{s.t. } \; & \left\|\mathbf{K}^{X\frac{1}{2}}\alpha\right\|_2 \leq 1 \\
& \left\|\mathbf{K}^{Y\frac{1}{2}}\beta\right\|_2 \leq 1
\end{aligned}$$

Let us denote C the quantity

$$\begin{aligned}
\max_{\tilde{\alpha},\tilde{\beta}\in\mathbb{R}^n} \; & \frac{1}{n}\tilde{\alpha}^\top\mathbf{K}^{X\frac{1}{2}}\left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right)\mathbf{K}^{Y\frac{1}{2}}\tilde{\beta} \\
\text{s.t. } \; & \left\|\tilde{\alpha}\right\|_2 \leq 1 \\
& \left\|\tilde{\beta}\right\|_2 \leq 1
\end{aligned}$$

We immediately have $C_n^K\left(X, Y\right) \leq C$. The other inequality can be obtained using the eigendecompositions of $\mathbf{K}^{X\frac{1}{2}}$ and $\mathbf{K}^{Y\frac{1}{2}}$. Besides, by definition, $C = \frac{1}{n}\sigma_{max}\left[\mathbf{K}^{X\frac{1}{2}}\left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right)\mathbf{K}^{Y\frac{1}{2}}\right]$ where $\sigma_{max}\left(A\right)$ is the largest singular value of the matrix A. Hence, we can conclude:

$$\boxed{C_n^K\left(X, Y\right) = \frac{1}{n}\sigma_{max}\left[\mathbf{K}^{X\frac{1}{2}}\left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right)\mathbf{K}^{Y\frac{1}{2}}\right] = \frac{1}{n}\left\|\mathbf{K}^{X\frac{1}{2}}\left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right)\mathbf{K}^{Y\frac{1}{2}}\right\|_2}$$

Exercice 4. Dual coordinate ascent algorithms for SVMs

We recall the primal formulation of SVMs seen in the class (slide 148).

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \max(0, 1 - y_i f(\mathbf{x}_i)) + \lambda \|f\|_{\mathcal{H}}^2,$$

and its dual formulation (slide 158)

$$\max_{\boldsymbol{\alpha} \in \mathbb{R}^n} 2\boldsymbol{\alpha}^\top \mathbf{y} - \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha} \quad \text{such that} \quad 0 \leq y_i \alpha_i \leq \frac{1}{2\lambda n}, \quad \text{for all } i.$$

1. The coordinate ascent method consists of iteratively optimizing with respect to one variable, while fixing the other ones. Assuming that you want to maximize the dual by following this approach. Find (and justify) the update rule for α_j.

2. Consider now the primal formulation of SVMs with intercept

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

   Can we still apply the representer theorem? Why? Derive the corresponding dual formulation by using Lagrangian duality. Can we apply the coordinate ascent method to this dual? If yes, what are the update rules?

3. Consider a coordinate ascent method to this dual that consists of updating two variables (α_i, α_j) at a time (while fixing the $n-2$ other variables). What are the update rules for these two variables?

Solution 4.

1. **Note:** As the dual's formulation was given, it has been kept for the following proof, but the results would have been simpler if the formulation from slide 157 was used, as it is done in question 3.

   Let $j \in [\![1, n]\!]$. We denote $\boldsymbol{\alpha}^t$ the vector at the t-th iteration and $\mathbf{e}_j$ the j-th vector of the $\mathbb{R}^n$'s canonical basis. To update the j-th coordinate, we take δ such that it maximizes the following optimization problem:

   $$\max_{\delta \in \mathbb{R}} 2 \left(\boldsymbol{\alpha}^t + \delta \mathbf{e}_j\right)^\top \mathbf{y} - \left(\boldsymbol{\alpha}^t + \delta \mathbf{e}_j\right)^\top \mathbf{K} \left(\boldsymbol{\alpha}^t + \delta \mathbf{e}_j\right) \text{ such that } 0 \leq y_j \left(\alpha_j^t + \delta\right) \leq \frac{1}{2\lambda n}$$

   or, equivalently:

   $$\max_{\delta \in \mathbb{R}} 2\delta \left(y_j - \left[\mathbf{K}\boldsymbol{\alpha}^t\right]_j\right) - \delta^2 \mathbf{K}_{jj} \text{ such that } 0 \leq y_j \left(\alpha_j^t + \delta\right) \leq \frac{1}{2\lambda n}$$

