Apprentissage: cours 3a
Méthodes par moyennage local

Guillaume Obozinski

1er mars 2012

Réference : chap. 6 of [Hastie et al., 2009] and chap. 6 of [Devroye et al., 1996].

Algorithmes par moyennage local

On considère la régression au sens des moindres carrés avec des entrées dans $\mathcal{X} = \mathbb{R}^d$ et des sorties réelles bornées : $\mathcal{Y} = [-B, B]$ pour $B > 0$ et $\ell(y, y') = (y - y')^2$. Une fonction cible est donc $f^\star(x) = \mathbb{E}[Y|X = x]$. On considère un ensemble d'entraînement $D_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$.

Principe des méthodes par moyennage local : Prédire par la moyenne pondérée des Y_i pour des X_i voisins de x. On considère les prédicteurs de la forme

$$\widehat{\eta} : x \mapsto \sum_{i=1}^{n} W_i(x)\, Y_i$$

Algorithme par partition : méthode d'histogrammes

On se donne une partition $\{A_1, A_2, \dots\}$ finie ou dénombrable de $\mathcal{X}$. Soit $A(x)$ l'élément de la partition contenant x. On choisit les poids :

$$W_i(x) = \frac{1_{\{X_i \in A(x)\}}}{\sum_{l=1}^{n} 1_{\{X_l \in A(x)\}}},$$

avec la convention $\frac{0}{0} = 0$.

Algorithme des k plus proches voisins (k-p.p.v.)

On suppose que $\mathcal{X} = \mathbb{R}^p$. On définit les k plus proches voisins de x comme un ensemble de k éléments de $\mathcal{X}_n = \{X_1, \dots, X_n\}$ tel que $\forall (X_i, X_j) \in V_k(x) \times \mathcal{X}_n \backslash V_k(x),\ \|X_i - x\| \leqslant \|X_j - x\|$. Ce sont exactement les k-p.p.v. s'il n'y a pas d'ex æquo.

On définit alors les poids :

$$W_i(x) = \frac{1_{\{X_i \in V_k(x)\}}}{k}.$$

Algorithme par noyau : méthode de Nadaraya-Watson

On considère une fonction $K : \mathbb{R}^p \to \mathbb{R}_+$ appelé *noyau de convolution*. Les *noyaux de convolution* on d'abord été utilisé par Parzen et Rosenblatt pour faire de l'estimation de densité (méthode des fenêtres de Parzen). On appelle h la *largeur de bande* du noyau.

Les poids de la méthode sont alors définis comme :

$$W_i(x) = \frac{K(\frac{x - X_i}{h})}{\sum_{l=1}^{n} K(\frac{x - X_l}{h})}$$

Quelques noyaux classiques sur $\mathbb{R}$:

- le noyau gaussien : $t \mapsto \exp(-t^2)$
- le noyau quadratique d'Epanechnikov : $t \mapsto (1-t^2)_+$
- le noyau "tricube" : $t \mapsto (1-t^3)_+^3$

où on a noté la fonction *partie positive* par $(x)_+ = \max(0, x)$. Pour une liste d'autres noyaux on pourra consulter `http://en.wikipedia.org/wiki/Kernel_(statistics)`.

En dimension p on utilisera typiquement des noyaux de la forme $K : x \mapsto K_1(\|x\|)$ pour K_1 l'un des noyaux définis pour la dimension 1.

Exercice 1. A quoi correspond la méthode de Nadaraya-Watson pour un noyau gaussien lorsque $h \to 0$?

