MVA "Kernel methods in machine learning" Exercices

Michael Arbel, Julien Mairal, and Jean-Philippe Vert

Exercice 1. Kernels

Study whether the following kernels are positive definite:

1. $\mathcal{X} = (-1, 1), \quad K(x, x') = \frac{1}{1 - xx'}$
2. $\mathcal{X} = \mathbb{N}, \quad K(x, x') = 2^{x+x'}$
3. $\mathcal{X} = \mathbb{N}, \quad K(x, x') = 2^{xx'}$
4. $\mathcal{X} = \mathbb{R}_+, \quad K(x, x') = \log(1 + xx')$
5. $\mathcal{X} = \mathbb{R}, \quad K(x, x') = \exp(-|x - x'|^2)$
6. $\mathcal{X} = \mathbb{R}, \quad K(x, x') = \cos(x + x')$
7. $\mathcal{X} = \mathbb{R}, \quad K(x, x') = \cos(x - x')$
8. $\mathcal{X} = \mathbb{R}_+, \quad K(x, x') = \min(x, x')$
9. $\mathcal{X} = \mathbb{R}_+, \quad K(x, x') = \max(x, x')$
10. $\mathcal{X} = \mathbb{R}_+, \quad K(x, x') = \min(x, x')/\max(x, x')$
11. $\mathcal{X} = \mathbb{N}, \quad K(x, x') = GCD(x, x')$
12. $\mathcal{X} = \mathbb{N}, \quad K(x, x') = LCM(x, x')$
13. $\mathcal{X} = \mathbb{N}, \quad K(x, x') = GCD(x, x')/LCM(x, x')$
14. Given a probability space $(\Omega, \mathcal{A}, P)$, on $\mathcal{X} = \mathcal{A}$:

$$\forall A, B \in \mathcal{A}, \quad K(A, B) = P(A \cap B) - P(A)P(B).$$

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

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

16. 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 $\frac{0}{0} = 0$.

Exercice 2. Function and kernel boundedness

Consider a p.d. kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that $K(x, z) \leq b^2$ for all x, z in $\mathcal{X}$. Show that $\|f\|_\infty = \sup_{x \in \mathcal{X}} |f(x)| \leq b$ for any function f in the unit ball of the corresponding RKHS.

Exercice 3. Non-expansiveness of the Gaussian kernel

Consider the Gaussian kernel $K : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ such that for all pair of points x, x' in $\mathbb{R}^p$,

$$K(x, x') = e^{-\frac{\alpha}{2}\|x - x'\|^2},$$

where $\|.\|$ is the Euclidean norm on $\mathbb{R}^p$. Call $\mathcal{H}$ the RKHS of K and consider its RKHS mapping $\varphi : \mathbb{R}^p \to \mathcal{H}$ such that $K(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}}$ for all x, x' in $\mathbb{R}^p$. Show that

$$\|\varphi(x) - \varphi(x')\|_{\mathcal{H}} \leq \sqrt{\alpha}\|x - x'\|.$$

The mapping is called non-expansive whenever $\alpha \leq 1$.

Exercice 4. 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
- $K(x, x') = K(x', x'') = 1 \Rightarrow K(x, x'') = 1$.

Exercice 5. RKHS

1. Let K_1 and K_2 be two positive definite kernels on a set $\mathcal{X}$, and α, β two positive scalars. Show that $\alpha K_1 + \beta K_2$ is positive definite, and describe its RKHS.

2. Let $\mathcal{X}$ be a set and $\mathcal{F}$ be a Hilbert space. Let $\Psi : \mathcal{X} \to \mathcal{F}$, and $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be:

$$\forall x, x' \in \mathcal{X}, \quad K(x, x') = \langle \Psi(x), \Psi(x') \rangle_{\mathcal{F}} \, .$$

Show that K is a positive definite kernel on $\mathcal{X}$, and describe its RKHS.

3. Prove that for any p.d. kernel K on a space $\mathcal{X}$, a function $f : \mathcal{X} \to \mathbb{R}$ belongs to the RKHS $\mathcal{H}$ with kernel K if and only if there exists $\lambda > 0$ such that $K(\mathbf{x}, \mathbf{x}') - \lambda f(\mathbf{x}) f(\mathbf{x}')$ is p.d.

Exercice 6. Completeness of the RKHS

We want to finish the construction of the RKHS associated to a positive definite kernel K given in the course. Remember we have defined the set of functions:

$$\mathcal{H}_0 = \left\{ \sum_{i=1}^{n} \alpha_i K_{x_i} \, : \, n \in \mathbb{N}, \alpha_1, \ldots, \alpha_n \in \mathbb{R}, x_1, \ldots, x_n \in \mathcal{X} \right\}$$

and for any two functions $f, g \in \mathcal{H}_0$, given by:

$$f = \sum_{i=1}^{m} a_i K_{\mathbf{x}_i}, \quad g = \sum_{j=1}^{n} b_j K_{\mathbf{y}_j} \, ,$$

we have defined the operation:

$$\langle f, g \rangle_{\mathcal{H}_0} := \sum_{i,j} a_i b_j K\left(\mathbf{x}_i, \mathbf{y}_j\right) .$$

In the course we have shown that $\mathcal{H}_0$ endowed with this inner product is a pre-Hilbert space. Let us now show how to finish the construction of the RKHS from $\mathcal{H}_0$

1. Show that any Cauchy sequence (f_n) in $\mathcal{H}_0$ converges pointwisely to a function $f : \mathcal{X} \to \mathbb{R}$ defined by $f(x) = \lim_{n\to+\infty} f_n(x)$.

2. Show that any Cauchy sequence $(f_n)_{n\in\mathbb{N}}$ in $\mathcal{H}_0$ which converges pointwise to 0 satisfies:
$$\lim_{n\to+\infty} \| f_n \|_{\mathcal{H}_0} = 0 \, .$$

3. Let $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ be the set of functions $f : \mathcal{X} \to \mathbb{R}$ which are pointwise limits of Cauchy sequences in $\mathcal{H}_0$, i.e., if (f_n) is a Cauchy sequence in $\mathcal{H}_0$, then $f(x) = \lim_{n\to+\infty} f_n(x)$. Show that $\mathcal{H}_0 \subset \mathcal{H}$.

4. If (f_n) and (g_n) are two Cauchy sequences in $\mathcal{H}_0$, which converge pointwisely to two functions f and $g \in \mathcal{H}$, show that the inner product $\langle f_n, g_n \rangle_{\mathcal{H}_0}$ converges to a number which only depends on f and g. This allows us to define formally the operation:
$$\langle f, g \rangle_{\mathcal{H}} = \lim_{n\to+\infty} \langle f_n, g_n \rangle_{\mathcal{H}_0} \, .$$

5. Show that $\langle ., . \rangle_{\mathcal{H}}$ is an inner product on $\mathcal{H}$.

6. Show that $\mathcal{H}_0$ is dense in $\mathcal{H}$ (with respect to the metric defined by the inner product $\langle ., . \rangle_{\mathcal{H}}$)

7. Show that $\mathcal{H}$ is complete.

8. Show that $\mathcal{H}$ is a RKHS whose reproducing kernel is K.

Exercice 7. Uniqueness of the RKHS

Prove that if $K : \mathcal{X} \times \mathcal{X}$ is a positive definite function, then it is the r.k. of a unique RKHS. (Hint: consider the linear space spanned by the functions $K_x : t \mapsto K(x,t)$, and use the fact that a linear subspace $\mathcal{F}$ of a Hilbert space $\mathcal{H}$ is dense in $\mathcal{H}$ if and only 0 is the only vector orthgonal to all vectors in $\mathcal{F}$)

Exercice 8. Conditionally positive definite kernels

Let $\mathcal{X}$ be a set. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called *conditionally positive* *definite* (c.p.d.) if and only if it is symmetric and satisfies:

$$\sum_{i,j=1}^{n} a_i a_j k(x_i, x_j) \geq 0$$

for any $n \in \mathbb{N}$, $x_1, x_2, \ldots, x_n \in \mathcal{X}^n$ and $a_1, a_2, \ldots, a_n \in \mathbb{R}^n$ with $\sum_{i=1}^{n} a_i = 0$
.

