Statistical learning, course No.5

November 19, 2014

1 Summary of previous class

We saw why

$$\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} L(Y_i, f(X_i)) + \lambda\Omega(f)$$

was interesting.

Questions :

- How can we solve this ?
- What is a good regularization for Ω ?

Definition : A kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a *positive definite kernel* if it is symmetric and $\forall\alpha \in \mathbb{R}^n \forall x \in \mathcal{X}^n$

$$\sum_{i,j=1}^{n} \alpha_i\alpha_j K(x_i, x_j) \geq 0$$

Theorem : $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive definite if and only if there is a Hilbert space $\mathcal{H}$ and a mapping $\varphi : \mathcal{X} \to \mathcal{H}$ such that $\forall x, y \in \mathcal{X}$

$$K(x, y) =< \varphi(x), \varphi(y) >$$

Definition : Reproducing Kernel Hilbert Space (RKHS)
Let $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$ be a Hilbert space. $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a *reproducing kernel for* $\mathcal{H}$:

1. $\mathcal{H}$ contains the functions $K_x : t \longmapsto K(x, t)$
2. $\forall f \in \mathcal{H}, \forall\ x\ \in \mathcal{X},\ f(x) = \langle f; K_x\rangle$

Theorem : A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive definite if and only if it is the reproducing kernel of a RKHS.

Theorem : If an Hilbert space has a reproducing kernel, it is unique. Conversely, a function is reproducing for at most one Hilbert space, with an unique mapping φ as above.

Example-exercise : $\mathcal{X} = \mathbb{R}^p$ and $K(x,y) = (x^T y)^2$. Is it a RKHS ?

Solution : $K(x,y) = \text{trace}(xx^T yy^T) = \langle xx^T | yy^T \rangle_F$ *(F stands for "Frobenius") What do we know about the Hilbert space* $\mathcal{H}$ *?*
It must contain the functions $K_x : t \in \mathbb{R}^p \longmapsto \langle xx^T | tt^T \rangle_F$*, and thus, as it is a vector space, must contain* $Span(K_x, x \in \mathbb{R}^p$*, that is, all functions* $t \longmapsto t^T (\sum_{i=1}^n \alpha_i x_i x_i^T) t$*,* $\alpha_i \in \mathbb{R}$*,* $x_i \in \mathcal{X} = \mathbb{R}^p$*.*
Remark: a matrix Z *is symmetric iff it can be written* $Z = U \Delta U^T = \sum_{i=1}^p \Delta_i U_i U_i^T$*.*
Now, let us prove that this we have obtained the sought space :

$$\mathcal{H} := \{ f_Z : t \longmapsto t^T Z t, \forall Z \in M_p(\mathbb{R}) \text{ symmetric} \}$$

with inner product $\langle f_Z | f_{Z'} \rangle = \text{trace}(Z^T Z')$ *It is an Hilbert space (exercise). Now, we need only to check properties 1. and 2. :*

1. $K_x = f_{xx^T}$*, OK*
2. $f_Z(x) = x^T Z x = \text{trace}(x^T Z x) = \text{trace}(Z^T x x^T) = \langle f_Z | f_{xx^T} \rangle = \langle f_Z | K_x \rangle$

✓

2 Kernel methods

2.1 Kernel trick

Use the Aronszajn theorem, i.e. we use the mapping $\varphi : x \in \mathcal{X} \to K_x \in \mathcal{H}$. The method of the *kernel trick* consists to do operations in $\mathcal{H}$, computing kernel evaluations only.

Examples :

- Compute distances in $\mathcal{H}$. $x, x \in \mathcal{X}$, $d_K(x,y) = \|\varphi(x) - \varphi(y)\|^2 = ... = K(x,x) - 2K(x,y) + K(y,y)$
- Compute the barycenter, in $\mathcal{H}$ of points $x_i, i = 1...n$ in $\mathcal{X}$. This barycenter is $\mu = \frac{1}{n} \sum_i \varphi(x_i)$.
- data centering: We would like to consider $\tilde{\varphi} = \varphi - \mu$.

  Thus, we change of kernel : consider $\tilde{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that $\tilde{K}(x_i, x_j) =< \tilde{\varphi}(x_i); \tilde{\varphi}(x_j) >\in \mathbb{R}$. Calculation gives us:

$$\tilde{K}(x,y) = K(x,y) - \frac{1}{n} \sum_{k=1}^n K(x_k, x_j) - \frac{1}{n} \sum_{l=1}^n K(x_i, x_l) + \frac{1}{n^2} \sum_{k,l} K(x_k, k_l)$$

Thus, the new Kernel matrix is $\tilde{K} = (I - G)K(I - G)$ where $G = \frac{1}{n}$? (exercise : find G).

2.2 Representer theorem

Consider a RKHS with reproducing kernel K and some data points $x_i \in \mathcal{X}$.

Let $\psi : \mathbb{R}^{n+1} \to \mathbb{R}$ be strictly increasing with respect to its last parameter. We consider the minimization problem:

$$\min_{f \in \mathcal{H}} \psi(f(x_1), ..., f(x_n), \|f\|) \qquad (**)$$

Theorem : Any solution f to ** has the form $f = \sum_i \alpha_i K_{x_i}$, i.e. is in $Span(K_{x_i})$. (Does not state existence.)

