Cours Apprentissage - ENS Math/Info
Méthodes à noyaux

Francis Bach

24 mai 2012

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., $x \in \mathbb{R}^p$) et la fonction de prédiction est linéaire, i.e., $f(x) = w^\top x$ pour $w \in \mathbb{R}^p$. 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 $\varphi(x) \in \mathbb{R}^p$, *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 Méthodes à noyaux et dualité convexe

Soit $\Phi \in \mathbb{R}^{n\times p}$, la matrice des "features" (descripteurs), dont les lignes sont les $\varphi(x_i) \in \mathbb{R}^p$, $i = 1,\dots,n$. 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

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

avec $w = \Phi^\top \alpha$.

– 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)

2 Théorème du représentant

Théoreme 1 *Théorème du représentant (1971) :*
Soit $\varphi : \mathcal{X} \to \mathbb{R}^p$. Soit $(x_1, .., x_n) \in \mathcal{X}^n$, 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 $w \in \mathbb{R}^p$, 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

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

■

Corollaire 1 $\min_{w\in\mathbb{R}^p} \frac{1}{n}\sum \ell(y_i, w^\top\varphi(x_i)) + \frac{\lambda}{2} w^\top w$ *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).
– Le résultat est généralisable aux espaces de Hilbert (RKHS).
– On a : $\forall j \in \{1,\ldots,n\}$, $w^\top \varphi(x_j) = \sum_{i=1}^n \alpha_i k(x_i,x_j) = (K\alpha)_j$ où K est la matrice de noyau et $w^\top w = \alpha^\top K \alpha$. On peut alors réécrire :

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

L'astuce du noyau permet donc de :
– remplacer $\mathbb{R}^p$ par $\mathbb{R}^n$
– 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 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, $(A^\top A + I)^{-1}A = A(A^\top A+I)^{-1}$*

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

4 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 $\forall x, y,\ k(x, y) = \langle \Phi(x), \Phi(y) \rangle$.*
- Noyau linéaire : $k(x, y) = x^\top y$
- Noyau polynomial : $k(x, y) = (x^\top y)^r$

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

  $\Phi(x) = \{ \binom{r}{\alpha_1, \ldots, \alpha_p}^{\frac{1}{2}} x_1^{\alpha_1} \ldots x_p^{\alpha_p} \}$
- Noyaux invariants par translation : Noyau invariant par translation : $\mathcal{X} = \mathbb{R}^p$,$k(x, y) = q(x - y)$ avec $q : \mathbb{R}^p \to \mathbb{R}$,
  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 $x_1, \ldots x_n \in \mathbb{R}^p$, soit $\alpha_1, .., \alpha_n \in \mathbb{R}$,

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

■

- Beaucoup d'applications de l'astuce du noyau !
- Données non vectorielles

5 Application à la reconnaissance d'objets

Voir code MATLAB associé.