Homework for the Course
"Machine Learning with Kernel Methods"

Maha ELBAYAD
M2 MVA, ENS Cachan
maha.elbayad@student.ecp.fr

1 Combination Rules for Kernels

We consider a set $\mathcal{X}$ and two p.d. kernels $K_1, K_2 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.

For $n \in \mathbb{N}$ and $(x_1, ..., x_n) \in \mathcal{X}^n$ we consider the kernels matrices $\mathbf{K}_1$ and $\mathbf{K}_2$

1. The linear combination: Let α, β be two non-negative scalars, the kernel $K = \alpha K_1 + \beta K_2$ is also positive semi-definite as:

$$(\alpha \mathbf{K}_1 + \beta \mathbf{K}_2)^T = \alpha \mathbf{K}_1^T + \beta K_2^T = \alpha \mathbf{K}_1 + \beta \mathbf{K}_2$$

And for a vector $a \in \mathbb{R}^n$

$$\begin{aligned}\sum_{1 \le i,j \le n} a_i a_j K_{(i,j)} &= \sum_{1 \le i,j \le n} a_i a_j (\alpha K_{1,(i,j)} + \beta K_{2,(i,j)}) \\ &= \alpha \sum_{1 \le i,j \le n} a_i a_j K_{1,(i,j)} + \beta \sum_{1 \le i,j \le n} a_i a_j K_{2,(i,j)} \ge 0\end{aligned}$$

Hence K is a p.d. kernel.

2. The elementwise product:

Symmetry: $\forall x, x' \in \mathcal{X} : K(x, x') = K_1(x, x').K_2(x, x') = K_1(x', x).K_2(x', x) = K(x', x)$

The matrix representation of K associated with $\{x_1, ...x_n\}$ is $\mathbf{K} = \mathbf{K}_1 \odot \mathbf{K}_2$ the hadamard product. We consider the eigendecomposition of $\mathbf{K}_1$

$$\mathbf{K}_1 = U \Lambda U^T = \sum_{e=1}^{n} \lambda_e u_e u_e^T$$

Where $U = (u_1, ..., u_n)$ is a unitary matrix and Λ the diagonal matrix of the non-negatives eigenvalues $(\lambda_1, ...\lambda_n)$.

For $a \in \mathbb{R}^n$:

$$\begin{aligned}a^T \mathbf{K} a &= a^T (\mathbf{K}_1 \odot \mathbf{K}_2) a \\ &= \sum_{1 \le i,j \le n} \sum_e \lambda_e a_i a_j u_e^{(i)} u_e^{(j)} \mathbf{K}_{2,(i,j)} \\ &= \sum_{1 \le i,j \le n} b_i b_j \mathbf{K}_{2,(i,j)},\ b_i = \sum_e \sqrt{\lambda_e} a_i u_e^{(i)} \\ &\ge 0\end{aligned}$$

An other proof would be to consider two random vectors $(X_1, ..X_n)$ and $(Y_1, ..Y_n)$ each with mean zero and respective covariance matrices $\mathbf{K}_1$ and $\mathbf{K}_2$, then $\mathbf{K} = \mathbf{K}_1 \odot \mathbf{K}_2$ is the covariance matrix of the vector $(X_1 Y_1, ...X_n Y_n)$ which mean $\mathbf{K}$ is positive definite.

3. The limit: We consider a sequence $(K_m)_{m\geq 0}$ of p.d. kernels such that:

$$\forall x, y \in \mathcal{X},\ K_m(x,y) \xrightarrow[m\to+\infty]{} K(x,y)$$

The symmetry holds for the limit :

$$\forall x, x' \in \mathcal{X},\ K(x,x') = \lim_{m\to+\infty} K_m(x,x') = \lim_{m\to+\infty} K_m(x',x) = K(x',x)$$

And for $a \in \mathbb{R}^n$:

$$\sum_{1\leq i,j\leq n} a_i a_j K(x_i,x_j) = \lim_{m\to+\infty} \sum_{1\leq i,j\leq n} a_i a_j K_m(x_i,x_j) \geq 0$$

2. The exponenetial: We consider the kernel $K = e^{K_1}$

We can write K as:

$$K = \lim_{N\to+\infty} \left[\sum_{m=0}^{N} \frac{K_1^m}{m!}\right]$$

Using the product rule we can show by induction on $m \in \mathbb{N}$ that K_1^m is a p.d kernel.

By the sum (linear combination with positive scalars) and the limit rules above K is a p.d. kernel.

