Kernel Methods in Machine Learning Homework 2

Master Data Science – Université Paris-Saclay

Peter MARTIGNY : peter.martigny@gmail.com
Benoît CHOFFIN : benoit.choffin@ensae.fr

Wednesday, February 8th 2017

1 Dual coordinate ascent algorithms for SVMs

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

$$\min_{f\in H} \frac{1}{n}\sum_{i=1}^{n} \max(0, 1-y_i f(x_i)) + \lambda ||f||_H{}^2$$

and its dual formulation (slide 152) :

$$\max_{\alpha\in\mathbb{R}^n} 2\alpha^T y - \alpha^T K\alpha$$

such that $\forall i, 0 \leq y_i\alpha_i \leq \frac{1}{2\lambda n}$
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 :

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

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 ?

1. Let $\phi : \alpha \mapsto 2\alpha^T y - \alpha^T K\alpha$ the function to be optimized. Let us optimize it w.r.t. α_j. We are looking for α_j such that :

$$0 = \frac{\partial\phi}{\partial\alpha_j} = 2y_j - 2(K\alpha)_j = 2y_j - 2\sum_{i=1}^{n} K(j,i)\alpha_i$$

Hence, we can rewrite this as :

$$y_j = K(j,j)\alpha_j + \sum_{i\neq j} K(j,i)\alpha_i$$

Because K is a p.d. kernel, we can solve this by :

$$\alpha_j = \frac{y_j - \sum_{i \neq j} K(j,i)\alpha_i}{K(j,j)}$$

Moreover, α_j needs to satisfy the constraints : $0 \leq y_j \alpha_j \leq \frac{1}{2\lambda n}$.
From the right comparison, we must have :

$$\alpha_j = \min(\frac{y_j}{2\lambda n}, \frac{y_j - \sum_{i \neq j} K(j,i)\alpha_i}{K(j,j)})$$

Then, from the left comparison, we obtain the update rule for α_j :

$$\alpha_j = \max(0, \min(\frac{y_j}{2\lambda n}, \frac{y_j - \sum_{i \neq j} K(j,i)\alpha_i}{K(j,j)}))$$

2. Now, we optimize w.r.t. both f and b. Because the probem is separable, we can optimize w.r.t. b, and then w.r.t. f. Hence, while the optimization in b is done, we obtained problem which lies in the context of the representer theorem, that we can apply. Hence, from the representer theorem we have :

$$\hat{f}(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x)$$

We wish then to solve the following problem :

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

This is a convex optimization problem. However, the objective function is not smooth. We introduce additional slack variables $\xi_1, ..., \xi_n \in \mathbb{R}$. The problem is hence equivalent to :

$$\min_{\alpha \in \mathbb{R}^n, b \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \xi_i + \lambda \alpha^T K \alpha$$

subject to :

$$\forall i, 0 \leq \xi_i$$
$$\forall i, 1 - y_i([K\alpha]_i + b) \leq \xi_i$$

where we have smoothed the constraints from the hinge loss. The Lagrangian writes (with positive Lagrange multipliers μ and ν) :

$$L(\alpha, b, \xi, \mu, \nu) = \frac{1}{n} \sum_{i=1}^{n} \xi_i + \lambda \alpha^T K \alpha - \sum_{i=1}^{n} \mu_i [y_i([K\alpha]_i + b) + \xi_i - 1] - \sum_{i=1}^{n} \nu_i \xi_i$$

which can be rewritten in matrix form :

$$L(\alpha, b, \xi, \mu, \nu) = \xi^T \frac{\mathbb{1}}{n} + \lambda \alpha^T K \alpha - (\text{diag}(y)\mu)^T K \alpha - b \mu^T y - (\mu + \nu)^T \xi + \mu^T \mathbb{1}$$

L is convex quadratic function in α, so it is minimized when the the gradient is null :

$$\nabla_\alpha L = 2\lambda K \alpha - K \text{diag}(y)\mu = K(2\lambda\alpha - \text{diag}(y)\mu)$$

Hence the value of α that optimizes the Lagrangian is :

$$\alpha^* = \frac{\text{diag}(y)\mu}{2\lambda}$$

Then, optimizing w.r.t. ξ gives :
its minimum is $-\infty$ except when it is constant, ie when :

$$\frac{\mathbb{1}}{n} = \mu + \nu$$

Finally, optimizing w.r.t. b gives the condition :

$$\mu^T y = 0$$

Therefore, the Lagrange dual function writes :

$$q(\mu,\nu) = \inf_{\alpha\in\mathbb{R}^n, \xi\in\mathbb{R}^n, b\in\mathbb{R}} L(\alpha, b, \xi, \mu, \nu) = \begin{cases} \mu^T\mathbb{1} - \frac{1}{4\lambda}\mu^T \text{diag}(y) K \text{diag}(y)\mu & \text{if } \mu+\nu = \frac{\mathbb{1}}{n} \\ -\infty \quad \text{otherwise} \end{cases}$$

