Cours Apprentissage - ENS Math/Info
Méthodes à noyaux

Francis Bach

13 et 20 Novembre 2015

Pour approfondir ce cours, on pourra consulter les documents suivants :
- `http://cbio.ensmp.fr/~jvert/svn/kernelcourse/slides/master/master.pdf`
- `http://www.di.ens.fr/~fbach/rasma_fbach.pdf`

Dans ce cours, l'accent a souvent été mis sur les méthodes de prédiction dites *linéaires* : les données d'entrées sont vectorielles (i.e., $x \in \mathbb{R}^d$) et la fonction de prédiction est linéaire, i.e., $f(x) = w^\top x$ pour $w \in \mathbb{R}^d$. Dans ce cadre, à partir d'observations (x_i, y_i), $i = 1, \dots, n$, le vecteur w est obtenu en minimisant

$$\frac{1}{n} \sum_{i=1}^{n} \ell(y_i, w^\top x_i) + \lambda \Omega(w)$$

(exemple de la régression logistique et moindres carrés).

Ces méthodes sont en apparence limitées, car
- Les données ne sont pas forcément vectorielles.
- Les bonnes fonctions de prédictions ne sont pas forcément linéaires.

Le but des méthodes à noyaux est d'aller au-delà de ces limitations tout en en conservant les bons aspects. Leur principe sous-jacent est de remplacer x par n'importe quelle fonction $\varphi(x) \in \mathbb{R}^d$, *explicitement* ou *implicitement*, et considérer des prédicteurs linéaires en $\Phi(x)$, i.e., $f(x) = w^\top \varphi(x)$. On appelle $\varphi(x)$ le "feature" (ou vecteur de caractéristiques) associée à x.

Exemple : régression polynomiale homogène de degré r, en considérant $x \in \mathbb{R}^d$ et

$$\varphi(x) = \left(x_1^{\alpha_1} \cdots x_d^{\alpha_d} \right)_{\sum_{i=1}^d \alpha_i = r} \quad .$$

Dans ce cas, $p = C_{d+r-1}^r$ (nombre de k-combinaisons avec répétition d'un ensemble de cardinal d), peut être très/trop grand pour qu'une représentation explicite soit faisable.

1 Support Vector Machine

On considère n points x_i dans $\mathbb{R}^d$, et une étiquette $y_i \in \{-1, 1\}$, et le problème d'optimisation suivant

$$\begin{array}{ll} \min\limits_{w\in\mathbb{R}^d,\ b\in\mathbb{R}} & \frac{1}{2}\|w\|^2 + C\sum\limits_{i=1}^{n}\xi_i \\ \text{tel que} & \xi_i \geqslant 0 \\ \text{tel que} & y_i(w^\top x_i + b) \geqslant 1 - \xi_i \end{array}$$

— Interprétation géométrique dans les cas séparables et non séparables.
— Dérivation du dual par dualité Lagrangienne

$$\max_{\alpha\in\mathbb{R}^n} -\frac{1}{2}\alpha^\top D(y)KD(y) + \alpha^\top 1 \text{ tel que } \alpha^\top y = 1,\ 0 \leqslant \alpha \leqslant C,$$

avec valeur optimal du paramètre primal égale à $w = \sum_{i=1}^n \alpha_i y_i x_i$.
— Conditions de KKT ("support vectors") : $(C - \alpha_i)\xi_i = \alpha_i(y_i(w^\top x_i + b) - 1 + \xi_i) = 0$. Ceci implique que si $y_i(w^\top x_i + b) > 1$, alors $\alpha_i = 0$, si $y_i(w^\top x_i + b) < 1$ alors $\alpha_i = C$. Sinon $\alpha_i \in [0, C]$.
— Les données d'emtrées x_i n'interviennent quà travers les produits scalaires $x_i^\top x_j$. C'est la première version du "kernel trick" (astuce du noyau).

2 Théorème du représentant

Théoreme 1 *Théorème du représentant (1971) :*
Soit $\varphi : \mathcal{X} \to \mathbb{R}^d$. Soit $(x_1, \ldots, x_n) \in \mathcal{X}^n$, soit $\Psi : \mathbb{R}^{n+1} \to \mathbb{R}$ strictement croissante par rapport à sa dernière variable,
alors le minimum de $\Psi(w^\top\varphi(x_1), ..., w^\top\varphi(x_n), w^\top w)$ est atteint pour $w = \sum_{i=1}^n \alpha_i\Phi(x_i)$ avec $\alpha \in \mathbb{R}^n$.

