# MVA "Kernel methods in machine learning" Homework 2

Julien Mairal and Jean-Philippe Vert

Upload your answers (in PDF) to:
`http://goo.gl/cXHXhw`
before February 8th, 2017, 1pm (Paris time).

**Exercice 1. 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 $\alpha_j$.

**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 $(\alpha_i, \alpha_j)$ at a time (while fixing the $n-2$ other variables). What are the update rules for these two variables?

**Exercice 2. 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

$$\mu(P) : y \to \mathbb{E}_{X\sim P}[k(X, y)].$$

**1.** Explain why $\mu(P)$ is in $\mathcal{H}$.

**2.** Show that if $P$ and $Q$ are two Borel probability measures,

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

Hint: Use the relation $\|f\|_{\mathcal{H}} = \sup_{\|g\|_{\mathcal{H}}\leq 1}\langle f, g\rangle_{\mathcal{H}}$ for all $f$ in $\mathcal{H}$.

Remark: 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.

**3. (Bonus)** 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 $\delta_{x_i}$ is a Dirac distribution centered at $x_i$. Show that

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

where $\mathbb{E}_{\mathcal{S}}$ is the expectation by randomizing over the training set (each $x_i$ is a r.v. distributed according to $P$).

Hint: You may use the fact that

$$\mathbb{E}_{\mathcal{S}}\left[\sup_{f\in\mathcal{F}}\left|\mathbb{E}[f(X)] - \frac{1}{n}\sum_{i=1}^{n} f(X_i)\right|\right] \leq 2\mathrm{Rad}_n(\mathcal{F}),$$

where $\mathrm{Rad}_n(\mathcal{F})$ is the Rademacher complexity of the class of functions $\mathcal{F}$.