   Let's suppose that $\mathbf{K}_{jj} \neq 0$. If we consider the unconstrained problem, the maximum is reached when the gradient is null. This yields $\delta = \frac{y_j - \left[\mathbf{K}\boldsymbol{\alpha}^t\right]_j}{\mathbf{K}_{jj}}$. Let's suppose that $y_j = 1$

(the case $y_j = -1$ is similar). If we are strictly inside the set of feasible points – i.e. when $0 < y_j \left(\alpha_j^t + \delta\right) < \frac{1}{2\lambda n}$ – then the update can be done as $\alpha_j^{t+1} = \alpha_j^t + \delta$. If $y_j \left(\alpha_j^t + \delta\right) \leq 0$, then $\alpha_j^{t+1} = 0$ and if $y_j \left(\alpha_j^t + \delta\right) \geq \frac{1}{2\lambda n}$, then $\alpha_j^{t+1} = \frac{1}{2\lambda n}$. Hence, for $y_j = 1$, we can update α_j as:

$$\alpha_j^{t+1} = \min \left[\max \left(\alpha_j^t + \frac{y_j - [\mathbf{K}\boldsymbol{\alpha}^t]_j}{\mathbf{K}_{jj}}, 0 \right), \frac{1}{2\lambda n} \right]$$

Similarly, if $y_j = -1$, the condition becomes $0 \geq \alpha_j^t + \delta \geq -\frac{1}{2\lambda n}$ and we have:

$$\alpha_j^{t+1} = \max \left[\min \left(\alpha_j^t + \frac{y_j - [\mathbf{K}\boldsymbol{\alpha}^t]_j}{\mathbf{K}_{jj}}, 0 \right), -\frac{1}{2\lambda n} \right]$$

Using the fact that for $x, y \in \mathbb{R}$, $\max(x, y) = -\min(-x, -y)$ and $\min(x, y) = -\max(-x, -y)$, we get:

$$\begin{aligned} \alpha_j^{t+1} &= -\min \left[\max \left(-\alpha_j^t - \frac{y_j - [\mathbf{K}\boldsymbol{\alpha}^t]_j}{\mathbf{K}_{jj}}, 0 \right), \frac{1}{2\lambda n} \right] \\ &= y_j \min \left[\max \left\{ y_j \left(\alpha_j^t + \frac{y_j - [\mathbf{K}\boldsymbol{\alpha}^t]_j}{\mathbf{K}_{jj}} \right), 0 \right\}, \frac{1}{2\lambda n} \right] \end{aligned}$$

If $\mathbf{K}_{jj} = 0$, we have to solve a linear problem and distinguishing six cases (depending on the sign of y_j and if $y_j - [\mathbf{K}\boldsymbol{\alpha}^t]_j$ is negative, null or positive), we can show that:

$$\alpha_j^{t+1} = \alpha_j^t \mathbb{1}\left(y_j - \left[\mathbf{K}\boldsymbol{\alpha}^t\right]_j = 0\right) + \frac{y_j}{2\lambda n} \mathbb{1}\left(y_j \left\{y_j - \left[\mathbf{K}\boldsymbol{\alpha}^t\right]_j\right\} > 0\right)$$

Finally, we have:

$$\boxed{\forall j \in [\![1, n]\!], \alpha_j^{t+1} = \begin{cases} \alpha_j^t \mathbb{1}\left(y_j - [\mathbf{K}\boldsymbol{\alpha}^t]_j = 0\right) + \frac{y_j}{2\lambda n} \mathbb{1}\left(y_j \left\{y_j - [\mathbf{K}\boldsymbol{\alpha}^t]_j\right\} > 0\right) & \text{if } \mathbf{K}_{jj} = 0 \\ y_j \min \left[\max \left\{ y_j \left(\alpha_j^t + \frac{y_j - [\mathbf{K}\boldsymbol{\alpha}^t]_j}{\mathbf{K}_{jj}} \right), 0 \right\}, \frac{1}{2\lambda n} \right] & \text{if } \mathbf{K}_{jj} \neq 0 \end{cases}}$$