1. Show that a positive definite (p.d.) function is c.p.d.

2. Is a constant function p.d.? Is it c.p.d.?

3. If $\mathcal{X}$ is a Hilbert space, then is $k(x, y) = -||x - y||^2$ p.d.? Is it c.p.d.?

4. Let $\mathcal{X}$ be a nonempty set, and $x_0 \in \mathcal{X}$ a point. For any function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, let $\tilde{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be the function defined by:

   $$\tilde{k}(x, y) = k(x, y) - k(x_0, x) - k(x_0, y) + k(x_0, x_0).$$

   Show that k is c.p.d. if and only if $\tilde{k}$ is p.d.

5. Let k be a c.p.d. kernel on $\mathcal{X}$ such that $k(x, x) = 0$ for any $x \in \mathcal{X}$. Show that there exists a Hilbert space $\mathcal{H}$ and a mapping $\Phi : \mathcal{X} \to \mathcal{H}$ such that, for any $x, y \in \mathcal{X}$,

   $$k(x, y) = -||\Phi(x) - \Phi(y)||^2.$$

6. Show that if k is c.p.d., then the function $\exp(tk(x, y))$ is p.d. for all $t \geq 0$

7. Conversely, show that if the function $\exp(tk(x, y))$ is p.d. for any $t \geq 0$, then k is c.p.d.

8. Show that the negative shortest-path distance on a tree[1] is c.p.d over the set of vertices (a tree is an undirected graph without loops). Is the negative shortest-path distance over graphs c.p.d. in general?

[1] I.e., the function $k(x, y) = -d(x, y)$, where $d(x, y)$ is the shortest-path distance between x and y, that is, the minimum number of edges of any path that connects x to y.

Exercice 9. 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

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

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:

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

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.

Exercice 10. RKHS-induced semi-metrics

Let $\mathcal{H}$ be a RKHS of functions with domain $\mathcal{X}$, associated to a measurable p.d. kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Consider two probability distributions $\mathbb{P}$ and $\mathbb{Q}$ on $\mathcal{X}$. Show that

$$\sup_{\|f\|_{\mathcal{H}} \leq 1} \left|\mathbb{E}_{\mathbb{P}}[f(X)] - \mathbb{E}_{\mathbb{Q}}[f(Z)]\right|^2 = \mathbb{E}[K(X, X') + K(Z, Z') - 2K(X, Z)],$$

where $X, X' \sim \mathbb{P}$, and $Z, Z' \sim \mathbb{Q}$ are jointly independent.

Exercice 11. Kernel PCA for data denoising

Let $\mathcal{X}$ be a space endowed with a p.d. kernel K, and $\Phi : \mathcal{X} \to \mathcal{H}$ a mapping to a Hilbert space $\mathcal{H}$ such that for all $x, x' \in \mathcal{X}$,

$$\langle \Phi(x), \Phi(x') \rangle = K(x, x')\,.$$

Let $\mathcal{S} = \{x_1, \ldots, x_n\}$ be a set of points in $\mathcal{X}$, and

$$m = \frac{1}{n} \sum_{i=1}^n \Phi(x_i)$$

their barycenter in the feature space.

1. For $x \in \mathcal{X}$, let
$$\Psi(x) = P_d\left(\Phi(x) - m\right) + m$$
where P_d is the projection onto the linear span of the first d kernel principal components of $\mathcal{S}$. Show that $\Psi(x)$ can be expressed as
$$\Psi(x) = \sum_{i=1}^{n} \gamma_i \Phi(x_i)\,,$$
for some γ_i to be explicitly computed.

2. For $y \in \mathcal{X}$, express
$$f(y) = \parallel \Phi(y) - \Psi(x) \parallel^2$$
in terms of kernel evaluations. Explain why minimizing $f(y)$ can be thought of as a method to "denoise" x.

3. Express f and ∇f in the case $\mathcal{X} = \mathbb{R}^p$ and $K(x, x') = \exp\left(-\frac{\parallel x - x' \parallel^2}{2\sigma^2}\right)$. Propose an iterative algorithm (for example gradient descent) to find a local minimum of f in that case.

4. Download the USPS ZIP code data from `http://statweb.stanford.edu/~tibs/ElemStatLearn/data.html` Visualize (a subset of) the dataset in two dimensions with kernel PCA, for different kernels. Implement the procedure discussed in question 4, and test it on some data that you have corrupted with noise. Compute how similar the denoised images are from the original (uncorrupted) images as a function of the number of principal components used.

Exercice 12. Kernel k-means

In order to cluster a set of vectors $x_1, \ldots, x_n \in \mathbb{R}^p$ into K groups, we consider the minimization of:
$$C(z, \mu) = \sum_{i=1}^{n} \parallel x_i - \mu_{z_i} \parallel^2$$
over the cluster assignment variable z_i (taking values in $1, \ldots, K$ for all $i = 1, \ldots, n$) and over the cluster means $\mu_i \in \mathbb{R}^p, i = 1, \ldots, K$.

1. Starting from an initial assignment z^0, we can try to minimize $C(z, \mu)$ by iterating:

$$\mu^i = \underset{\mu}{\operatorname{argmin}}\ C(z^i, \mu)\,, \qquad z^{i+1} = \underset{z}{\operatorname{argmin}}\ C(z, \mu^i)\,.$$

   Explicit how both minimization can be carried out (note: this method is called k-means).

2. Propose a similar iterative algorithm to perform k-means in the RKHS $\mathcal{H}$ of a p.d. kernel K over $\mathbb{R}^p$, i.e., to minimize:

$$C_K(z, \mu) = \sum_{i=1}^{n} \| \Phi(x_i) - \mu_{z_i} \|^2\,,$$

   where $\Phi : \mathbb{R}^p \to \mathcal{H}$ satisfies $\Phi(x)^\top \Phi(x') = K(x, x')$.

3. Let Z be the $n \times K$ assignment matrix with values $Z_{ij} = 1$ if x_i is assigned to cluster j, 0 otherwise. Let $N_j = \sum_{i=1}^{n} Z_{ij}$ be the number of points assigned to cluster j, and L be the $K \times K$ diagonal matrix with entries $L_{ii} = 1/N_i$. Show that minimizing $C_K(z, \mu)$ is equivalent to maximizing over the assignment matrix Z the trace of $L^{1/2} Z^\top K Z L^{1/2}$.

4. Let $H = ZL^{1/2}$. What can we say about $H^\top H$? Do you see a connection between kernel k-means and kernel PCA? Propose an algorithm to estimate Z from the solution of kernel PCA.