Proof soit $w \in \mathbb{R}^d$, soit $\mathcal{F}_D = \{\sum \alpha_i\Phi(x_i)/\alpha \in \mathbb{R}^n\}$,
soit $w_D \in \mathcal{F}_D$ et $w_\perp \in \mathcal{F}_D^\perp$ tel que $w = w_D + w_\perp$,
alors $\forall i$, $w^\top\varphi(x_i) = w_D^\top\varphi(x_i) + w_\perp^\top\varphi(x_i)$ avec $w_\perp^\top\varphi(x_i) = 0$
D'après le théorème de Pythagore, on a : $w^\top w = w_D^\top w_D^2 + w_\perp^\top w_\perp$. Par conséquent, on a :

$$\begin{aligned} \Psi(w^\top\varphi(x_1), ..., w^\top\varphi(x_n), w^\top w) =& \Psi(w_D^\top\varphi(x_1), ..., w_D^\top\varphi(x_n), w_D^\top w_D + w_\perp^\top w_\perp) \\ \geq& \Psi(w_D^\top\varphi(x_1), ..., w_D^\top\varphi(x_n), w_D^\top w_D) \end{aligned}$$

Donc

$$\inf_{w\in\mathbb{R}^d} \Psi(w^\top\varphi(x_1), ..., w^\top\varphi(x_n), w^\top w) = \inf_{w\in\mathcal{F}_D} \Psi(w^\top\varphi(x_1), ..., w^\top\varphi(x_n), w^\top w)$$

■

Corollaire 1 *Pour $\lambda > 0$, $\min_{w\in\mathbb{R}^d} \frac{1}{n}\sum \ell(y_i, w^\top\varphi(x_i)) + \frac{\lambda}{2}w^\top w$ est atteint en $w = \sum_{i=1}^n \alpha_i\varphi(x_i)$.*

— Il est important de remarquer qu'il n'y a aucune hypothèsee sur ℓ (pas de convexité). De plus, on obtient un problème de minimisation (et pas de maximisation). Voir Sections 4 et 5.
— Le résultat est généralisable aux espaces de Hilbert (RKHS).
— On a : $\forall j \in \{1,\dots,n\}$, $w^\top \varphi(x_j) = \sum_{i=1}^n \alpha_i k(x_i,x_j) = (K\alpha)_j$ où K est la matrice de noyau et $w^\top w = \alpha^\top K\alpha$. On peut alors réécrire :

$$\min_{w\in\mathbb{R}^d} \frac{1}{n}\sum \ell(y_i, w^\top\varphi(x_i)) + \frac{\lambda}{2} w^\top w = \min_{\alpha\in\mathbb{R}^n} \frac{1}{n}\sum \ell(y_i,(K\alpha)_i) + \frac{\lambda}{2}\alpha^\top K\alpha$$

L'astuce du noyau permet donc de :
— remplacer $\mathbb{R}^d$ par $\mathbb{R}^n$; intéressant surtout quand d est très grand.
— séparer le problème de représentation (définir un noyau sur un ensemble $\mathcal{X}$) et des problèmes d'algorithmes et d'analyse (qui n'utilisent que la matrice de noyau K).

3 Noyaux

— **Définition** : k est un noyau ssi toutes les matrices de noyau sont semi-définies positives.
Théoreme 2 *Théorème d'Aronszajn (1950) : k est un noyau défini positif si et seulement si il existe un espace de Hilbert $\mathcal{F}$, et $\Phi : \mathcal{X} \to \mathcal{F}$ tel que $\forall x,y$, $k(x,y) = \langle \Phi(x), \Phi(y)\rangle$.*
— Propriétés : la somme et le produit de noyaux sont des noyaux.
— Noyau linéaire : $k(x,y) = x^\top y$
— Noyau polynomial : $k(x,y) = (x^\top y)^r$

$$k(x,y) = (\sum_{i=1}^d x_i y_i)^r = \sum_{\alpha_1+\dots+\alpha_p=r} \binom{r}{\alpha_1,\dots,\alpha_p} \underbrace{(x_1y_1)^{\alpha_1}\dots(x_py_p)^{\alpha_p}}_{(x_1^{\alpha_1}\dots x_p^{\alpha_p})(y_1^{\alpha_1}\dots y_p^{\alpha_p})}$$

$\Phi(x) = \{\binom{r}{\alpha_1,\dots,\alpha_p}^{\frac{1}{2}} x_1^{\alpha_1}\dots x_p^{\alpha_p}\}$
— **Noyaux invariants par translation sur** $[0,1]$. $k(x,y) = q(x-y)$ où t est 1-périodique. k est un noyau ssi la série de Fourier de q est positive (en passant par la représentation complexe), i.e.,

$$k(x,y) = \nu_0 + \sum_{m\geqslant 1} 2\nu_m \cos 2\pi mx \cos 2\pi my + 2\nu_m \sin 2\pi mx \sin 2\pi my$$

avec $\nu \geqslant 0$.
Le vecteur (dimension infinie) de "features" est composé de $\nu_0^{1/2}$, et les $\sqrt{2\nu_m}\cos 2\pi mx$ et $\sqrt{2\nu_m}\sin 2\pi mx$, pour $m \geqslant 1$.
Si $f(x)$ s'écrit $f(x) = \Phi(x)^\top w$, alors

$$\|w\|^2 = \big(\int_0^1 f(x)\big)^2 + \sum_{m\geqslant 1} \frac{2}{\nu_m}\big(\int_0^1 f(x)\cos 2\pi mx\big)^2 + \frac{2}{\nu_m}\big(\int_0^1 f(x)\sin 2\pi mx\big)^2.$$