Théorème de Stone

Théorème 1. *(Stone) Supposons que les poids W_i et la loi des données d'entraînement satisfont*

(i) $\exists c > 0, \quad \forall f : \mathcal{X} \to \mathbb{R}_+, \quad \forall n \in \mathbb{N}, \quad \mathbb{E}\Big[\sum_{i=1} |W_i(X)| f(X_i)\Big] \leqslant c\,\mathbb{E}[f(X)],$

(ii) $\exists D > 0, \quad \forall n \in \mathbb{N}, \quad \sum_{i=1}^n |W_i(X)| \leqslant D \quad \mathbb{P}\text{-}p.s.,$

(iii) $\forall a > 0, \quad \mathbb{E}\Big[\sum_{i=1}^n |W_i(X)|\, 1_{\{\|X_i - X\| > a\}}\Big] \underset{n\to\infty}{\longrightarrow} 0,$

(iv) $\sum_{i=1}^n W_i(X) \xrightarrow{\mathbb{P}} 1,$

(v) $\mathbb{E}\Big[\sum_{i=1}^n |W_i(X)|^2\Big] \underset{n\to\infty}{\longrightarrow} 0.$

Alors $\hat{f} : x \mapsto \sum_{i=1}^n W_i(x)\, Y_i$ est consistant pour la loi des données d'entraînement.

Démonstration. Pour montrer la consistance on veut montrer que l'espérance de l'excès de risque

$$\mathbb{E}[\mathcal{R}\,(\widehat{\eta})] - \mathcal{R}\,(\eta^\star) = \mathbb{E}[(\widehat{\eta}(X) - \eta^\star(X))^2]$$

tend vers 0. On écrit

$$\widehat{\eta}(X) - \eta^\star(X) = \underbrace{\sum_{i=1}^n W_i(X)(Y_i - \eta^\star(X_i))}_{\alpha_2} + \underbrace{\sum_i W_i(X)(\eta^\star(X_i) - \eta^\star(X))}_{\alpha_3} + \underbrace{\big(-1 + \sum_{i=1}^n W_i\big)\eta^\star(X)}_{\alpha_1}.$$

Nous avons donc $\mathbb{E}[(\widehat{\eta}(X) - \eta^\star(X))^2] = \mathbb{E}[(\alpha_1 + \alpha_2 + \alpha_3)^2] \leqslant 3(\mathbb{E}[\alpha_1^2] + \mathbb{E}[\alpha_2^2] + \mathbb{E}[\alpha_3^2])$. Nous allons montrer que ces trois termes tendent vers 0.

Pour α_1, comme $\eta^\star(X) \in [-B, B]$, (iv) montre que $\alpha_1 \xrightarrow{\mathbb{P}} 0$. Mais on a la borne supérieure suivante

$$\alpha_1^2 = (-1 + \sum_{i=1}^n W_i)^2 \eta^\star(X)^2 \leqslant (D+1)^2 B^2$$

qui donne une hypothèse de domination. D'après le théorème de convergence dominée de Lebesgue $\mathbb{E}[\alpha_1^2] \underset{n\to\infty}{\longrightarrow} 0$.

Pour α_2, en développant on obtient :

$$\mathbb{E}[\alpha_2^2] = \sum_{i=1}^n \mathbb{E}[W_i(X)^2 (Y_i - \eta^\star(X_i))^2] + \sum_{i \neq j} \mathbb{E}[W_i(X) W_j(X)(Y_i - \eta^\star(X_i))(Y_j - \eta^\star(X_j))].$$

Le deuxième terme est nul car $\eta^\star(X_i) = E[Y_i|X_i]$ (exercice).

Pour le premier terme, d'après l'hypothèse (v) :

$$\sum_{i=1}^n \mathbb{E}[W_i(X)^2(Y_i - \eta^\star(X_i))^2] \leqslant 4B^2\,\mathbb{E}\Big[\sum_{i=1}^n W_i(X)^2\Big] \underset{n\to\infty}{\longrightarrow} 0.$$