5. Implement the two variants of kernel k-means (Questions **2** and **4**). Test them with different kernels (linear, Gaussian) on the *Libras Movement Data Set*[2] ($n = 360, p = 90, K = 15$). Visualize the data mapped to the first two principal components for different kernels, and check how well clustering recovers the 15 classes. (note: only use the first 90 attributes for clustering, the 91*st* one is the class label).

Exercice 13. Kernel LDA

Fisher's linear discriminant analysis (LDA) is a method for supervised binary classification of finite-dimensional vectors. Given two sets of points

[2] `http://archive.ics.uci.edu/ml/datasets/Libras+Movement`

$\mathcal{S}_1 = \left\{x_1^1, \ldots, x_{n_1}^1\right\}$ and $\mathcal{S}_2 = \left\{x_1^2, \ldots, x_{n_2}^2\right\}$ in $\mathbb{R}^p$, let us denote by $m_i = \frac{1}{n_i}\sum_{j=1}^{l_i} x_j^i$, and by:

$$S_B = (m_1 - m_2)(m_1 - m_2)^\top \,, \tag{1}$$

$$S_W = \sum_{i=1,2}\sum_{x\in\mathcal{S}_i}(x - m_i)(x - m_i)^\top \,, \tag{2}$$

the *between* and *within* class scatter matrices, respectively. LDA constructs the function

$$f_w(x) = w^\top x \,,$$

where w is the vector which maximizes

$$J(w) = \frac{w^\top S_B w}{w^\top S_W w} \,.$$

1. Why does it make sense to maximize $J(w)$? What do we expect to find? (you can take as example the case where the two sets $\mathcal{S}_1$ and $\mathcal{S}_2$ form two clusters, e.g., two Gaussians).

2. We want to extend LDA to the feature space $\mathcal{H}$ induced by a positive definite kernel K by the relations $K(x, x') =< \Phi(x), \Phi(x') >_{\mathcal{H}}$. For a vector $w \in \mathcal{H}$ that is a linear combination of the form

$$w = \sum_{i=1,2}\sum_{j=1}^{n_i} \alpha_j^i \Phi(x_j^i) \,,$$

express $J(w)$ and $f_w(x)$ as a function of α and K.

Exercice 14. Rademacher complexity

A Rademacher variable is a random variables σ that can take two possible values, -1 and $+1$, with equal probability $1/2$.

1. Let $(u_1, u_2, \ldots, u_N)$ be N vectors in a Hilbert space endowed with an inner product $< .,. >$, and let $\sigma_1, \sigma_2, \ldots, \sigma_N$ be N independent Rademacher variables. Show that:

$$E\left(\sum_{i=1}^{N}\sum_{j=1}^{N} \sigma_i\sigma_j < u_i, u_j >\right) = \sum_{i=1}^{N} \| u_i \|^2 \,.$$

2. Let K be a positive definite kernel on a space $\mathcal{X}$, $\mathcal{H}_K$ denote the associated reproducing kernel Hilbert space, and $B_R = \{f \in \mathcal{H}_K, \| f \|_{\mathcal{H}_K} \leq R\}$. Let a set of points $\mathcal{S} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$ with $\mathbf{x}_i \in \mathcal{X}$ $(i = 1, \ldots, N)$, and let $\sigma_1, \sigma_2, \ldots, \sigma_N$ be N independent Rademacher variables. Show that:

$$E \sup_{f \in B_R} \left| \sum_{i=1}^{N} \sigma_i f(\mathbf{x}_i) \right| \leq R \sqrt{\sum_{i=1}^{N} K(\mathbf{x}_i, \mathbf{x}_i)}\,.$$

Exercice 15. Some upper bounds for learning theory

Let K be a positive definite kernel on a measurable set $\mathcal{X}$, $(\mathcal{H}_K, \| \,.\, \|_{\mathcal{H}_K})$ denote the corresponding reproducing kernel Hilbert space, $\lambda > 0$, and $\varphi : \mathbb{R} \to \mathbb{R}$ a function. We assume that:

$$\kappa = \sup_{\mathbf{x} \in \mathcal{X}} K(\mathbf{x}, \mathbf{x}) < +\infty\,,$$

and we note $B_R = \{f \in \mathcal{H}_K, \| f \|_{\mathcal{H}_K} \leq R\}$. Let us define, for all $f \in \mathcal{H}$ and $\mathbf{x} \in \mathcal{X}$,

$$R_\varphi(f, \mathbf{x}) = \varphi(f(\mathbf{x})) + \lambda \| f \|^2_{\mathcal{H}_K}\,.$$

1. φ is said to be Lipschitz if there exists a constant $L > 0$ such that, for all $u, v \in \mathbb{R}$, $|\varphi(u) - \varphi(v)| \leq L |u - v|$. Show that, in that case, there exists a constant C_1 to be determined such that, for all $\mathbf{x} \in \mathcal{X}$ and $f, g \in B_R$:

$$|R_\varphi(f, \mathbf{x}) - R_\varphi(g, \mathbf{x})| \leq C_1 \| f - g \|_{\mathcal{H}_K}\,.$$

2. φ is said to be convex if for all $u, v \in \mathbb{R}$ and $t \in [0, 1]$, $\varphi(tu + (1-t)v) \leq t\varphi(u) + (1-t)\varphi(v)$. We assume that φ is convex, and that for all $\mathbf{x} \in \mathcal{X}$, there exists $f_\mathbf{x} \in \mathcal{H}$ which minimizes $f \mapsto R_\varphi(f, \mathbf{x})$. Show that there exists a constant $C_2 > 0$ to be determined, such that:

$$\psi(f, \mathbf{x}) \stackrel{\Delta}{=} R_\varphi(f, \mathbf{x}) - R_\varphi(f_\mathbf{x}, \mathbf{x}) \geq C_2 \| f - f_\mathbf{x} \|^2_{\mathcal{H}_K}\,.$$

3. Under the hypothesis of questions **2.1** and **2.2**, show that there exists a constant C, to be determined, such that if X is a random variable with values in $\mathcal{X}$, then:

$$\forall f \in B_R, \quad E\psi(f, X)^2 \leq C E\psi(f, X)\,.$$

Exercice 16. Dual coordinate ascent algorithms for SVMs

1. We recall the primal formulation of SVMs seen in the class (slide 142).

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

   and its dual formulation (slide 152)

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

   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?

Exercice 17. 2-SVM

The 2-SVM algorithm is a method for supervised binary classification. Given a training set $(x_i, y_i)_{i=1,\ldots,n}$ of training patterns $x_1, \ldots, x_n$ in a space X endowed with a positive definite kernel K, and a set of corresponding labels $y_1, \ldots, y_n \in \{-1, 1\}$, it solves the following problem:

$$\min_{f \in H_K} \left\{ \frac{1}{n} \sum_{i=1}^{n} L(f(x_i), y_i) + \lambda ||f||^2 \right\},$$

where $||f||$ is the norm of f in the RKHS H_K of the kernel K, and L is the *square* hinge loss function:

$$L(u, y) = \max(1 - uy, 0)^2 .$$

Write the primal and dual problems associated to the 2-SVM, and compare the result with the SVM studied in the course.