2 Quizz: Positive Definite Kernels

•$K(x,y) = \dfrac{1}{1-xy},\ \mathcal{X} = (-1,1)$
$\forall x, y \in (-1,1),\ xy \in (-1,1)$ thus we can expand K as the series:

$$K(x,y) = \frac{1}{1-xy} = \sum_{n=1}^{+\infty} (xy)^n$$

Each component of the summation is a monomial of the dot product xy (with coefficent 1) by the sum and the limit rule K **is a p.d. kernel**.

•$K(x,y) = 2^{xy},\ \mathcal{X} = \mathbb{N}$

$$K(x,y) = 2^{xy} = \exp(xy\ln(2))$$

$(x,y) \mapsto \ln(2)xy$ **is a p.d. kernel** and so is its exponential K

•$K(x,y) = \log(1+xy),\ \mathcal{X} = \mathbb{R}_+$

$$K(x,y) = \log(1+xy) = \int_0^{+\infty} (1-e^{-txy})\frac{e^{-t}}{t} d\lambda_t$$

$(x,y) \mapsto -xy$ is negative definite, hence for $t \in \mathbb{R}^+$, the kernel $(x,y) \mapsto \exp(-txy)$ is negative definite which means $(x,y) \mapsto 1 - \exp(-txy)$ is positive definite. Thus the integrand above is positive definite, by the sum rule K **is a p.d. kernel**.

•$K(x,y) = e^{-(x-y)^2},\ \mathcal{X} = \mathbb{R}$

$$K(x,y) = e^{-(x-y)^2} = [e^{-x^2}e^{-y^2}][\sum_{n=0}^{+\infty} 2^n \frac{(xy)^n}{n!}]$$

The first term is a trivial p.d. kernel of the form $K_1(x,y) = g(x)g(y),\ g : \mathcal{X} \to \mathbb{R}$
In fact, K_1 is symmetric and for $\{x_1, ...x_n\} \in \mathcal{X}^n$ and $a \in \mathbb{R}^n$:

$$\sum_{1\leq i,j\leq n} a_i a_j K_1(x_i,x_j) = \sum_{1\leq i,j\leq n} a_i g(x_i) a_j g(x_j) = \left(\sum_{i=1}^{n} a_i g(x_i)\right)^2 \geq 0$$

While the second term is a p.d kernel by the sum and the limit rules. Thus K **is a p.d. kernel**.

$\bullet K(x, y) = cos(x + y),\ \mathcal{X} = \mathbb{R}$
For the set $\{0, \frac{\pi}{2}\}$ the matrix representation of K is $\mathbf{K} = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}$ which is negative definite, thus K **is not a p.d. kernel**

$\bullet K(x, y) = cos(x - y),\ \mathcal{X} = \mathbb{R}$

$$K(x, y) = cos(x)cos(y) + sin(x)sin(y)$$

Which is the sum of two p.d kernels of the form $K(x, y) = g(x)g(y)$, thus K **is a p.d. kernel**.

$\bullet K(x, y) = \min(x, y),\ \mathcal{X} = \mathbb{R}_+$
K is symmetric and for $n \in \mathbb{N}$,$(x_1, ..x_n) \in \mathbb{R}_+$ and $a \in \mathbb{R}^n$ we have:

$$\begin{aligned} \sum_{1 \le i,j \le n} a_i a_j K(x_i, x_j) &= \sum_{1 \le i,j \le n} a_i a_j \int_0^{+\infty} \mathbb{1}_{[0,x_i]}(t) \mathbb{1}_{[0,x_j]}(t) d\lambda_t \\ &= \int_0^{+\infty} \left(\sum_{1 \le i \le n} a_i \mathbb{1}_{[0,x_i]}(t) \right)^2 d\lambda_t \\ &\ge 0 \end{aligned}$$

Thus K **is a p.d. kernel**.

$\bullet K(x, y) = \max(x, y),\ \mathcal{X} = \mathbb{R}_+$
For the set $\{1, 2\}$ the assoicated matrix is $\mathbf{K} = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$

For $(\alpha, \beta) \in \mathbb{R}^2$,

$$(\alpha, \beta)\mathbf{K} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = (\alpha + \beta)^2 + 2\alpha\beta$$

Which might be negative say for $(\alpha, \beta) = (-1, 1)$. Thus, K **is not a p.d. kernel**

$\bullet K(x, y) = \dfrac{\min(x, y)}{\max(x, y)},\ \mathcal{X} = \mathbb{R}_+^*$

$$K(x, y) = \frac{min(x, y)^2}{xy}$$

The denominator is a p.d kernel by the product rule and the previous result. And $\dfrac{1}{xy}$ is a trivial p.d. kernel of the form $g(x)g(y)$.

Thus the product K **is a p.d. kernel**.