2. Let (f^*, b^*) be a solution to the problem. Then f^* is a solution to the minimization problem:

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \max\left(0, 1 - y_i \left(f\left(\mathbf{x}_i\right) + b^*\right)\right) + \lambda \left\|f\right\|_{\mathcal{H}}^2$$

Hence, we can use the representer theorem on f. There exists $\boldsymbol{\alpha}^*$ such that the solution f^* satisfies:

$$f^* = \sum_{i=1}^{n} \alpha_i^* \mathbf{K}_{\mathbf{x}_i}$$

The problem becomes:

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^n, b \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \max\left(0, 1 - y_i \left([\mathbf{K}\boldsymbol{\alpha}]_i + b\right)\right) + \lambda \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha}$$

We introduce slack variables $\xi_1, \ldots, \xi_n \in \mathbb{R}$. The problem is equivalent to:

$$\begin{aligned}\min_{\boldsymbol{\alpha} \in \mathbb{R}^n, b \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^n} & \frac{1}{n} \sum_{i=1}^{n} \xi_i + \lambda \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha} \\ \text{s.t. } & \xi_i \geq 0 & \text{for } i = 1, \ldots, n \\ & \xi_i \geq 1 - y_i \left([\mathbf{K}\boldsymbol{\alpha}]_i + b \right) & \text{for } i = 1, \ldots, n\end{aligned}$$

Let $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathbb{R}^n$. The Lagrangian is given by:

$$\begin{aligned}\mathcal{L}\left(\boldsymbol{\alpha}, b, \boldsymbol{\xi}, \boldsymbol{\mu}, \boldsymbol{\nu}\right) &= \frac{1}{n} \sum_{i=1}^{n} \xi_i + \lambda \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha} - \sum_{i=1}^{n} \mu_i \left[y_i \left([\mathbf{K}\boldsymbol{\alpha}]_i + b \right) + \xi_i - 1 \right] - \sum_{i=1}^{n} \nu_i \xi_i \\ &= \boldsymbol{\xi}^\top \frac{\mathbf{1}}{n} + \lambda \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha} - \left(\operatorname{Diag}(\mathbf{y}) \boldsymbol{\mu} \right)^\top \left(\mathbf{K}\boldsymbol{\alpha} + b \cdot \mathbf{1} \right) - \left(\boldsymbol{\mu} + \boldsymbol{\nu} \right)^\top \boldsymbol{\xi} + \boldsymbol{\mu}^\top \mathbf{1}\end{aligned}$$

$\mathcal{L}$ is a convex quadratic function in $\boldsymbol{\alpha}$. It is minimized whenever the gradient is null:

$$\nabla_{\boldsymbol{\alpha}} \mathcal{L} = 2\lambda \mathbf{K} \boldsymbol{\alpha} - \mathbf{K} \operatorname{Diag}(\mathbf{y}) \boldsymbol{\mu} = \mathbf{K} \left(2\lambda \boldsymbol{\alpha} - \operatorname{Diag}(\mathbf{y}) \boldsymbol{\mu} \right) = 0$$

This gives:

$$\boldsymbol{\alpha}^* = \frac{\operatorname{Diag}(\mathbf{y}) \boldsymbol{\mu}}{2\lambda}$$

$\mathcal{L}$ is a linear function in b. Its minimum is $-\infty$ except when it is constant, i.e., when

$$\nabla_b \mathcal{L} = -\left(\operatorname{Diag}(\mathbf{y}) \boldsymbol{\mu} \right)^\top \mathbf{1} = 0 \quad \text{i.e.} \quad \sum_{i=1}^{n} y_i \mu_i = 0$$

$\mathcal{L}$ is a linear f unction in $\boldsymbol{\xi}$. Its minimum is $-\infty$ except when it is constant, i.e., when

$$\nabla_{\boldsymbol{\xi}} \mathcal{L} = \frac{\mathbf{1}}{n} - \boldsymbol{\mu} - \boldsymbol{\nu} = 0 \quad \text{i.e.} \quad \boldsymbol{\mu} + \boldsymbol{\nu} = \frac{\mathbf{1}}{n}$$