Exercice 18. Kernel mean embedding

Let us consider a Borel probability measure P of some random variable X on a compact set $\mathcal{X}$. Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a continuous, bounded, p.d. kernel and $\mathcal{H}$ be its RKHS. The kernel mean embedding of P is defined as the function

$$\begin{aligned} \mu(P) : \mathcal{X} &\to \mathbb{R} \\ y &\mapsto \mathbb{E}_{X\sim P}[K(X, y)]. \end{aligned}$$

1. Show that $\mu(P)$ is in $\mathcal{H}$ and that $\mathbb{E}_{X\sim P}[f(X)] = \langle f, \mu(P)\rangle_{\mathcal{H}}$ for any $f \in \mathcal{H}$.
   Remark: If P and Q are two Borel probability measures, then

   $$\mu(P) = \mu(Q) \quad \textit{implies} \quad \{\mathbb{E}_{X\sim P}[f(X)] = \mathbb{E}_{X\sim Q}[f(X)] \quad \textit{for all} \quad f \in \mathcal{H}\} .$$

   When $\mathcal{H}$ is dense in the space of continuous bounded functions on $\mathcal{X}$, this relation is sufficient to show that $P = Q$. Hence, the kernel mean embedding (single point in the RKHS!) carries all information about the distribution. We call such kernels "universal". It is possible to show that the Gaussian kernel is universal.

2. Consider the empirical distribution

   $$P_{\mathcal{S}} = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i},$$

   where $\mathcal{S} = \{x_1, \ldots, x_n\}$ is a finite subset of $\mathcal{X}$ and δ_{x_i} is a Dirac distribution centered at x_i. Show that

   $$\mathbb{E}_{\mathcal{S}}[\|\mu(P) - \mu(P_{\mathcal{S}})\|_{\mathcal{H}}] \le \frac{4\sqrt{\mathbb{E}K(X, X)}}{\sqrt{n}},$$

where $\mathbb{E}_{\mathcal{S}}$ is the expectation by randomizing over the training set (each x_i is a r.v. distributed according to P). Remember that you are allowed to (and you should!) use any existing result from the slides.

3. Consider the quantity

$$MMD(\mathcal{S}_1, \mathcal{S}_2) = \mathbb{E}_{\mathcal{S}}[\|\mu(P_{\mathcal{S}_1}) - \mu(P_{\mathcal{S}_2})\|_{\mathcal{H}}^2$$

for two sets $\mathcal{S}_1 = (x_1, \ldots, x_n)$ and $\mathcal{S}_2 = (y_1, \ldots, y_m)$. Show that

$$MMD(\mathcal{S}_1, \mathcal{S}_2) = \left(\sup_{\|f\|_{\mathcal{H}} \leq 1} \left\{ \frac{1}{n} \sum_{i=1}^{n} f(x_i) - \frac{1}{m} \sum_{j=1}^{m} f(y_j) \right\} \right)^2 ,$$

and give a formula for this quantity in terms of kernel evaluations only. *Remark: this is called the maximum mean discrepancy criterion, which can be used for statistical testing (are $\mathcal{S}_1$ and $\mathcal{S}_2$ coming from the same distribution?).*

4. We consider $\mathcal{X} = \mathbb{R}^d$ and the Gaussian kernel with bandwidth σ: $K(x, y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$. For any two sets $\mathcal{S}_1$ and $\mathcal{S}_2$, show that $MMD(\mathcal{S}_1, \mathcal{S}_2)$ is an increasing function of σ.

Exercice 19. Sobolev spaces

1. Let

$$\mathcal{H} = \left\{ f : [0,1] \to \mathbb{R} \,,\ \text{absolutely continuous}, f' \in L^2([0,1]), f(0) = 0 \right\} ,$$

endowed with the bilinear form

$$\forall f, g \in \mathcal{H} \,, \quad \langle f, g \rangle_{\mathcal{H}} = \int_0^1 f'(u) g'(u) du \,.$$

Show that $\mathcal{H}$ is an RKHS, and compute its reproducing kernel.

2. Same question when

$$\mathcal{H} = \left\{ f : [0,1] \to \mathbb{R} \,,\ \text{absolutely continuous}, f' \in L^2([0,1]), f(0) = f(1) = 0 \right\} ,$$

3. Same question, when $\mathcal{H}$ is endowed with the bilinear form:

$$\forall f, g \in \mathcal{H}\,, \quad \langle f, g\rangle_{\mathcal{H}} = \int_0^1 \left(f(u)g(u) + f'(u)g'(u)\right) du\,.$$

4. Same question when

$$\mathcal{H} = \left\{f : [0,1] \to \mathbb{R}\,,\ f' \text{ exists and absolutely continuous}, f'' \in L^2([0,1]), f(0) = f('0) = 0\right\}\,,$$

endowed with the bilinear form

$$\forall f, g \in \mathcal{H}\,, \quad \langle f, g\rangle_{\mathcal{H}} = \int_0^1 f''(u)g''(u)du\,.$$

Exercice 20. Splines

Let $H = C_2\left([0,1]\right)$ be the set of twice continuously differentiable functions $f : [0,1] \to \mathbb{R}$, and $H_1 \subset H$ be the set of functions $f \in H$ that satisfy:

$$f(0) = f'(0) = 0.$$

1. Show that H_1 endowed with the norm:

$$\| f \|_{H_1}^2 = \int_0^1 f''(t)^2 dt$$

is a reproducing kernel Hilbert space (RKHS), and compute the reproducing kernel K_1.

2. Let H_2 be the set of affine functions $f : [0,1] \to \mathbb{R}$ (i.e., the functions that can be written as $f(x) = ax + b$, with $a, b \in \mathbb{R}$). Show that H_2 endowed with the norm:

$$\| f \|_{H_2}^2 = f(0)^2 + f'(0)^2$$

is a RKHS and compute the corresponding kernel K_2.

3. Deduce that H endowed with the norm:

$$\| f \|_{H}^2 = \int_0^1 f''(t)^2 dt + f(0)^2 + f'(0)^2$$

is a RKHS and compute the reproducing kernel K.

4. Let $0 < x_1 < \ldots < x_n < 1$ and $(y_1, \ldots, y_n) \in \mathbb{R}^n$. In order to estimate a regression function $f : [0,1] \to \mathbb{R}$, we consider the following optimization problem:

$$\min_{f \in H} \frac{1}{n} \sum_{i=1}^{n} \left(f\left(x_i\right) - y_i\right)^2 + \lambda \int_0^1 f''(t)^2 dt. \tag{3}$$

Show that any solution of (3) can be expanded as:

$$\hat{f}(x) = \sum_{i=1}^{n} \alpha_i K_1(x_i, x) + \beta_1 x + \beta_2,$$

with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)' \in \mathbb{R}^n$ et $\boldsymbol{\beta} = (\beta_0, \beta_1)' \in \mathbb{R}^2$.