and the dual problem is :

$$\text{maximize} \quad q(\mu,\nu)$$

subject to $0 \leq \mu, 0 \leq \nu, \mu^T y = 0$
Moreover, if $\frac{1}{n} \leq \mu_i$ for some i, there is no $0 \leq \nu_i$ such that $\mu_i + \nu_i = \frac{1}{n}$, hence the Lagrange dual function is $-\infty$. Hence, the dual problem becomes :

$$\max_{0 \leq \mu \leq \frac{1}{n}, \mu^T y = 0} \mu^T 1 - \frac{1}{4\lambda}\mu^T \text{diag}(y) K \text{diag}(y) \mu$$

We can rewrite the dual program with α as :

$$\max_{\alpha\in\mathbb{R}^n} 2\alpha^T y - \alpha^T K \alpha$$

subject to :

$$\forall i, 0 \leq y_i \alpha_i \leq \frac{1}{2\lambda n}$$

$$\sum_{i=1}^n \alpha_i = 0$$

Because of the second constraint, we cannot use coordinate ascent for this dual program.

3. A solution to the problem raised in the previous question is the sequential minimal optimization (SMO).

   In this method, we want to optimize the objective function, with a box constraint and, in addition, a linear constraint. Hence, let us take two coefficients α_i and α_j. We first update the first coefficient within the box via the update rule found in question 1 :

$$\alpha_i = \max(0, \min(\frac{y_i}{2\lambda n}, \frac{y_i - \sum_{k\neq i} K(i,k)\alpha_i}{K(i,i)}))$$

   Then, we use the linear constraint to calculate the update for α_j :

$$\alpha_j = -\alpha_i - \sum_{k\neq i,j} \alpha_k$$

   or, in other words :

$$\alpha_j = -\max(0, \min(\frac{y_i}{2\lambda n}, \frac{y_i - \sum_{k\neq i} K(i,k)\alpha_i}{K(i,i)})) - \sum_{k\neq i,j} \alpha_k$$

2 Kernel mean embedding

Let us consider a Borel probability measure P of some random variable X on a compact set $\mathcal{X}$. Let $K : \mathcal{X}^2 \to \mathbb{R}$ be a continuous, bounded, p.d. kernel and $\mathcal{H}$ be its RKHS. The kernel mean embedding of P is defined as the function :

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

1. Explain why $\mu(P)$ is in $\mathcal{H}$.
2. Show that if P and Q are two Borel probability measures, $\mu(P) = \mu(Q)$ implies $\forall f \in \mathcal{H}, \mathbb{E}_{X\sim P}[f(X)] = \mathbb{E}_{X\sim Q}[f(X)]$.
   Hint : Use the relation $\forall f \in \mathcal{H}, ||f||_{\mathcal{H}} = \sup_{||g||_{\mathcal{H}}\leq 1}\langle f, g\rangle_{\mathcal{H}}$.
   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 univeral.
3. (Bonus) Consider the empirical distribution :

$$P_S = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i},$$

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

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

where $\mathbb{E}_S$ 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 :

$$\mathbb{E}_S[\sup_{f\in\mathcal{F}}|\mathbb{E}[f(X)] - \frac{1}{n}\sum_{i=1}^{n}f(X_i)|] \leq 2Rad_n(\mathcal{F})$$

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

1. To prove that $\mu(P)$ is in $\mathcal{H}$, we will use the representer theorem of Riesz on the linear form $f \in \mathcal{H} \to \mathbb{E}_{X\sim P}[f(X)]$. It is assumed that k is bounded ; as a consequence :

$$\begin{aligned}
\forall f \in \mathcal{H}, \qquad |\mathbb{E}_{X\sim P}[f(X)]| &\leq \mathbb{E}_{X\sim P}[|f(X)|] \quad \text{by Jensen's inequality}\\
&= \mathbb{E}_{X\sim P}[|\langle f, k(X, .)\rangle_{\mathcal{H}}|] \quad \text{by the reproducing property of } k\\
&\leq \mathbb{E}_{X\sim P}[||f||_{\mathcal{H}}\ ||k(X, .)||_{\mathcal{H}}] \quad \text{by the Cauchy-Schwartz inequality}\\
&= \mathbb{E}_{X\sim P}[||f||_{\mathcal{H}}\ \sqrt{\langle k(X, .), k(X, .)\rangle_{\mathcal{H}}}]\\
&= \mathbb{E}_{X\sim P}[||f||_{\mathcal{H}}\ \sqrt{k(X, X)}]
\end{aligned}$$

and therefore, $f \in \mathcal{H} \to \mathbb{E}_{X\sim P}[f(X)]$ is a continuous linear mapping. We apply the Riesz representation theorem and hence, for all $f \in \mathcal{H}$, there exists a unique $g_x \in \mathcal{H}$ such that $\mathbb{E}_{X\sim P}[f(X)] = \langle f, g_x\rangle_{\mathcal{H}}$. Let us put now $f = k(x, .)$ with $x \in \mathcal{X}$. We can then use the reproducing property and write :

$$g_x(y) = \mathbb{E}_{X\sim P}[k(X, y)] = \int_{\mathcal{X}} k(X, y)dP(y)$$

Then, $g_x = \int_{\mathcal{X}} k(., y)dP(y) = \mathbb{E}_{X\sim P}[k(X, .)] = \mu(P)$. We conclude that $\mu(P)$ lives in $\mathcal{H}$, since g_x lives in $\mathcal{X}$.