Hence, we obtain the Lagrange dual function:

$$\begin{aligned}q\left(\boldsymbol{\mu}, \boldsymbol{\nu}\right) &= \inf_{\boldsymbol{\alpha} \in \mathbb{R}^n, b \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^n} \mathcal{L}\left(\boldsymbol{\alpha}, b, \boldsymbol{\xi}, \boldsymbol{\mu}, \boldsymbol{\nu}\right) \\ &= \begin{cases} \boldsymbol{\mu}^\top \mathbf{1} - \frac{1}{4\lambda} \boldsymbol{\mu}^\top \operatorname{Diag}(\mathbf{y}) \mathbf{K} \operatorname{Diag}(\mathbf{y}) \boldsymbol{\mu} & \text{if } \boldsymbol{\mu} + \boldsymbol{\nu} = \frac{\mathbf{1}}{n} \text{ and } \sum_{i=1}^{n} y_i \mu_i = 0 \\ -\infty & \text{otherwise} \end{cases}\end{aligned}$$

The dual problem is:

$$\max_{\boldsymbol{\mu} \geq 0, \boldsymbol{\nu} \geq 0} q\left(\boldsymbol{\mu}, \boldsymbol{\nu}\right)$$

If $\mu_i > \frac{1}{n}$ for some i, then there is no $\nu_i \geq 0$ such that $\mu_i + \nu_i = \frac{1}{n}$, hence $q\left(\boldsymbol{\mu}, \boldsymbol{\nu}\right) = -\infty$. Otherwise, the dual function takes finite values that depend only on $\boldsymbol{\mu}$ by taking $\nu_i = \frac{1}{n} - \mu_i$. The dual problem is therefore equivalent to:

$$\begin{aligned}\max_{\boldsymbol{\mu} \in \mathbb{R}^n} & \boldsymbol{\mu}^\top \mathbf{1} - \frac{1}{4\lambda} \boldsymbol{\mu}^\top \operatorname{Diag}(\mathbf{y}) \mathbf{K} \operatorname{Diag}(\mathbf{y}) \boldsymbol{\mu} \\ \text{s.t. } & 0 \leq \boldsymbol{\mu} \leq \frac{\mathbf{1}}{n} \quad \text{and} \quad \sum_{i=1}^{n} y_i \mu_i = 0\end{aligned}$$

As the link between $\boldsymbol{\alpha}$ and $\boldsymbol{\mu}$ is simple, we can formulate the dual as an optimization problem on $\boldsymbol{\alpha}$:

$$\boxed{\begin{aligned} &\max_{\boldsymbol{\alpha}\in\mathbb{R}^n} 2\boldsymbol{\alpha}^\top \mathbf{y} - \boldsymbol{\alpha}^\top \mathbf{K}\boldsymbol{\alpha} \\ &\text{s.t. } 0 \leq y_i\alpha_i \leq \frac{1}{2\lambda n} \quad \text{for } i = 1, \ldots, n \\ &\qquad \sum_{i=1}^{n} \alpha_i = 0 \end{aligned}}$$

Because of the supplemental condition on $\boldsymbol{\alpha}$, if we fix $n-1$ variables, then the last one is fixed as well. So <u>we cannot use the coordinate ascent method to this dual</u>.

3. For more simplicity, we use this dual's formulation:

$$\begin{aligned} &\max_{\boldsymbol{\mu}\in\mathbb{R}^n} \boldsymbol{\mu}\top\mathbf{1} - \frac{1}{4\lambda}\boldsymbol{\mu}^\top \operatorname{Diag}(\mathbf{y})\,\mathbf{K}\operatorname{Diag}(\mathbf{y})\,\boldsymbol{\mu} \\ &\text{s.t. } 0 \leq \mu_k \leq \frac{1}{n} \quad \text{for } k = 1, \ldots, n \\ &\qquad \sum_{k=1}^{n} y_k\mu_k = 0 \end{aligned}$$