5. Let I be the $n \times n$ identity matrix, M be the square $n \times n$ matrix defined by:

$$M_{i,j} = \begin{cases} K_1(x_i, x_j) & \text{si } i \neq j, \\ K_1(x_i, x_j) + n\lambda & \text{si } i = j, \end{cases}$$

T be the $n \times 2$ matrix:

$$T = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix},$$

and $\mathbf{y} = (y_1, \ldots, y_n)'$. Show that $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ satisfy:

$$\begin{cases} T'\boldsymbol{\alpha} = 0, \\ M\boldsymbol{\alpha} + T\boldsymbol{\beta} = \mathbf{y}. \end{cases}$$

6. Deduce that $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are given by:

$$\begin{cases} \boldsymbol{\alpha} = M^{-1}\left(I - T\left(T'M^{-1}T\right)^{-1} T'M^{-1}\right)\mathbf{y}, \\ \boldsymbol{\beta} = \left(T'M^{-1}T\right)^{-1} T'M^{-1}\mathbf{y}. \end{cases}$$

7. Show that

   - $\hat{f} \in C_2\left([0,1]\right)$;

- $\hat{f}$ is a polynomial of degree 3 on each interval $[x_i, x_{i+1}]$ for $i = 1, \ldots, n-1$;
- $\hat{f}$ is an affine function on both intervals $[0, x_1]$ and $[x_n, 1]$.

$\hat{f}$ is called a *spline*.

Exercice 21. 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:

$$\min_{f \in \mathcal{H}_K} \frac{1}{n} \sum_{i=1}^{n} \ell_{y_i}\left(f(x_i)\right) \quad \text{such that} \quad \| f \|_{\mathcal{H}_K} \leq B \,, \tag{4}$$

where ℓ_y is a convex loss functions (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 (7) can be found be solving the following problem:

$$\min_{\alpha \in \mathbb{R}^n} R(K\alpha) + \lambda \alpha^\top K \alpha \,, \tag{5}$$

   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[3] 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 (7) can be rewritten as a constrained optimization problem:

$$\min_{\alpha \in \mathbb{R}^n, u \in \mathbb{R}^n} R(u) + \lambda \alpha^\top K \alpha \quad \text{such that} \quad u = K\alpha \,. \tag{6}$$

   Express the dual problem of (6) in terms of R^*, and explain how a solution to (6) can be found from a solution to the dual problem.

[3] 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

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

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

$$\ell_y(u) = \log(1 + e^{-yu})\,.$$
$$\ell_y(u) = \max(0, 1 - yu)^2\,.$$

Exercice 22. ***B_n*-splines**

The convolution between two functions $f, g : \mathbb{R} \to \mathbb{R}$ is defined by:

$$f \star g(x) = \int_{-\infty}^{\infty} f(u)g(x-u)du,$$

when this integral exists.

Let now the function:

$$I(x) = \begin{cases} 1 & \text{si } -1 \leq x \leq 1, \\ 0 & \text{si } x < -1 \text{ ou } x > 1, \end{cases}$$

and $B_n = I^{\star n}$ for $n \in \mathbb{N}_*$ (that is, the function I convolved n times with itself: $B_1 = I, B_2 = I \star I, B_3 = I \star I \star I$, etc...).

Is the function $k(x, y) = B_n(x - y)$ a positive definite kernel over $\mathbb{R} \times \mathbb{R}$? If yes, describe the corresponding reproducing kernel Hilbert space.

Exercice 23. Semigroup kernels

1. Are the following functions positive definite kernels?

$$\forall x, y \in \mathbb{R}, \quad K_2(x, y) = \frac{1}{2 - e^{-\| x-y \|^2}}$$
$$\forall x, y \in \mathbb{R}, \quad K_3(x, y) = \max\left(0, 1 - |x - y|\right)$$

2. For any $n > 0$, show that the $n \times n$ Hankel matrix $A_{ij} = \frac{1}{1+i+j}$ is positive semidefinite.

3. Describe the functions $\varphi : [0, 1] \mapsto \mathbb{R}$ such that:

$$K(x, y) = \varphi\left(\max(x + y - 1, 0)\right)$$

is a positive definite kernel on $[0, 1]$.

4. Can you describe the functions $\varphi : \mathbb{R}^+ \mapsto \mathbb{R}$ such that:

$$K(x,y) = \varphi\left(\max(x,y)\right)$$

is a positive definite kernel on $\mathbb{R}^+$?

Exercice 24. Gaussian RKHS

For any $\sigma > 0$, let K_σ be the normalized Gaussian kernel on $\mathbb{R}^d$:

$$\forall x, y \in \mathbb{R}^d \quad K_\sigma(x,y) = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^d} \exp\left(-\frac{\| x - y \|^2}{2\sigma^2}\right),$$

and let $\mathcal{H}_\sigma$ be its reproducing kernel Hilbert space (RKHS).

1. Recall a proof of the positive definiteness of K.

2. For any $0 < \sigma < \tau$, show that

$$\mathcal{H}_\tau \subset \mathcal{H}_\sigma \subset L_2(\mathbb{R}^d),$$

3. For any $0 < \sigma < \tau$ and $f \in \mathcal{H}_\tau$, show that

$$\| f \|_{\mathcal{H}_\tau} \geq \| f \|_{\mathcal{H}_\sigma} \geq \| f \|_{L_2(\mathbb{R}^d)},$$

and that

$$0 \leq \| f \|^2_{\mathcal{H}_\sigma} - \| f \|^2_{L_2(\mathbb{R}^d)} \leq \frac{\sigma^2}{\tau^2}\left(\| f \|^2_{\mathcal{H}_\tau} - \| f \|^2_{L_2(\mathbb{R}^d)}\right).$$

4. For any $\tau > 0$ and $f \in \mathcal{H}_\tau$, show that

$$\lim_{\sigma \to 0} \| f \|_{\mathcal{H}_\sigma} = \| f \|_{L_2(\mathbb{R}^d)}.$$

Exercice 25. Kernel for sets

We wish to construct positive definite kernels for finite sets of points in the interval $[0,1]$. Let $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_n)$ be two such sets of length n and m.

1. Show that the following kernel is positive definite for any $\sigma > 0$:

$$K_1(X,Y) = \sum_{x\in X}\sum_{y\in Y} \exp\left(-\frac{(x-y)^2}{2\sigma^2}\right).$$

2. To any finite set X of length n we associate the function $g_X : \mathbb{R} \to \mathbb{R}$ defined by:

$$g_X(t) = \frac{1}{n}\sum_{x\in X} \exp\left(-\frac{(x-t)^2}{2\sigma^2}\right).$$

Show that the following kernel is positive definite for any $\sigma > 0$:

$$K_2(X,Y) = \int_{\mathbb{R}} g_X(t) g_Y(t) dt\,.$$

Is there a simple relation between $K_1(X,Y)$ and $K_2(X,Y)$?

3. Let $\mathcal{P}$ be a partition of $[0,1]$. For any bin $p \in \mathcal{P}$, let $n_p(X)$ be the number of points of X which are in p. Show that the following kernels are positive definite:

$$K_3(X,Y) = \sum_{p\in\mathcal{P}} \min(n_p(X), n_p(Y))\,,$$

$$K_4(X,Y) = \prod_{p\in\mathcal{P}} \min(n_p(X), n_p(Y))\,.$$

4. Let T_D be a complete binary tree of depth D, that is, a directed graph such that, starting from the root, each node has two children, until the nodes in the D-th generation which have no children (nodes with no children are called *leaves*). The nodes of T_D are denoted $s \in T_D$. How many nodes are there in T_D?

5. We denote by $S(T_D)$ the set of connected subgraphs of T_D which contain the root and such that all their nodes have either 0 or 2 children. What is the size of $S(T_D)$ for $D = 10$?