2. Let us suppose that P and Q are two Borel measures. Then, by the result of question 1, we know that $\mu(P)$ and $\mu(Q)$ are both in $\mathcal{H}$. The preceding result has also showed that for all $f \in \mathcal{H}$, $\mathbb{E}_{X\sim P}[f(X)] = \langle f, \mathbb{E}_{X\sim P}[k(X,.)]\rangle_{\mathcal{H}}$. Hence, $\forall f \in \mathcal{H}$:

$$\begin{aligned}\mathbb{E}_{X\sim P}[f(X)] &= \langle f, \mathbb{E}_{X\sim P}(k(X,.))\rangle_{\mathcal{H}} \\ &= \langle f, \mu(P)\rangle_{\mathcal{H}} \\ &= \langle f, \mu(Q)\rangle_{\mathcal{H}} \\ &= \langle f, \mathbb{E}_{X\sim Q}(k(X,.))\rangle_{\mathcal{H}} \\ &= \mathbb{E}_{X\sim Q}[f(X)]\end{aligned}$$

3. We have :

$$\mathbb{E}_S[||\mu(P) - \mu(P_S)||_{\mathcal{H}}] = \mathbb{E}_S[||\mathbb{E}_{X\sim P}[k(X,.)] - \mathbb{E}_{X\sim P_S}[k(X,.)]||_{\mathcal{H}}]$$

Let us derive $\mathbb{E}_{X\sim P_S}[k(X,.)]$:

$$\begin{aligned}\mathbb{E}_{X\sim P_S}[k(X,.)] &= \mathbb{E}_{X\sim \frac{1}{n}\sum_{i=1}^n \delta_{x_i}}[k(X,.)] \\ &= \frac{1}{n}\sum_{i=1}^n k(X_i,.)\end{aligned}$$

Then, we can use the hint given in the indications :

$$\begin{aligned}\mathbb{E}_S[||\mathbb{E}_{X\sim P}[k(X,.)] - \mathbb{E}_{X\sim P_S}[k(X,.)]||_{\mathcal{H}}] &= \mathbb{E}_S[||\mathbb{E}_{X\sim P}[k(X,.)] - \frac{1}{n}\sum_{i=1}^n k(X_i,.)||_{\mathcal{H}}] \\ &= \mathbb{E}_S[\sup_{||g||_{\mathcal{H}}\leq 1} \langle \mathbb{E}_{X\sim P}[k(X,.)] - \frac{1}{n}\sum_{i=1}^n k(X_i,.), g\rangle] \\ &\leq \mathbb{E}_S[\sup_{||g||_{\mathcal{H}}\leq 1} |\langle \mathbb{E}_{X\sim P}[k(X,.)] - \frac{1}{n}\sum_{i=1}^n k(X_i,.), g\rangle|] \\ &= \mathbb{E}_S[\sup_{||g||_{\mathcal{H}}\leq 1} |\langle \mathbb{E}_{X\sim P}[k(X,.)], g\rangle - \langle\frac{1}{n}\sum_{i=1}^n k(X_i,.), g\rangle|] \\ &= \mathbb{E}_S[\sup_{||g||_{\mathcal{H}}\leq 1} |\mathbb{E}_{X\sim P}[\langle k(X,.), g\rangle] - \frac{1}{n}\sum_{i=1}^n \langle k(X_i,.), g\rangle|] \\ &= \mathbb{E}_S[\sup_{||g||_{\mathcal{H}}\leq 1} |\mathbb{E}_{X\sim P}[g(X)] - \frac{1}{n}\sum_{i=1}^n g(X_i)|] \\ &\leq 2Rad_n(\mathcal{F}_\infty)\end{aligned}$$

We know from the lectures that if we consider the ball of radius B in the RKHS as function class ($\mathcal{F}_B = \{f \in \mathcal{H} : ||f||_{\mathcal{H}} \leq B\}$), then we have :

$$Rad_n(\mathcal{F}_B) \leq \frac{2B\sqrt{\mathbb{E}K(X,X)}}{\sqrt{n}}$$

Therefore, since we restrict here to the ball of radius 1, then $B = 1$ and we obtain the desired inequality.