We want to optimize this dual with respect to μ_i and μ_j, while the other variables are fixed. We have:

$$y_i\mu_i + y_j\mu_j = -\sum_{k\neq i,j} y_k\mu_k \triangleq A$$

Hence, we can write μ_j as a function of μ_i:

$$\mu_j = y_j\left(A - y_i\mu_i\right)$$

which means that we can optimize the dual with respect to μ_i. We denote $\boldsymbol{\mu}_{-(i,j)}$ the n-dimensional vector $\boldsymbol{\mu}$ whose i-th and j-th components are 0, and $\tilde{\mathbf{K}} = \operatorname{Diag}(\mathbf{y})\,\mathbf{K}\operatorname{Diag}(\mathbf{y})$. We first solve the unconstrained problem:

$$\max_{\mu_i\in\mathbb{R}} \left(\boldsymbol{\mu}_{-(i,j)} + \mu_i\mathbf{e}_i + \mu_j\mathbf{e}_j\right)^\top \mathbf{1} - \frac{1}{4\lambda}\left(\boldsymbol{\mu}_{-(i,j)} + \mu_i\mathbf{e}_i + \mu_j\mathbf{e}_j\right)^\top \tilde{\mathbf{K}}\left(\boldsymbol{\mu}_{-(i,j)} + \mu_i\mathbf{e}_i + \mu_j\mathbf{e}_j\right)$$

where $\mu_j = y_j\left(A - y_i\mu_i\right)$. This problem is equivalent to:

$$\max_{\mu_i\in\mathbb{R}} \mu_i + \mu_j - \frac{1}{4\lambda}\left[\left(2\boldsymbol{\mu}_{-(i,j)}^\top\tilde{\mathbf{K}}\mathbf{e}_i\right)\mu_i + \left(2\boldsymbol{\mu}_{-(i,j)}^\top\tilde{\mathbf{K}}\mathbf{e}_j\right)\mu_j + \tilde{\mathbf{K}}_{ii}\mu_i^2 + 2\tilde{\mathbf{K}}_{ij}\mu_i\mu_j + \tilde{\mathbf{K}}_{jj}\mu_j^2\right]$$

or, using the expression of μ_j and the fact that $y_i^2 = y_j^2 = 1$:

$$\begin{aligned} \max_{\mu_i\in\mathbb{R}} \left(1 - y_iy_j\right)\mu_i - \frac{1}{4\lambda}\Big[&\left(2\boldsymbol{\mu}_{-(i,j)}^\top\tilde{\mathbf{K}}\left(\mathbf{e}_i - y_iy_j\mathbf{e}_j\right)\right)\mu_i + \\ &\left(\tilde{\mathbf{K}}_{ii} + \tilde{\mathbf{K}}_{jj} - 2y_iy_j\tilde{\mathbf{K}}_{ij}\right)\mu_i^2 + 2A\left(y_j\tilde{\mathbf{K}}_{ij} + \tilde{\mathbf{K}}_{jj}y_i\right)\mu_i\Big] \end{aligned}$$

This can be written as:

$$\max_{\mu_i \in \mathbb{R}} a\mu_i^2 + b\mu_i$$

where $a \leq 0, b \in \mathbb{R}$. The maximum is reached for $\mu_i^{unconst.} = -\frac{b}{2a}$. Hence, given the previous constraints, the update for μ_i and μ_j is given by:

$$\boxed{\begin{aligned} &\mu_i = \min\left[\max\left(0, \mu_i^{unconst.}\right), \frac{1}{n}\right] \\ &\mu_j = y_j\left(A - y_i\mu_i\right) \end{aligned}}$$

Exercice 5. Duality

Let $(x_1, y_1), \ldots, (x_n, y_n)$ a training set of examples where $x_i \in \mathcal{X}$, a space endowed with a positive definite kernel K, and $y_i \in \{-1, 1\}$, for $i = 1, \ldots, n$. $\mathcal{H}_K$ denotes the RKHS of the kernel K. We want to learn a function $f : \mathcal{X} \mapsto \mathbb{R}$ by solving the following optimization problem:

$$\min_{f \in \mathcal{H}_K} \frac{1}{n} \sum_{i=1}^{n} \ell_{y_i}\left(f(x_i)\right) \quad \text{such that} \quad \| f \|_{\mathcal{H}_K} \leq B\,, \tag{1}$$

where ℓ_y is a convex loss function (for $y \in \{-1, 1\}$) and $B > 0$ is a parameter.