Finalement pour α_3, nous allons nous ramener à une somme sur i :

$$\begin{aligned}
\mathbb{E}[\alpha_3^2] &= \mathbb{E}\Big[\Big(\sum_{i=1}^n W_i(X)(\eta^\star(X_i) - \eta^\star(X))\Big)^2\Big] \\
&\leqslant \mathbb{E}\Big[\Big(\sum_{i=1}^n \sqrt{|W_i(X)|}\,\sqrt{|W_i(X)|}|\eta^\star(X_i) - \eta^\star(X)|\Big)^2\Big] \\
&\underset{\text{Cauchy-Schwarz}}{\leqslant} \mathbb{E}\Big[\Big(\sum_{i=1}^n |W_i(X)|\Big)\Big(\sum_{i=1}^n |W_i(X)|(\eta^\star(X_i) - \eta^\star(X))^2\Big)\Big] \\
&\leqslant D\,\mathbb{E}\Big[\Big(\sum_{i=1}^n |W_i(X)|(\eta^\star(X_i) - \eta^\star(X))^2\Big)\Big]
\end{aligned}$$

On introduit ensuite une fonction $\tilde\eta$ continue à support compact qui approxime η^* dans $L^2(P_X)$ où P_X est la loi commune des données d'entrée. Ceci est effectivement possible car les fonctions continues à support compact de $\mathbb{R}^d \to \mathbb{R}$ sont denses dans $L^q(P_X)$.

Soit donc pour $\varepsilon > 0$ $\tilde\eta \in L^2(P)$ continue à support compact telle que $\mathbb{E}[(\tilde\eta - \eta^\star)^2] \leqslant \varepsilon$. Comme $\tilde\eta$ est à support compact, elle est également uniformément continue, d'après le théorème de Heine. Donc, il existe $a > 0$, tel que

$$\forall x, x' \in \mathbb{R}^d, \quad (\|x - x'\| \leqslant a) \Rightarrow (|\tilde\eta(x) - \tilde\eta(y)| \leqslant \varepsilon).$$

Ecrivons donc :

$$\eta^\star(X_i) - \eta^\star(x) = \underbrace{\eta^\star(X_i) - \tilde\eta(X_i)}_{\beta_{1i}} + \underbrace{(\tilde\eta(X_i) - \tilde\eta(X))1_{\{\|X_i - X\| \leqslant a\}}}_{\beta_{2i}} + \underbrace{(\tilde\eta(X_i) - \tilde\eta(X))1_{\{\|X_i - X\| > a\}}}_{\beta_{3i}} + \underbrace{\tilde\eta(X) - \eta^\star(X)}_{\beta_4}$$

D'où $\mathbb{E}[\alpha_3^2] \leqslant D\,\mathbb{E}\big[\sum_{i=1}^n |W_i(X)|(\beta_{1i} + \beta_{2i} + \beta_{3i} + \beta_4)^2\big] \leqslant 4D\,\mathbb{E}\big[\sum_{i=1}^n |W_i(X)|(\beta_{1i}^2 + \beta_{2i}^2 + \beta_{3i}^2 + \beta_4^2)\big]$.

– En appliquant (i) à $f = (\eta^\star - \tilde\eta)^2$, on a :

$$\mathbb{E}\big[\sum_{i=1}^n |W_i(X)|\beta_{1i}^2\big] = \mathbb{E}\big[\sum_{i=1}^n W_i(X)(\eta^\star(X_i) - \tilde\eta(X_i))^2\big] \leqslant c\mathbb{E}[(\eta^\star(X) - \tilde\eta(X))^2] \leqslant c\varepsilon,$$

puisque $\tilde\eta$ approche $\eta^\star$ dans $L^2(P)$.

– Ensuite la continuité uniforme de $\tilde\eta$ et (ii) impliquent que $\mathbb{E}\big[\sum_{i=1}^n |W_i(X)|\beta_{2i}^2\big] \leqslant D\varepsilon^2$.

– On a $\mathbb{E}\big[\sum_{i=1}^n |W_i(X)|\beta_{3i}^2\big] \leqslant 4B^2\mathbb{E}\big[\sum_{i=1}^n |W_i(X)|1_{\{\|X_i - X\| > a\}}\big]$.

