MVA "Kernel methods in machine learning" Homework

Julien Mairal and Jean-Philippe Vert

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

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

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

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

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

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 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?

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 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 (**??**) 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 (**??**) can be rewritten as a constrained optimization problem:

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

---

[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

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

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

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

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

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