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

   $$\min_{\alpha \in \mathbb{R}^n} R(K\alpha) + \lambda \alpha^\top K \alpha\,, \tag{2}$$

   where K is the $n \times n$ Gram matrix and $R : \mathbb{R}^n \mapsto \mathbb{R}$ should be explicited.

2. Compute the Fenchel-Legendre transform[1] R^* of R in terms of the Fenchel-Legendre transform ℓ_y^* of ℓ_y.

3. Adding the slack variable $u = K\alpha$, the problem (1) can be rewritten as a constrained optimization problem:

   $$\min_{\alpha \in \mathbb{R}^n, u \in \mathbb{R}^n} R(u) + \lambda \alpha^\top K \alpha \quad \text{such that} \quad u = K\alpha\,. \tag{3}$$

   Express the dual problem of (3) in terms of R^*, and explain how a solution to (3) can be found from a solution to the dual problem.

4. Explicit the dual problem for the logistic and squared hinge losses:

   $$\ell_y(u) = \log(1 + e^{-yu})\,.$$

   $$\ell_y(u) = \max(0, 1 - yu)^2\,.$$

[1] For any function $f : \mathbb{R}^N \mapsto \mathbb{R}$, the *Fenchel-Legendre transform* (or *convex conjugate*) of f is the function $f^* : \mathbb{R}^N \mapsto \mathbb{R}$ defined by

$$f^*(u) = \sup_{x \in \mathbb{R}^N} x^\top u - f(x)\,.$$

Solution 5.

1. Let $i \in \llbracket 1, n \rrbracket$. By the reproducing property, we have $f(x_i) = \langle f, \mathbf{K}_{x_i} \rangle$. The function $f \mapsto \ell_{y_i}(\langle f, \mathbf{K}_{x_i} \rangle)$ is the composition of a convex function and a linear function, hence it is convex. The objective function being a positively weighted sum of convex functions, it is convex as well. The (convex) inequality constraint can be equivalently written as $\|f\|_{\mathcal{H}_K}^2 \leq B^2$. Hence, this is a convex optimization problem, and as the Slater's condition is verified (e.g. with $f = 0$), strong duality holds.

   Given a dual optimal λ^*, a primal optimal f^* minimizes the Lagrangian $f \mapsto \mathcal{L}(f, \lambda^*)$, where:

$$\mathcal{L}(f, \lambda^*) = \frac{1}{n} \sum_{i=1}^{n} \ell_{y_i}(f(x_i)) + \lambda^* \|f\|_{\mathcal{H}_K}^2 - \lambda^* B^2$$

   Hence, one can find the solution to the problem (1) by solving:

$$\min_{f \in \mathcal{H}_K} \frac{1}{n} \sum_{i=1}^{n} \ell_{y_i}(f(x_i)) + \lambda^* \|f\|_{\mathcal{H}_K}^2$$

   By the representer theorem, there exists $\alpha \in \mathbb{R}^n$ such that the solution f^* to this problem has the form $f^* = \sum_{i=1}^{n} \alpha_i \mathbf{K}_{x_i}$. We denote K the kernel matrix. As $f(x_i) = [K\alpha]_i$ and $\|f\|_{\mathcal{H}_K}^2 = \alpha^\top K \alpha$, the previous optimization problem is equivalent to:

$$\boxed{\min_{\alpha \in \mathbb{R}^n} R(K\alpha) + \lambda^* \alpha^\top K \alpha, \text{ where } R : u \longmapsto \frac{1}{n} \sum_{i=1}^{n} \ell_{y_i}(u_i)}$$

2. Let $u \in \mathbb{R}^n$. We have:

$$\begin{aligned} R^*(u) &= \sup_{x \in \mathbb{R}^n} x^\top u - R(x) \\ &= \sup_{x \in \mathbb{R}^n} \sum_{i=1}^{n} x_i u_i - \frac{1}{n} \sum_{i=1}^{n} \ell_{y_i}(x_i) \\ &= \frac{1}{n} \sup_{x \in \mathbb{R}^n} \sum_{i=1}^{n} (n x_i u_i - \ell_{y_i}(x_i)) \\ &= \frac{1}{n} \sum_{i=1}^{n} \sup_{x_i \in \mathbb{R}} (n x_i u_i - \ell_{y_i}(x_i)) \end{aligned}$$