Examples :

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_i L(Y_i, f(X_i)) + \frac{\lambda}{2}\|f\|^2$$

Remark: $\|f\|$ is a good regularization funtion since it penalizes large variations of f with respect to the geometry induced by K :

$|f(x) - f(y)| = |\langle f | K_x - K_y \rangle| \leq \|f\| \|K_x - K_y\|$ (Cauchy-Schwarz inequality)

Proof of theorem : *Define $\mathcal{H}_0 \subseteq \mathcal{H}$ by $\mathcal{H}_0 = Span(K_{x_i})$. It is a subspace of $\mathcal{H}$ of finite dimension.*

*Now, consider a solution f of ** Decompose $f = f_\perp + f_{//}$ with $f_{//}$ the orthogonal projection of f in $\mathcal{H}_0$. Note that $\langle f_\perp | K_x \rangle = 0$ by definitions of $\mathcal{H}_0$ and $f_\perp$. Thus, $f(x_i) = \langle f | K_{x_i} \rangle = \langle f_{//} | K_x \rangle$. Thus ** is equivalent to :*

$$\min \psi(f_{//}(x1), ..., f_{//}(xn), \|f_{//} + f_\perp\|) \qquad (**')$$

and $\|f_\perp + f_{//}\|$ is minimized for $f_\perp = 0$, due to Pythagora theorem. This, combined with the hypothesis on ψ, gives the result. ✓

Consider now $\hat{f} = \sum_i \alpha_i K_{x_i}$ Remark that $\hat{f}(x_j) = \langle \hat{f}; K_{x_j} \rangle = \sum_i \alpha_i K(x_{i,j}) = [K\alpha]_j$

With this, we can express $\|\hat{f}\|$, and this implies that ** can be solved by solving

$$\min_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, [K\alpha]_i) + \lambda \alpha^T K \alpha / 2$$

Examples of **' :

- Kernel ridge regression : $L(u, v) = (u - v)^2/2$. A solution is $\alpha^* = (K + \lambda_n I)^{-1} Y$ (exercise).

- Kernel logistic regression : $L(u,v) = \log(1+e^{-uv})$ This is a convex optimization problem.

- Support vector machines (SVM) $L(u,v) = \max(0, 1-uv)$ The problem is equivalent to

$$\min_{\substack{\xi\in\mathbb{R}^n, \alpha\in\mathbb{R}^n \\ \xi_i\geq 0 \\ \xi_i \geq 1-y_i[K\alpha]_i}} \frac{1}{n}\sum_i \xi_i + \lambda\alpha^T K\alpha/2$$

  This is a quadratic program (QP).

3 Kernel examples

3.1 Kernel for DNA sequences

Question : what is a good kernel for DNA sequences ?

Consider the alphabet $\mathcal{A} = \{A, C, G, T\}$ and the mapping $\varphi : \mathcal{A}^n \to \mathbb{R}^p$ such that for all $x \in \mathcal{A}^n$, $\varphi(x) = [\varphi_u(x)]_{u\in\mathcal{A}^k}$ where $\varphi_u(x)$ is the number of occurrences of u in x.

Definition : The *spectrum kernel* for some $k > 0$ is

$$K(x,y) = \sum_{u\in\mathcal{A}^k} \varphi_u(x)\varphi_u(y)$$

Question : How can we compute $K(x,y)$ efficiently ?

$$K(x,y) = \sum_{i=1}^{|x|-k+1} \sum_{j=1}^{|y|-k+1} 1_{x[i,i+k-1]=y[j,j+k-1]}$$

Complexity : $O(k|x||y|)$ polynomial but not optimal. We can do better with a retrieval tree. This method allows to search for patterns of length k in x in time complexity $O(k|x|)$, thus we obtain a complexity $O(k(|x|+|y|))$.

3.2 Kernel for graphs

Set $G = (V, E)$ a graph, with labels on vertices.

Definition A *walk* is a sequence of joint vertices. Denote $w_n(G)$ the set of walks on G of length n.

We define $\varphi_s(G) = \sum_{w\in w_n(G)} 1_{s \text{ is the label sequence of } w}$, S the set of sequences of labels of length M, and $\varphi(G) = [\varphi_s(G)]_{s\in S}$

Question : How can we compute the *walk kernel* $K(G,H) = \sum_{s \in S} \varphi_s(G)\varphi_s(H)$ **?** We define product graph $G \times H$ with $V(G \times H) = \{(x,y) \in V(G) \times V(H)$ with some labels$\}$ and for $X = (u_1, u_2), Y = (v_1, v_2) \in V(G \times H)$, $X, Y \in E(G \times H)$ if and only if $u_1, u_2 \in E(G)$ and $v_1, v_2 \in E(H)$.

Idea : There is a bijection between the walks in GxH and walks in G and H with some labels.

Let A be the adjacency matrix of $G \times H$. Then, $[A^n]_{i,j}$ is the number of walks of length n starting at node i and ending at j.

Exercise : Prove that with this, the complexity for computing K is $O(n|G||H|d(G)d(H))$, where $|G|$ is the number of vertices (or maybe the edges, or maybe something else) of G and $d(G)$ is the maximum degree in G.