– Et finalement $\mathbb{E}\big[\sum_{i=1}^n |W_i(X)|\beta_4^2\big] \leqslant D\varepsilon$ puisque $\tilde\eta$ approche $\eta^\star$ dans $L^2(P_X)$.

En réunissant les différent termes on a

$$\mathbb{E}\alpha_3^2 \leqslant D\,(c\varepsilon + D\varepsilon^2 + \mathbb{E}\big[\sum_{i=1}^n |W_i(X)|1_{\{\|X_i - X\| > a\}}\big] + D\varepsilon).$$

Comme par (iii) le troisième terme tend vers zéro quand $n \to \infty$, on a montré que pour tout $\varepsilon > 0$, le terme $\mathbb{E}\alpha_3^2$ est plus petit qu'une constante fois ε. On a donc montré $\mathbb{E}\alpha_3^2 \underset{n\to\infty}{\longrightarrow} 0$. Le théorème de Stone est prouvé. □

Le théorème de Stone s'applique aux plus proches voisins, aux méthodes d'histogrammes et au prédicteurs de Nadaraya-Watson. Nous allons montrer comment il s'applique dans le cas des plus proches voisins.

Théorème 2. *Soit P_X une loi commune de $X, X_1, \ldots, X_n$ i.i.d. telle que presque surement les distances $\|X - X_i\|$ soient toutes distinctes (pas d'ex æquos), alors l'algorithme des k-p.p.v.est consistent par rapport à P de marginale P_X sur X si $k \underset{n\to\infty}{\longrightarrow} \infty$ et $\frac{k}{n} \underset{n\to\infty}{\longrightarrow} 0$.*

Démonstration. Théorème démontré en cours. □

pour cela nous aurons besoin du lemme technique :

Lemme 1. *Soit ν une probabilité sur $\mathbb{R}^d$ et l'événement $\big\{x \in \mathbb{R}^d \mid \nu\big(B(x, \|x - x'\|)\big) \leqslant a\big\}$.*
Il existe $\gamma > 0$ tel que pour tout $a > 0$ et tout $x' \in \mathbb{R}^d$,

$$\nu\Big(\Big\{x \in \mathbb{R}^d \mid \nu\Big(B(x, \|x - x'\|)\Big) \leqslant a\Big\}\Big) \leqslant \gamma a.$$

Démonstration. Lemme admis (voir chap. 5.3 [Devroye et al., 1996]) □

Minimisation du "risque empirique local"

La régression linéaire locale est un hybride entre les méthodes de moyennage local et les méthodes de minimisation du risque empirique. Il s'agit de faire de la minimisation du "risque local". Soit des poids $W_i(x)$ dont on suppose qu'ils somment à 1.

On considère le "risque empirique local" (REL) :

$$\widehat{\mathcal{R}}_{n,W(x)}(f) = \sum_{i=1}^{n} W_i(x)\, \ell(f(x_i), y_i)$$

Soit $\widehat{\eta}_x = \operatorname{argmin}_{f\in S} \widehat{\mathcal{R}}_{n,W(x)}(f)$, alors le prédicteur minimisant le REL est $\widehat{\eta} : x \mapsto \widehat{\eta}_x(x)$. Notons la double dépendance en x dans $\widehat{\eta}_x(x)$. On retrouve le risque empirique classique lorsque l'on prends des poids $W_i(x) = \frac{1}{n}$.

Exercice 2. Si on restreint S à l'ensemble des fonctions constantes de la forme $f_x : z \mapsto \mu(x)$, montrer que pour la perte quadratique on retrouve les prédicteurs de Nadaraya-Watson.