   Finally,

$$\boxed{\forall u \in \mathbb{R}^n, R^*(u) = \frac{1}{n} \sum_{i=1}^{n} \ell_{y_i}^*(n u_i)}$$

3. Let $\alpha \in \mathbb{R}^n, u \in \mathbb{R}^n, \nu \in \mathbb{R}^n$. The Lagrangian $\mathcal{L}$ of the problem is given by:

$$\mathcal{L}(\alpha, u, \nu) = R(u) + \lambda \alpha^\top K \alpha + \nu^\top (K\alpha - u)$$

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

$$\nabla_\alpha \mathcal{L} = 2\lambda K\alpha + K\nu = 0 \quad \text{i.e.} \quad \alpha = -\frac{1}{2\lambda}\nu$$

Hence,

$$\inf_{\alpha\in\mathbb{R}^n} \mathcal{L}(\alpha, u, \nu) = R(u) - \nu^\top u - \frac{1}{4\lambda}\nu^\top K\nu$$

Then, minimizing with respect to u:

$$\begin{aligned}\inf_{\alpha\in\mathbb{R}^n, u\in\mathbb{R}^n} \mathcal{L}(\alpha, u, \nu) &= \inf_{u\in\mathbb{R}^n}\big(R(u) - \nu^\top u\big) - \frac{1}{4\lambda}\nu^\top K\nu \\ &= -\sup_{u\in\mathbb{R}^n}\big(\nu^\top u - R(u)\big) - \frac{1}{4\lambda}\nu^\top K\nu \\ &= -R^*(\nu) - \frac{1}{4\lambda}\nu^\top K\nu\end{aligned}$$

Hence, the dual problem is:

$$\max_{\nu\geq 0} -R^*(\nu) - \frac{1}{4\lambda}\nu^\top K\nu$$

which is equivalent to:

$$\boxed{\min_{\nu\geq 0} R^*(\nu) + \frac{1}{4\lambda}\nu^\top K\nu}$$

Once we have a solution ν^* to the dual problem, we can find a solution to (3) with:

$$\boxed{\alpha^* = -\frac{1}{2\lambda}\nu^*}$$

4. - Logistic loss $\ell_y(u) = \log(1 + e^{-yu})$.

  For $x \in \mathbb{R}$, $\ell_y^*(x) = \sup_{u\in\mathbb{R}}\{xu - \log(1 + e^{-yu})\} \triangleq \sup_{u\in\mathbb{R}} g_x(u)$.

  - If $y = 1$. If $x > 0$, $\ell_1^*(x) = +\infty$. If $x = 0$, $\ell_1^*(x) = \sup_{u\in\mathbb{R}}\left\{\log\left(\frac{1}{1+e^{-u}}\right)\right\} = 0$. Let us suppose $x < 0$. We have:

    $$xu - \log\left(1 + e^{-u}\right) = \log\left(\frac{e^{xu}}{1+e^{-u}}\right) = \log\left(\frac{e^{(x+1)u}}{e^u+1}\right) = (x+1)u + \log\left(\frac{1}{e^u+1}\right)$$

    If $x < -1$, $\ell_1^*(x) = +\infty$. If $x = -1$, $\ell_1^*(x) = 0$. If $x \in (-1, 0)$, then we can compute the derivative of g_x with respect to u:

    $$g_x'(u) = x + \frac{e^{-u}}{1+e^{-u}}$$

    We have:

    $$\begin{aligned} g_x'(u) = 0 &\text{ iff } x + xe^{-u} + e^{-u} = 0 \\ &\text{iff } e^u = -\frac{x+1}{x} \\ &\text{iff } u = \log\left(-\frac{x+1}{x}\right)\end{aligned}$$