6. For $0 < p < 1$, we consider the following rule to generate randomly a tree in $S(T_D)$. We start at the root, and give it two children with probability p, and no child with probability $1-p$. If it has no child, then

the process stops and the tree generated is the root only. Otherwise, the same rule is applied independently to both children, which have themselves 0 or 2 children with probability $1-p$ and p. The process is repeated iteratively to all new children, until no more child is generated, or until we reach the D-th generation where nodes have no children with probability 1. For any $T \in S(T_D)$ we denote by $\pi(T)$ the probability of generating T by this process. For any real-valued function h defined over the set of nodes $s \in T_D$, propose a factorization to compute the following sum efficiently:

$$\sum_{T \in S(T_D)} \pi(T) \prod_{s \in leaves(T)} h(s) \, .$$

7. Suppose that each leaf $s \in leaves(T_D)$ is associated to a interval $p(s)$ of $[0,1]$ which together form a partition. For any node $s \in T_D$ we denote by $D(s)$ the set of leaves of T_D which are descendant of s, and we associate to s the subset $p(s) \subset [0,1]$ defined by:

$$p(s) = \bigcup_{l \in D(s)} p(l) \, .$$

   For any $T \in S(T_D)$, show that the following function is a positive definite kernel:

$$K_T(X,Y) = \prod_{s \in leaves(T)} \min(n_{p(s)}(X), n_{p(s)}(Y)) \, .$$

8. Show that the following function is a positive definite kernel and propose an efficient implementation to compute it

$$K_5(X,Y) = \sum_{T \in S(T_D)} \pi(T) K_T(X,Y) \, .$$

Exercice 26. Rademacher complexity of MKL

Given a fixed sample of n points $S = (x_1, \ldots, x_n)$ in a space $\mathcal{X}$, the empirical Rademacher complexity of a set of function $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ is:

$$R(\mathcal{F}) = \frac{1}{n} \mathbb{E} \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \sigma_i f(x_i) \right] ,$$

where the expectation is taken over $\sigma_i \in \{-1, +1\}$ for $i = 1, \ldots, n$, which are independent uniform Rademacher random variables. The following result can be used without proof:

Lemma 1. *For any $n \times n$ symmetric p.s.d. matrix K, and $\sigma = (\sigma_1, \ldots, \sigma_n)^\top$ a vector of independent Rademacher random variables, the following holds:*

$$\forall r \in \mathbb{N}^*, \quad \mathbb{E}\left(\sigma^\top K \sigma\right)^r \leq (2r \, \mathit{trace}\,(K))^r \, .$$

1. Let k be a p.d. kernel over $\mathcal{X}$ with RKHS $\mathcal{H}_k$, K its Gram matrix on S, and $\mathcal{B}(k,t) = \{f \in \mathcal{H}_k \, : \, \| f \|_{\mathcal{H}_k} \leq t\}$. Show that

$$\forall t > 0 \, , \quad R\left(\mathcal{B}(k,t)\right) \leq \frac{t\sqrt{\text{trace}\,(K)}}{n} \, .$$

2. If, in addition, there exists $M > 0$ such that $\forall x \in \mathcal{X}, k(x,x) \leq M^2$, show that

$$\forall t > 0 \, , \quad R\left(\mathcal{B}(k,t)\right) \leq \frac{tM}{\sqrt{n}}$$

3. Let now $k_1, \ldots, k_p$ be p p.d. kernel on $\mathcal{X}$, and $k_\eta = \sum_{i=1}^n \eta_i k_i$ for any $\eta \in \Delta = \{\eta \in \mathbb{R}^p \, : \, \forall i = 1, \ldots, p, \, \eta_i \geq 0 \text{ and } \sum_{i=1}^p \eta_i = 1\}$. Show that k_η is a p.d. kernel for any $\eta \in \Delta$, and that, for any non-zero integer $r \in \mathbb{N}^*$

$$\forall t > 0 \, , \quad R\left(\bigcup_{\eta \in \Delta} \mathcal{B}(k_\eta, t)\right) \leq \frac{t\sqrt{2r}\left(\sum_{i=1}^p \text{trace}\,(K_i)^r\right)^{\frac{1}{2r}}}{n} \, .$$

4. If there exists $M > 0$ such that $\forall i = 1, \ldots, p, \forall x \in \mathcal{X}, k_i(x,x) < M^2$, show that

$$\forall t > 0 \, , \quad R\left(\bigcup_{\eta \in \Delta} \mathcal{B}(k_\eta, t)\right) \leq tM\sqrt{\frac{2e(\ln p + 1)}{n}} \, .$$

5. (Bonus) Prove Lemma 1.

Exercice 27. MKL on a DAG

Let $V = (v_1, \ldots, v_M)$ be the vertices of a directed acyclic graph (DAG). For any $v \in V$, we denote by $D(v) \subset V$ the set of descendants of v (including itself), and let $d_v \geq 0$ be a weight associated to each vertex v. We assume that to each vertex $v \in V$ is associated a positive definite kernel K_v over a space $\mathcal{X}$.

1. Using the notations of the course (slide 159), show that the following *weighted* MKL with the set of kernels $\{K_v : v \in V\}$:

$$\min_{(f_{v_1},\ldots,f_{v_M}) \in \mathcal{H}_{K_{v_1}} \times \ldots \times \mathcal{H}_{K_{v_M}}} \left\{ R\left(\sum_{v \in V} f_v^n\right) + \lambda \left(\sum_{v \in V} d_v \| f_v \|_{\mathcal{H}_{K_v}}\right)^2 \right\}$$

   is equivalent to solving:

$$\min_{\eta \in \Sigma} \min_{f \in \mathcal{H}_{K_\eta}} \left\{ R(f^n) + \lambda \| f \|^2_{\mathcal{H}_{K_\eta}} \right\}$$

   for some set Σ to be determined.

2. We now consider the following variant of MKL which takes the graph structure into account:

$$\min_{(f_{v_1},\ldots,f_{v_M}) \in \mathcal{H}_{K_{v_1}} \times \ldots \times \mathcal{H}_{K_{v_M}}} \left\{ R\left(\sum_{v \in V} f_v^n\right) + \lambda \left(\sum_{v \in V} d_v \left(\sum_{w \in D(v)} \| f_w \|^2_{\mathcal{H}_{K_w}}\right)^{\frac{1}{2}}\right)^2 \right\}. \tag{7}$$

   Can you intuitively explain why we may want to do this, and what we can expect from the solution of this formulation?

3. Show that the MKL formulation (7) is equivalent to solving:

$$\min_{\eta \in \Sigma_V} \min_{f \in \mathcal{H}_{K_\eta}} \left\{ R(f^n) + \lambda \| f \|^2_{\mathcal{H}_{K_\eta}} \right\}$$

   for some set Σ_V to be determined.

4. Show that if the DAG is a tree, then Σ_V is convex. Is it also convex for a general DAG?

Exercice 28. Properties of the dot-product kernel

Consider the dot-product kernel on the sphere $K_1 : \mathbb{S}^{p-1} \times \mathbb{S}^{p-1} \to \mathbb{R}$ such that for all pair of points x, x' in $\mathbb{S}^{p-1}$ (unit sphere of $\mathbb{R}^p$),

$$K_1(x, x') = \kappa(\langle x, x' \rangle),$$

where $\kappa : [-1, 1] \to \mathbb{R}$ is an infinitely differentiable function that admits a polynomial expansion on $[-1, 1]$:

$$\kappa(u) = \sum_{i=0}^{+\infty} a_i u^i, \tag{8}$$

where the a_i's are real coefficients and the sum above is always converging.