$\bullet K(x, y) = GCD(x, y) = x \wedge y,\ \mathcal{X} = \mathbb{N}$
For $n \in \mathbb{N}$ and a subset $X = \{x_1, ..x_n\}$. we denote with D_{x_i} the set of non-negative x_i divisors and define $D = \cup_{i=1}^n D_{x_i} = \{d_1, ..., d_m\}$.
We consider the $n \times m$ matrix Z such that:

$$\forall i, j \in [\![1, n]\!] \times [\![1, m]\!],\ z_{i,j} = \mathbb{1}_{d_j | x_i} . \sqrt{\varphi(d_j)}$$

Where φ is Euler's totient function.

The (i, j) entry of ZZ^T is

$$\begin{aligned} (ZZ^T)_{i,j} = \sum_{k=1}^m z_{ik} z_{jk} &= \sum_{\substack{d_k | x_i \\ d_k | x_j}} \sqrt{\varphi(d_k)}^2 \\ &= \sum_{d_k | x_i \wedge x_j} \varphi(d_k) = x_i \wedge x_j = K(x_i, x_j) \end{aligned}$$

Consequently $\mathbf{K} = ZZ^T$ is a positive definite matrix and K **is a p.d. kernel**.

$\bullet K(x,y) = LCM(x,y) = x \vee y,\ \mathcal{X} = \mathbb{N}$
For the set $\{1, 2, 15, 42\}$ the LCM matrix is:

$$\mathbf{K} = \begin{pmatrix} 1 & 2 & 15 & 42 \\ 2 & 2 & 30 & 42 \\ 15 & 30 & 15 & 210 \\ 42 & 42 & 210 & 42 \end{pmatrix}$$

$\mathbf{K}$ is singular ($det(\mathbf{K}) = 0$) which means $\mathbf{K}$ is not positive definite. Thus, K **is not a p.d. kernel**.

$\bullet K(x,y) = \dfrac{GCD(x,y)}{LCM(x,y)} = \dfrac{x \wedge y}{x \vee y},\ \mathcal{X} = \mathbb{N}$

$$K(x,y) = \frac{(x_i \wedge x_j)^2}{x_i x_j}$$

The denominator is a p.d kernel (monomial of the gcd kernel) and $\dfrac{1}{xy}$ is a p.d kernel, thus, K **is a p.d. kernel**.

3 Covariance Operators in RKHS

Let us consider the linear kernel on $\mathbb{R} : K(a,b) = ab$. The RKHS generated by K is:

$$\mathcal{H} = span\{K(x,.),\ x \in \mathbb{R}\} = \{K_a \,|K_a(x) = ax,\ a \in \mathbb{R}\}$$

And the unit ball is:

$$\mathcal{B}_K = \{f \in \mathcal{H} \,|\| f\|_{\mathcal{H}} \leq 1\} = \{K_a \,|K_a(x) = ax,\ |a| \leq 1\}$$

In this case, the covariance takes the form:

$$\begin{aligned} C_n^K(X,Y) &= \max_{f,g \in \mathcal{B}_K} \left(\frac{1}{n}\sum_{i=1}^n f(x_i)g(y_i) - (\frac{1}{n}\sum_{i=1}^n f(x_i))(\frac{1}{n}\sum_{i=1}^n g(y_i)) \right) \\ &= \max_{|a|,|b| \leq 1} ab \left(\frac{1}{n}\sum_{i=1}^n x_i y_i - (\frac{1}{n}\sum_{i=1}^n x_i)(\frac{1}{n}\sum_{i=1}^n y_i) \right) \\ &= \max_{|a|,|b| \leq 1} ab.cov_n(X,Y) \\ &= |cov_n(X,Y)| \\ &= |\frac{1}{n} XHY^T| \end{aligned}$$

Where $H = I_n - \mathbb{1}_n$ ($\mathbb{1}_n$ matrix $n \times n$ of ones).

In the general case:

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

We can write:

$$f = \sum_{i=1}^n \alpha_i K_{x_i},\ \alpha \in \mathbb{R}^n$$

And:

$$g = \sum_{i=1}^{n} \beta_i K_{y_i},\ \beta \in \mathbb{R}^n$$

Since any component in the orhtogonal of $span\{K_{x_i},\ i = 1,..n\}$ (resp. $span\{K_{y_i},\ i = 1,..n\}$) vanishes in the inner product above.