Pour $\nu_m = \frac{1}{m^{2s}}$, $m \geqslant 1$, cette norme est égale à

$$\|w\|^2 = \big(\int_0^1 f(x)\big)^2 + \frac{1}{(2\pi)^{2s}}\int_0^2 |f^{(s)}(x)|^2 dx$$

et le noyau s'écrit en formule analytique $k(x,y) = \nu_0 + (-1)^{s-1}\frac{(2\pi)^{2s}}{(2s)!}B_{2s}(\{x-y\})$, où B_{2s} est le polynôme de Bernoulli.

— **Noyaux invariants par translation sur** $\mathbb{R}^d$: Noyau invariant par translation : $\mathcal{X} = \mathbb{R}^d$,$k(x,y) = q(x-y)$ avec $q : \mathbb{R}^d \to \mathbb{R}$,
Théoreme 3 *Théorme de Böchner : k est défini positif $\Leftrightarrow$ q est la transformée de Fourier d'une mesure de Borel finie positive $\Leftarrow$ $q \in L^1$ et sa transformée de Fourier est positive.*

Proof (partielle) Soit $x_1, ...x_n \in \mathbb{R}^d$, soit $\alpha_1, .., \alpha_n \in \mathbb{R}$,

$$\begin{aligned}\sum \alpha_s \alpha_j k(x_s, x_j) &= \sum \alpha_s \alpha_j q(x_s - x_j) \\ &= \sum \alpha_s \alpha_j \int \exp^{-i\omega^\top (x_s - x_j)} \, d\mu(\omega) \\ &= \int (\sum \alpha_s \alpha_j \exp^{-i\omega^\top x_s} \overline{\exp^{-i\omega^\top x_j}}) d\mu(\omega) \\ &= \int |\sum \alpha_s \exp^{-i\omega^\top x_s}|^2 d\mu(\omega) \geq 0\end{aligned}$$

Raisonnement intuitif (non-rigoureux) : par ailleurs, si q est dans L^1, alors $\hat{q}(\omega)$ existe et, avec $d\mu(\omega) = \hat{q}(\omega)d\omega$, on a une representation explicite de $k(x,y) = \int \langle \sqrt{\hat{q}(\omega)} \exp^{-i\omega^\top x}, \sqrt{\hat{q}(\omega)} \exp^{-i\omega^\top x} \rangle d\omega = \int \langle \varphi_\omega(x), \varphi_\omega(y) \rangle d\omega = \langle \varphi(x), \varphi(y) \rangle$.
Si on considère $f(x) = \int \varphi_\omega(x) w_\omega d\omega$, alors $w_\omega = \hat{f}(\omega)/\sqrt{\hat{q}(\omega)}$, et la norme au carré de w est égale à $\int \frac{|\hat{f}(w)|^2}{\hat{q}(w)} dw$, where $\hat{f}$ denotes the Fourier transform of f.
Exemple : noyau exponentiel $\exp(-\alpha|x-y|)$ et noyau Gaussien $\exp(-\alpha|x-y|^2)$.

■

— Beaucoup d'applications de l'astuce du noyau !
— Données non vectorielles (séquences, graphes, images)

4 Méthodes à noyaux et dualité convexe

Soit $\Phi \in \mathbb{R}^{n \times d}$, la matrice des "features" (descripteurs), dont les lignes sont les $\varphi(x_i) \in \mathbb{R}^d$, $i = 1, \ldots, n$. On peut alors écrire

$$\frac{1}{n} \sum_{i=1}^{n} \ell(y_i, w^\top \varphi(x_i)) + \frac{\lambda}{2} w^\top w = g(\Phi w) + \frac{\lambda}{2} w^\top w.$$

Par dualité convexe, on a

$$\begin{aligned} & \min_{w \in \mathbb{R}^d} g(\Phi w) + \frac{\lambda}{2} w^\top w \\ = & \min_{w \in \mathbb{R}^d, u \in \mathbb{R}^n} \max_{\alpha \in \mathbb{R}^n} g(u) + \frac{\lambda}{2} w^\top w + \lambda \alpha^\top (u - \Phi w) \\ = & \max_{\alpha \in \mathbb{R}^n} \min_{w \in \mathbb{R}^d, u \in \mathbb{R}^n} g(u) + \frac{\lambda}{2} w^\top w + \lambda \alpha^\top (u - \Phi w) \\ = & \max_{\alpha \in \mathbb{R}^n} -g^*(-\lambda\alpha) - \frac{\lambda}{2} \alpha^\top \Phi \Phi^\top \alpha \end{aligned}$$

avec $w = \Phi^\top \alpha$, où par définition, $-g^*(-\lambda\alpha) = \min_u g(u) + \lambda \alpha^\top u$ est concave en α.