Régression linéaire locale

On considère le cas de la perte quadratique et des prédicteurs linéaires de la forme $f_{\boldsymbol{\theta}} : \mathbf{x} \mapsto \boldsymbol{\theta}^\top \mathbf{x}$. On note $\mathbf{z}$ une nouvelle entrée, $\mathbf{w}(\mathbf{z})$ son vecteur de poids et $\mathbf{W}(\mathbf{z}) = \mathrm{Diag}(\mathbf{w}(\mathbf{z}))$. On note $(\mathbf{x}_i, y_i)$ les paires de données d'entraînement.

$$\widehat{\mathcal{R}}_{n,\mathbf{w}(\mathbf{z})}(f_{\boldsymbol{\theta}}) = \sum_{i=1}^{n} w_i(\mathbf{z})(y_i - \boldsymbol{\theta}^\top \mathbf{x}_i)^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^\top \mathbf{W}(\mathbf{z})(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})$$

Le problème de minimisation du "risque local" est donc un problème de minimisation des *moindres carrés pondérés*.

Si la matrice $\mathbf{X}^\top \mathbf{W}(\mathbf{z})\mathbf{X}$ est inversible on a

$$\widehat{\eta}^{\mathrm{RLL}}(\mathbf{z}) = (\mathbf{X}^\top \mathbf{W}(\mathbf{z})\mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W}(\mathbf{z})\mathbf{y}$$

Exercices pour le 8/3/2012

Exercice 3. (Fléau de la dimension pour l'algorithme du plus proche voisin) Soit X de loi uniforme sur l'hypercube $[-1,1]^p$ et $Y = f^*(X)$ pour $f = \exp(-\|x\|^2)$. On considère la méthode k$-p.p.v.$ avec un seul voisin et on considère l'excès de risque en $x = 0$. Quelle est l'espérance de $\|X\|^2$? Utiliser un argument de concentration et un borne d'union pour montrer qu'à moins que le nombre n de données d'entraînement soit exponentiel en p le biais de la méthode en $x = 0$ tend rapidement vers 1.

Exercice 4. L'estimateur de densité par *fenêtres de Parzen* est un estimateur classique de la densité d'un variable aléatoire à valeurs dans $\mathbb{R}^d$ qui est construit, comme les estimateurs de Nadarya-Watson avec des noyaux. Plus précisément, étant donné un échantillon $X_1, \ldots, X_n$ i.i.d. de loi P, un noyau K normalisé de telle sorte que $\int K(x)dx = 1$ et une largeur de bande h, l'estimateur de la densité de Parzen est défini par :

$$\hat{p}(x) = \frac{1}{h^d} \sum_{i=1}^{n} K\Big(\frac{x - x_i}{h}\Big).$$

On considère le noyau gaussien (normalisé) sur $\mathbb{R}^d$ défini par $K(x) = (2\pi)^{-d/2} \exp(-\frac{1}{2}\|x\|_2^2)$.

Soit X une variable d'entrée à valeurs dans $\mathcal{X} = \mathbb{R}^d$ et Y une variable de sortie à valeur dans $\mathcal{Y} = \mathbb{R}$. On suppose que l'on dispose d'un ensemble d'entraînement $(X_1, Y_1), \ldots, (X_n, Y_n)$. Soit $\hat{p}$ l'estimateur de Parzen de la densité jointe de (X, Y) dans $\mathbb{R}^{d+1}$ pour le noyau gaussien. Montrer que l'espérance conditionnelle $\mathbb{E}_{\hat{p}}[Y \mid X]$ sous la loi jointe de densité $\hat{p}$ ainsi estimée est un prédicteur de Nadarya-Watson. (on pourra utiliser que $\int xK(x - \mu)dx = \mu$) Etendez ce résultat à la classification binaire en proposant un noyau sur $\mathcal{X} \times \mathcal{Y}$ dans le cas où $\mathcal{X} = \mathbb{R}^d$ et $\mathcal{Y} = \{0, 1\}$, de façon à retrouver le prédicteur de Nadarya-Watson pour la classification.

Références

[Devroye et al., 1996] Devroye, L., Györfi, L., and Lugosi, G. (1996). *A probabilistic theory of pattern recognition*, volume 31. Springer Verlag.

[Hastie et al., 2009] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning*. Springer.