1. Show that if all coefficients a_i are non-negative and $\kappa \neq 0$, then K_1 is p.d.

2. If K_1 is p.d., show that the homogeneous dot-product kernel $K_2 : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ is also p.d..

$$K_2(x, x') = \begin{cases} \|x\|\|x'\|\kappa\left(\frac{\langle x, x'\rangle}{\|x\|\|x'\|}\right) & \text{if } \ \|x\| \neq 0 \ \text{ and } \ \|x'\| \neq 0 \\ 0 & \text{otherwise} \end{cases}.$$

Remark: it is in fact possible to show that all coefficients a_i need to be non-negative for the positive definiteness to hold for all dimension p, but we do not ask for a proof of this result, which is due to Shoenberg, 1942.

3. Assume that all coefficients a_i are non-negative (K_1 is thus p.d.) and that $\kappa(1) = \kappa'(1) = 1$. Let $\mathcal{H}$ be the RKHS of K_1 and consider its RKHS mapping $\varphi : \mathbb{S}^{p-1} \to \mathcal{H}$ such that $K_1(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}}$ for all x, x' in $\mathbb{S}^{p-1}$. Show that:

$$\forall x, x' \in \mathbb{S}^{p-1}, \quad \|\varphi(x) - \varphi(x')\|_{\mathcal{H}} \leq \|x - x'\|.$$

4. Find an explicit feature map $\psi : \mathbb{S}^{p-1} \to \ell^2$, where ℓ^2 is the Hilbert space of real-valued sequences (see definition on slide 240), such that for all x, y in $\mathbb{S}^{p-1}$

$$K_1(x, y) = \langle \psi(x), \psi(y) \rangle_{\ell_2}.$$

Hint: remember that $\langle x, y \rangle^2 = \langle xx^\top, yy^\top \rangle_F$, where $\langle ., \rangle_F$ is the Frobenius inner-product. You may want to use the tensor product notation $x^{\otimes 2} = xx^\top$ and its generalization for degrees higher than 2.

5. Let us assume that you have found an explicit feature map ψ in the previous question. Remember from one of our previous homeworks that the RKHS $\mathcal{H}$ of K_1 can be characterized by

$$\mathcal{H} = \{f_w : w \in \ell_2\} \quad \text{such that} \quad f_w : x \mapsto \langle w, \psi(x) \rangle_{\mathcal{H}},$$

with

$$\|f_w\|_{\mathcal{H}}^2 = \inf_{w' \in \ell_2} \left\{ \|w'\|_{\ell_2}^2 : f_w = f_{w'} \right\}.$$

Consider then a function $g_z : \mathbb{S}^{p-1} \to \mathbb{R}$ of the form

$$g_z : x \mapsto \sigma(\langle z, x \rangle)$$

with z in $\mathbb{S}^{p-1}$ and σ admits a polynomial expansion $\sigma(u) = \sum_{i=0}^{+\infty} b_i u^i$. Could you find a sufficient condition on z and on the coefficients b_i for g_z to be in $\mathcal{H}$?

Remark: g_z can be interpreted as a one-layer neural network function. We could ask you to do the same analysis for the homogeneous kernel K_2, but this would be unnecessary technical for this homework which is already too long. This being said, if you found it too short, we're happy to see your analysis of K_2 and the type of functions g_z you will consider.

Exercice 29. Support Vector Classifier

Consider a dataset of N pairs (x_i, y_i) where each x_i is a vector of dimension d and y_i is a binary class, i.e. $y_i \in \{-1, 1\}$. We would like to separate the two classes of samples with a **separating hyper-surface** of equation $f(x_i) + b = 0$ such that $f(x_i) + b \leq 0$ if x_i belongs to the class $y_i = -1$ and $f(x_i) + b \geq 0$ if $y_i = 1$.

I. Maximum margin separator. We consider the simple case where the hyper-surface is a hyperplan of equation $f(x) = w^\top x = 0$. The **maximum margin** classifier finds a hyperplan that separates the two classes while being the furthest away from the data. This can be expressed as the following opitimization problem:

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}\|w\|^2 \\ \text{s.t.} \quad & y_i(x_i^T w + b) \geq 1, i \in \{1, ..., N\} \end{aligned} \tag{9}$$

II. Soft Margin Support Vector Classifier. When the classes are non-separable due to the presence of noise, one approach is to relax the hard constraints $y_i(x_i^T w + b) \geq 1$ to soft ones $y_i(x_i^T w + b) \geq 1 - \xi_i$ where ξ_i is a non-negative tolerance. This allows some outlier data points to fail the margin constraint up to ξ_i. To discourage high values of the tolerance ξ_i an additional penalty is introduced to the problem, thus yielding:

$$\begin{aligned} \min_{w,b,\xi_i} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad & y_i(x_i^T w + b) \geq 1 - \xi_i \\ & \xi_i \geq 0 \end{aligned} \tag{10}$$

III. Kernel Support Vector Classifier. For more complicated problems, the classes cannot be separated by a simple hyperplane. Instead, one needs to find a non-linear hyper-surface of equation $f(x) + b = 0$. To achieve this, we consider functions f that belong to a Reproducing Kernel Hilbert Space $\mathcal{H}$ of kernel k. Such choice allows to represent highly non-linear hyper-surfaces while still solving a convex problem of the form:

$$\begin{aligned} \min_{f,b,\xi_i} \quad & \frac{1}{2}\|f\|^2 + C\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad & y_i(f(x_i) + b) \geq 1 - \xi_i \\ & \xi_i \geq 0 \end{aligned} \tag{11}$$

1. Without providing the details of the calculations:
   (a) Provide an expression for the Lagrangian of the problems in **??????** in terms of N dual parameters $\alpha_i \geq 0$ corresponding the margin inequalities and N dual parameters $\mu_i \geq 0$ corresponding to the positivity constraints on ξ_i whenever applicable.
   (b) Using the optimality condition on the Lagrangian, express the dual problem as a **constrained minimization** over $(\alpha_i)_{i\in\{1,\dots,N\}}$ and express $f(x)$ in terms of α_i and relevant quantities.
   (c) Using Strong duality (KKT conditions), find a condition characterizing the **support vector** points x_i that are on the margin of the separating hyper-surface, i.e. the points satisfying the equation $y_i(f(x_i) + b) = 1$.

2. In the notebook, the classes `MMS` and `SVC` correspond to **??**??. For each class, implement the method `fit` that computes the optimal dual parameters α_i, the parameters w, b and the support vectors.

3. In the notebook, implement the method `kernel` of the class RBF, which takes as input two data matrices X and Y of size $N \times d$ and $M \times d$ and returns a gramm matrix G of shape $N \times M$ whose components are $k(x_i, y_j) = \exp(-\|x_i - y_i\|^2/(2\sigma^2))$. (The fastest solution does not use any for loop!)

4. In the notebook, the class `KernelSVC` corresponds to **??**:
   (a) Implement the method `fit` that computes the optimal dual parameters α_i, the parameter b and the support vectors.
   (b) Implement the method `separating_function` that takes a matrix of shape $N' \times d$ and returns a vector of size N' of evaluations of f.