This gives:

$$\begin{aligned}\ell_1^*(x) &= x\log\left(-\frac{x+1}{x}\right) - \log\left(1-\frac{x}{x+1}\right)\\ &= x\log(x+1) - x\log(-x) + \log(x+1)\\ &= -x\log(-x) + (x+1)\log(x+1)\end{aligned}$$

Hence, we have:

$$\ell_1^*(x) = \begin{cases} -x\log(-x) + (x+1)\log(x+1) & \text{if } x \in (-1,0)\\ 0 & \text{if } x \in \{-1,0\}\\ +\infty & \text{if } x < -1 \text{ or } x > 0\end{cases}$$

- If $y=-1$. Similarly, we can show that:

$$\ell_{-1}^*(x) = \begin{cases} x\log(x) + (1-x)\log(1-x) & \text{if } x \in (0,1)\\ 0 & \text{if } x \in \{0,1\}\\ +\infty & \text{if } x < 0 \text{ or } x > 1\end{cases}$$

Finally, we have:

$$\ell_y^*(x) = \begin{cases} -yx\log(-yx) + (yx+1)\log(yx+1) & \text{if } -1 < yx < 0\\ 0 & \text{if } x \in \{-y,0\}\\ +\infty & \text{otherwise}\end{cases}$$

Notice that 0 is the limit of the function $t \mapsto -t\log(-t) + (t+1)\log(t+1)$ in -1^+ and 0^-, which can then be extended to $[-1,0]$. Using the expression of R^* from question 2, the dual problem for the logistic loss is given by:

$$\boxed{\begin{aligned}&\min_{\nu\geq 0} \frac{1}{n}\sum_{i=1}^n \left(-ny_i\nu_i\log(-ny_i\nu_i) + (ny_i\nu_i+1)\log(ny_i\nu_i+1)\right) + \frac{1}{4\lambda}\nu^\top K\nu\\ &\text{s.t.} \quad -\frac{1}{n} \leq y_i\nu_i \leq 0 \quad \text{for} \quad i=1,\dots,n\end{aligned}}$$

- Squared hinge loss $\ell_y(u) = \max(0, 1-yu)^2$.

  For $x \in \mathbb{R}$, $\ell_y^*(x) = \sup_{u\in\mathbb{R}}\left\{xu - \max(0,1-yu)^2\right\} \triangleq \sup_{u\in\mathbb{R}} g_x(u)$.

  - If $y=1$. If $x>0$, $\ell_1^*(x) = +\infty$. If $x=0$, $\ell_1^*(x)=0$. Let us suppose $x<0$. We have

  $$\ell_1^*(x) = \sup_{u\in\mathbb{R}}\left\{xu - \max(0,1-u)^2\right\} = \sup_{u\leq 0}\left\{xu - (1-u)^2\right\}$$

  For $u \leq 0$, the derivative of g_x is given by $g_x'(u) = x + 2(1-u)$ and:

  $$g_x'(u) = 0 \quad \text{iff} \quad u = \frac{x}{2}+1$$

This gives:

$$\begin{aligned}\ell_1^*(x) &= x\left(\frac{x}{2}+1\right)-\left(1-\left(\frac{x}{2}+1\right)\right)^2\\ &= \frac{x^2}{2}+x-\frac{x^2}{4}\\ &= \frac{x^2}{4}+x\end{aligned}$$

Hence, we have:

$$\ell_1^*(x) = \begin{cases}\frac{x^2}{4}+x & \text{if } x \leq 0\\ +\infty & \text{if } x > 0\end{cases}$$

- If $y = -1$. Similarly, we can show that:

$$\ell_{-1}^*(x) = \begin{cases}\frac{x^2}{4}-x & \text{if } x \geq 0\\ +\infty & \text{if } x < 0\end{cases}$$

Finally, we have:

$$\ell_y^*(x) = \begin{cases}\frac{x^2}{4}+yx & \text{if } yx \leq 0\\ +\infty & \text{if } yx > 0\end{cases}$$

The dual problem for the square hinged loss is given by:

$$\boxed{\min_{\nu \geq 0} \frac{n}{4}\sum_{i=1}^{n}\nu_i^2 + \sum_{i=1}^{n} y_i\nu_i + \frac{1}{4\lambda}\nu^\top K\nu}$$