Now we have:

$$\|f\|_{\mathcal{H}} = \left\langle \sum_{i=1}^{n} \alpha_i K_{x_i}, \sum_{i=1}^{n} \alpha_i K_{x_i} \right\rangle^{1/2} = \left(\sum_{1 \le i,j \le n} \alpha_i \alpha_j K(x_i, x_j) \right)^{1/2} = (\alpha^T \mathbf{K}_X \alpha)^{1/2}$$

And similarly:

$$\|g\|_{\mathcal{H}} = (\beta^T \mathbf{K}_Y \beta)^{1/2}$$

Where $\mathbf{K}_X$ and $\mathbf{K}_Y$ denote the Gram matrices of the kernel K associated with the sets $\{x_1, ..x_n\}$ and $\{y_1, .., y_n\}$ respectivly.

Consequently,

$$C_n^K(X, Y) = \max_{\alpha, \beta \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} \left(\left([\mathbf{K}_X \alpha]_i - \frac{1}{n} \sum_{j=1}^{n} [\mathbf{K}_X \alpha]_j \right) \left([\mathbf{K}_Y \beta]_i - \frac{1}{n} \sum_{j=1}^{n} [\mathbf{K}_Y \beta]_j \right) \right)$$

$$\text{subject to } \alpha^T \mathbf{K}_X \alpha \le 1 \text{ and } \beta^T \mathbf{K}_Y \beta \le 1$$

Thus:

$$C_n^K(X, Y) = \max_{\alpha, \beta \in \mathbb{R}^n} \frac{1}{n} \alpha^T \mathbf{K}_X H \mathbf{K}_Y \beta$$

$$\text{subject to } \alpha^T \mathbf{K}_X \alpha \le 1 \text{ and } \beta^T \mathbf{K}_Y \beta \le 1$$

Where $H = I_n - \mathbb{1}_n$.

4 Some Basic Learning Bounds

1. Suppose ϕ is a L-Lipshitz function i.e.

$$\forall u, v \in \mathbb{R},\ |\phi(u) - \phi(v)| \le L|u - v|$$

$\forall f, g \in \mathcal{B}_R = \{f \in \mathcal{H}_K,\ \|f\|_{\mathcal{H}_K} \le R\}$:

$$\begin{aligned}
|R_\phi(f, x) - R_\phi(g, x)| &= |\phi(f(x)) - \phi(g(x)) + \lambda(\|f\|^2_{\mathcal{H}_K} - \|g\|^2_{\mathcal{H}_K})| \\
&\le L|(f - g)(x)| + \lambda.(\|f\|_{\mathcal{H}_K} + \|g\|_{\mathcal{H}_K})(\|f\|_{\mathcal{H}_K} - \|g\|_{\mathcal{H}_K}) \\
&\le L|\langle f - g, K_x \rangle| + 2\lambda.R\|f - g\|_{\mathcal{H}_K}. \qquad \text{(Triangular inequality)} \\
&\le (L.\|K_x\|_{\mathcal{H}_K} + 2\lambda.R)\|f - g\|_{\mathcal{H}_K}. \qquad \text{(Cauchy-Schwartz)} \\
&\le (L.K(x, x)^{1/2} + 2\lambda.R)\|f - g\|_{\mathcal{H}_K}. \\
&\le (L.\kappa + 2\lambda.R)\|f - g\|_{\mathcal{H}_K}. \\
&\le C_1\|f - g\|_{\mathcal{H}_K}.
\end{aligned}$$

With $C_1 = L.\kappa + 2\lambda.R$.

2. Suppose ϕ is convex and that $f_x = \arg\min_{f \in \mathcal{H}_K} R_\phi(f, x)$ exists.

R_ϕ is convex (the norm being convex too) w.r to f, we consider the subgradient of $\mathbb{R}_\phi$: $\nabla_f R_\phi(f, x)$

$$\psi(f, x) = R_\phi(f, x) - R_\phi(f_x, x) \ge \partial\phi(f_x(x))(f(x) - f_x(x)) + \lambda(\|f\|^2_{\mathcal{H}} - \|f_x\|^2_{\mathcal{H}})$$

Where $\delta_x : \begin{cases} \mathcal{H}_K \to \mathbb{R} \\ f \mapsto \delta_X(f) = \langle K_x, f \rangle = f(x) \end{cases}$

Using the representor theorem we can write f_x as $f_x = \alpha K_x,\ \alpha \in \mathbb{R}$

We set the subgradient of R_ϕ with respect to f to zero:

$$\alpha \partial \phi(\alpha K(x,x)) K(x,x) + 2\lambda \alpha^2 K(x,x) = 0$$
$$-2\lambda\alpha \in \partial\phi(f_x(x))$$

Which means:

$$\psi(f,x) \geq -2\lambda\alpha(f(x) - f_x(x)) + \lambda(\|f\|_{\mathcal{H}}^2 - \|f_x\|_{\mathcal{H}}^2)$$