5. Report the outputs for each code block that performs a classification.

Exercice 30. Kernel Ridge Regression

Given a dataset of N pairs (x_i, y_i) where x_i is a vector of dimension d and y_i is a real number. The regression is the task of finding a function f from $\mathbb{R}^d$ to $\mathbb{R}$ such that $f(x_i) + b \simeq y_i$ for some scalar b. Kernel Ridge Regression can model potentially complex/non-linear dependence between x_i and y_i by assuming the regression function f belongs to an RKHS $\mathcal{H}$ of kernel k and by solving a convex optimization problem:

$$\min_{f_j, b_j} \frac{1}{N} \sum_{i=1}^{N} \|f(x_i) + b - y_i\|^2 + \frac{\lambda}{2} \|f\|_{\mathcal{H}}^2 \tag{12}$$

When the variable y_i is a vector of dimension q a simple extension to **??** consists in finding q functions $(f_j)_{1 \le j \le q}$ in $\mathcal{H}$ and scalars $(b_j)_{1 \le j \le q}$ for regressing each dimension of $(y_i)_{1 \le j \le d}$, i.e. $f_j(x_i) + b_j \simeq (y_i)_j$. This can be achieved by solving the problem of the form:

$$\min_{f,b} \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{q} \|f_j(x_i) + b_j - (y_i)_j\|^2 + \frac{\lambda}{2} \sum_{j=1}^{q} \|f_j\|_{\mathcal{H}}^2 \tag{13}$$

1. Without providing the details of the calculations and using the Representer theorem, provide an equivalent finite-dimensional optimization problem for both **????** and find a closed-form expression for f and b in terms of the solutions of such problems.

2. In the notebook, the classes `KernelRR` and `MultivariateKernelRR` correspond to **????**:
   (a) Implement the method `fit`, which solves the finite dimensional problems obtained by the Representer theorem.
   (b) Implement the method `regression_function` that takes a matrix of shape $M \times d$ and returns a vector of size M of evaluations of f.

3. Report the outputs of each code block that performs a regression.

Exercice 31. Kernel Support Vector Regression
Given a dataset of N pairs (x_i, y_i), where x_i is a vector of dimension d and y_i is a scalar and an RKHS $\mathcal{H}$ of kernel k, the Kernel Support Vector Regression (Kernel SVR) finds a regression function $f \in \mathcal{H}$ and scalar b such that $f(x_i) + b - y_i$ are within and tube of size $\eta > 0$ with some tolerance. More precisely, the Kernel SVR solves the problem:

$$\begin{aligned} \min_{f,b,\xi^+,\xi^-} \quad & \frac{1}{2}\|f\|^2 + C\sum_{i=1}^{N} \xi_i^+ + \xi_i^- \\ \text{s.t.} \quad & y_i - f(x_i) - b \leq \epsilon + \xi_i^+ \\ & -y_i + f(x_i) + b \leq \epsilon + \xi_i^- \\ & \xi_i^+, \xi_i^- \geq 0 \end{aligned} \tag{14}$$

1. Without providing the details of the calculations:
   (a) Provide an expression for the Lagrangian of the problems in **??** in terms of:
   - $2N$ dual parameters $(\alpha_i^+)_{1\leq i\leq N} \geq 0$ and $(\alpha_i^-)_{1\leq i\leq N} \geq 0$ corresponding the tube inequalities $y_i - f(x_i) - b \leq \epsilon + \xi_i^+$ and $-y_i + f(x_i) + b \leq \epsilon + \xi_i^-$
   - N dual parameters μ_i^+ and μ_i^- corresponding to the positivity constraints on ξ_i^+ and ξ_i^-.
   (b) Using the optimality condition on the Lagrangian, express the dual problem as a **constrained minimization** over $(\alpha_i^+)_{i\in\{1,\ldots,N\}}$ and

$(\alpha_i^-)_{1\le i\le N}$, then provide an expression for $f(x)$ in terms of $(\alpha_i^+)_{1\le i\le N}$, $(\alpha_i^-)_{1\le i\le N}$ and relevant quantities.
(c) Using Strong duality (KKT conditions), find a condition characterizing the **support vector** points x_i that are on the boundary of the tube, i.e. the points satisfying the equation $y_i - f(x_i) - b = \eta$ or $-y_i + f(x_i) + b = \eta$.

2. In the notebook, the class `KernelSVR` corresponds to **??**:
(a) Implement the method `fit` that computes the optimal dual parameters α_i^+, α_i^-, the parameter, b and the support vectors.
(b) Implement the method `regression_function` that takes a matrix of shape $M\times d$ and returns a vector of size M of evaluations of f.

3. Report the output of the code block that performs the regression.

Exercice 32. Kernel PCA

One motivation for Kernel PCA is to perform non-linear dimensionality reduction of the data. This is relevant, for instance, when the data is concentrated on a lower dimensional manifold that is not a hyperplane. Given a dataset of N points x_i, the first step for performing kernel PCA is to map each data point x_i to some nonlinear feature $\varphi(x_i)$ in an RKHS $\mathcal{H}$ space corresponding to a kernel $k(x,y) = \langle \varphi(x), \varphi(y)\rangle$. One then define the centered features $\tilde{\varphi}(X_i) = \varphi(X_i) - \frac{1}{N}\sum_{j=1}^N \varphi(X_j)$ and the covariance operator C

$$C = \frac{1}{N}\sum_{i=1}^N \tilde{\varphi}(X_i)\otimes\tilde{\varphi}(X_i)$$

Where $\otimes$ denotes the tensor product associated to the inner-product $\langle .,.\rangle$, i.e. $\otimes$ is a binary operation on $\mathcal{H}\times\mathcal{H}$ such that for any u and v in $\mathcal{H}$, $u\otimes v$ is a linear operator from in $\mathcal{H}$ satisfying $(u\otimes v)f = \langle v, f\rangle u$ for any $f\in\mathcal{H}$.

Kernel PCA, consists in finding non-trivial eigenvectors of the operator C, i.e. elements $v\in\mathcal{H}$ such that $Cv = \lambda v$ for positive λ and $\|v\| = 1$.

1. Show that each non-trivial eigenvector of C can be expressed as a linear combination of the features $\tilde{\varphi}(X_i)$, with a vector of coefficients $\alpha = (\alpha_i)_{1:N}$ being an eigenvector of some square matrix G of size N and satisfying some normalization condition.

2. In the notebook, the class `Kernel_PCA` performs a Kernel PCA given some kernel as input:
   (a) Implement the method `compute_PCA` which finds the top r eigenvectors of the matrix G.
   (b) Implement the method `transform` which takes as input a data matrix of shape $M \times d$ and computes its representation of shape $M \times r$ along the r first eigenvectors of the covariance operator C.

3. Report the output of the code block that performs the PCA. What can you conclude?

4. **(Bonus).** The representation of the data obtained by kernel PCA can be interpreted as an r-dimensional encoding of the data (the encoder). From such encoding, it is possible to reconstruct the original data by solving a multivariate regression problem which can be interpreted as a decoder. The encoding-decoding of the data can be used in tasks such as de-noising. In the notebook `KernelPCA`, the class `Denoiser` achieves this by making use of the classes `KernelPCA` and `MultivariateKernelRR` previously implemented.
   (a) Implement the method `fit` that takes as input a noisy training set and learns both encoder and decoder.
   (b) Implement the method `denoise` which takes as input a noisy test dataset and returns a corresponding de-noised dataset.

5. **(Bonus).** Report the output of the code block that performs de-noising of a subset of MNIST digits dataset. To what extend the de-noising is successful? How can it be improved?