— Les données d'entrée ne sont utilisées qu'à travers la matrice de noyau $K = \Phi\Phi^\top$.
— K peut être plus facile à calculer que Φ (exemple du cas polynomial)

5 Cas des moindres carrés

Nous avons vu désormais deux problèmes d'optimisation :
— **problème dual (D)** : $\max_{\alpha\in\mathbb{R}^n} -g^*(-\lambda\alpha) - \frac{\lambda}{2}\alpha^\top K\alpha$
— **problème primal + représentant (P)** : $\min_{\alpha\in\mathbb{R}^n} g(\alpha) + \frac{\lambda}{2}\alpha^\top K\alpha$

Proposition 1 *Si α est optimal pour **(D)**, alors α est optimal pour **(P)**.*

Cas particulier (moindres carrés)
Soit $g(u) = \frac{1}{2n}\|y-u\|_2^2$. On obtient :

1. **problème dual** : $\max_{\alpha\in\mathbb{R}^n} -\frac{\lambda}{2}\alpha^\top K\alpha - \frac{1}{2n}||y-n\lambda\alpha||_2^2$
2. **problème primal + représentant** : $\min_{\alpha\in\mathbb{R}^n} \frac{1}{2n}||y-K\alpha||_2^2 + \frac{\lambda}{2}\alpha^\top K\alpha$

1. Méthode à noyaux (minimisation par rapport à α :
   gradient 1 /α : $-\lambda K\alpha - \frac{n\lambda}{n}(n\lambda\alpha - y) = 0 \Leftrightarrow (\lambda K + n\lambda^2)\alpha = \lambda y \Leftrightarrow \alpha = (K+n\lambda I)^{-1}y$ unique solution
   gradient 2 /α : $\frac{1}{n}K(K\alpha - y) + \lambda K\alpha = 0 \Leftrightarrow (K^2 + n\lambda K)\alpha = Ky \Leftrightarrow K((K+n\lambda I)\alpha - y) = 0$. Si K est non inversible, la solution n'est pas unique : $\alpha = (K+n\lambda I)^{-1}y + Ker(K)$. Par contre, la prdiction est unique : $K\alpha = K(K+n\lambda I)^{-1}y$.
2. Méthode directe. Minimisons par rapport à w.

   gradient / w : $\frac{1}{n}\Phi^\top(\Phi w - y)$
   ceci donne $w = (\frac{1}{n}\Phi^\top\Phi + \lambda I)^{-1}\frac{1}{n}\Phi^\top y \Leftrightarrow \Phi f = \Phi(\frac{1}{n}\Phi^\top\Phi + \lambda I)^{-1}\frac{1}{n}\Phi^\top y$.

En posant $K = \Phi\Phi^\top$ et en comparant les résultats donnés par les deux méthodes, on obtient l'égalité :

$$\overbrace{\Phi\Phi^\top(\underbrace{\Phi\Phi^\top}_{n\times n} + n\lambda I)^{-1}y}^{\text{noyau}} = \overbrace{\Phi(\underbrace{\Phi^\top\Phi}_{p\times p} + n\lambda I)^{-1}\Phi^\top y}^{\text{directe}}$$

Ce résultat n'est autre que le lemme suivant :

Lemma 1 *: lemme d'inversion de matrices : $\forall A$ matrice, $(AA^\top + I)^{-1}A = A(A^\top A + I)^{-1}$*

On a donc une "équivalence" entre ce lemme et le théorème du représentant.

6 Complexité des opérations d'algèbre linéaire

Si $K \in \mathbb{R}^{n\times n}$ and $L \in \mathbb{R}^{n\times n}$ sont deux matrices
— calculer KL a pour complexité $O(n^3)$

— calculer K^{-1} a pour complexité $O(n^3)$
— calculer Ky a pour complexité $O(n^2)$
— Résoudre $K^{-1}y$ a pour complexité $O(n^3)$
— Décomposition en une base de vecteurs propres $O(n^3)$
— "Plus grand" vecteur propre : $O(n^2)$

Approximation de rang faible
— Base de vecteurs propres (complexité $O(n^2 r)$)
— Projection orthogonales sur r premières colonnes : $O(nr^2)$