M2 Probabilités et Statistiques, Université Paris-Sud
Apprentissage statistique (Sylvain Arlot et Francis Bach)
Régularisation ℓ_2

Francis Bach

Cours 4, 22 février 2010

Références : [2, 1], `http://cbio.ensmp.fr/~jvert/teaching/2010mva`

1 Espaces de Hilbert à noyaux reproduisants (RKHS)

- $\mathcal{X}$ ensemble quelconque
- Définition : un espace vectoriel de fonctions de $\mathcal{X}$ dans $\mathbb{R}$ est un RKHS si les formes linéaires $f \mapsto f(x)$ sont continues pour tout $x \in \mathcal{X}$
- "Feature map" : $\Phi(x) : \mathcal{X} \to \mathcal{F}$ tel que $f(x) = \langle \Phi(x), f \rangle$ pour tout $x \in \mathcal{X}$, $f \in \mathcal{F}$
- Noyau défini positif $k(x, y) = \langle \Phi(x), \Phi(y) \rangle$
- Propriétés reproduisantes : pour tout $x, y \in \mathcal{X}$, $f \in \mathcal{F}$, $f(x) = \langle k(\cdot, x), f \rangle$, $k(x, y) = \langle k(\cdot, x), k(\cdot, y) \rangle$
- NB : les formes linéaires sont bien continues $|f(x)| \leqslant k(x, x)^{1/2} \|f\|$
- Théorème d'Aronszajn (1950) : k est un noyau défini positif si et seulement si il existe un espace de Hilbert $\mathcal{F}$, et $\Phi(x) : \mathcal{X} \to \mathcal{F}$ tel que $k(x, y) = \langle \Phi(x), \Phi(y) \rangle$
- Preuve constructive par construction du RKHS associé : complétion des combinaisons linéaires de fonctions $k(\cdot, x)$.
- Exemples de noyaux : linéaire, polynomial, invariant par translation : $k(x, y) = q(x - y)$, défini sur $\mathbb{R}^p$ est défini positif si et seulement si q est la transformée de Fourier d'une mesure de Borel finie positive (Théorème de Bochner).
- Lien avec espace de Sobolev ($k(x, y) = e^{-|x-y|}$).
- Autres propriétés : Noyaux de Mercer, liens avec les splines (noyau $k(x, y) = \max(x, y)$), noyaux sur données non vectorielles.

2 Théorème du représentant

- Soient $x_1, \dots, x_n \in \mathcal{X}$ et $\Psi : \mathbb{R}^{n+1} \to \mathbb{R}$ strictement croissante par rapport à la dernière variable. Alors $\inf_{f \in \mathcal{F}} \Psi(f(x_1), \dots, f(x_n), \|f\|^2)$ est atteint pour f de la forme $f = \sum_{i=1}^n \alpha_i k(x_i, \cdot)$.

– Reparamétrisation : $\Psi(f(x_1), \ldots, f(x_n), \|f\|) = \Psi(K\alpha, \alpha^\top K\alpha)$ où K est la matrice de noyau.
– Modularité entre représentation et algorithmes
– Différence avec les paramètres duaux α obtenus par dualité convexe
– Liens avec les méthodes à noyaux (e.g., kernel PCA)

3 Complexité de Rademacher

– Inégalité classique : $R_\phi(\hat{f}) - R_\phi^* \leqslant 2\sup_{f\in\mathcal{G}} |R_\phi(f) - \widehat{R}_\phi(f)| + \inf_{g\in\mathcal{G}} \left\{ R_\phi(g) - R_\phi^* \right\}$ où $R_\phi(f)$ est le ϕ-risque, $\widehat{R}_\phi(f)$ le ϕ-risque empirique et $\hat{f}$ le minimiseur de $\widehat{R}_\phi(f)$ sur $\mathcal{G} = \{f \in \mathcal{F}, \|f\| \leqslant B\}$ (boule du RKHS)
– Noyaux universels sur un compact de $\mathbb{R}^p$: le RKHS associé est dense dans l'espace des fonctions continues (pour la norme uniforme), et dense dans L^2 (pour la norme L^2). Noyaux invariants par translation $k(x, y) = q(x - y)$: la transformée de Fourier de q a un support de mesure non nulle. Si B tend vers $+\infty$, l'erreur d'approximation $\inf_{g\in\mathcal{G}} \left\{ R_\phi(g) - R_\phi^* \right\}$ tend alors vers zero.
– Complexité de Rademacher de $\mathcal{G}$ inférieure à $\frac{B}{n}(\operatorname{tr} K)^{1/2}$

4 Régression ridge

– Données $(X_i, Y_i) \in \mathcal{X} \times \mathbb{R}$, $i = 1, \ldots, n$.
– Minimisation de $\frac{1}{n}\sum_{i=1}^n (Y_i - f(X_i))^2 + \lambda\|f\|^2$ par rapport à f dans un RKHS.
– Par théorème du représentant, équivalent à minimiser $\frac{1}{n}\|Y - K\alpha\|^2 + \lambda\alpha^\top K\alpha$, avec solution $\alpha = (K + n\lambda I)^{-1}Y$ (défini à un élément du null-space de K près)
– Prédiction $\hat{Y} = K\alpha = K(K + n\lambda I)^{-1}Y$ (estimateur linéaire)
– Liens avec la régression linéaire par moindres carrés régularisé en dimension p : minimisation de $\frac{1}{n}\sum_{i=1}^n (Y_i - w^\top X_i)^2 + \lambda\|w\|^2$, de solution

$$\hat{w} = (\mathbf{X}^\top\mathbf{X} + n\lambda I)^{-1}\mathbf{X}^\top Y = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + n\lambda I)^{-1}Y = \mathbf{X}^\top\alpha$$

où $\mathbf{X} \in \mathbb{R}^{n\times p}$ est la matrice de design.

Références

[1] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification : A survey of some recent advances. *ESAIM : P&S*, 9 :323–375, 2005.

[2] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis.* Cambridge University Press, 2004.