Apprentissage: cours 2
Méthodes par moyennage local - Consistance des méthodes par partition

Guillaume Obozinski

21 février 2013

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.)

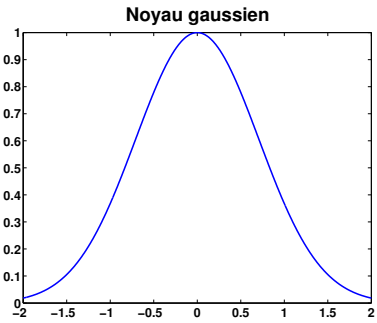
On suppose que $\mathcal{X} = \mathbb{R}^p$. On définit les k plus proches voisins de x comme un ensemble de k éléments de $\mathcal{X}_n = \{X_1, \dots, X_n\}$ tel que $\forall (X_i, X_j) \in V_k(x) \times \mathcal{X}_n \backslash V_k(x),\ \|X_i - x\| \leqslant \|X_j - x\|$. Ce sont exactement les k-p.p.v.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

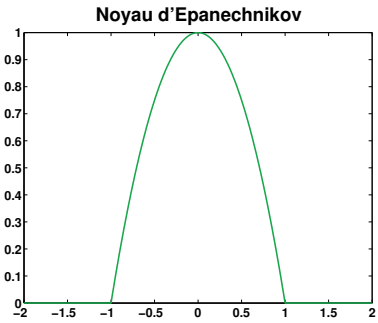
On définit alors les poids :

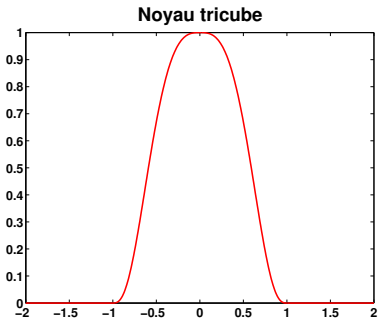
$$W_i(x) = \frac{1_{\{X_i \in V_k(x)\}}}{k}.$$

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* ont d'abord été utilisés 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 $h \to 0$?

1 Algorithmes par partition

Définition 1 (Algorithme par partition)**.** Soit $A = \{A_1, \ldots, A_k, \ldots\}$ une partition finie ou dénombrable de $\mathcal{X}$ et $D_n = (X_i, Y_i)_{1 \leqslant i \leqslant n}$ un échantillon. Pour tout $x \in \mathcal{X}$, on note $A(x)$ l'élément de la partition qui contient x et $N_A(x) := \text{card}\{i \in \{1, \ldots, n\} \text{ t.q. } X_i \in A(x)\}$. L'algorithme d'apprentissage par partition associé est défini comme suit.

- en régression :

$$\forall x \in \mathcal{X}, \quad \widehat{\eta}_A(x; D_n) := \frac{1}{N_A(x)} \sum_{i=1}^{n} 1_{\{X_i \in A(x)\}} Y_i \ . \tag{1}$$

- en classification :

$$\forall x \in \mathcal{X}, \quad \widehat{f}_A(x; D_n) := 1_{\{\widehat{\eta}(x; D_n; A) \geqslant \frac{1}{2}\}} \ . \tag{2}$$

Exemple classique dans $[0,1]^d$ ou $\mathbb{R}^d$: partition régulière de pas $h > 0$.

1.1 Rappel : lien entre régression et classification pour la méthode plug-in

L'algorithme de classification par partition est un algorithme de type "plug-in" : on apprend d'abord la fonction de régression par $\widehat{\eta}$, puis on considère le prédicteur pour la classification donné par $\widehat{f} = 1_{\{\widehat{\eta} \geqslant 1/2\}}$. Comme vu lors du premier cours, on a alors un lien entre le risque quadratique de $\widehat{\eta}$ en régression et le risque 0–1 de $\widehat{f}$ en classification :

$$\mathcal{R}^{0-1}\left(\widehat{f}(D_n)\right) - \mathcal{R}^{0-1}(f^\star) = \mathbb{E}\left[|2\eta^*(X) - 1| \, 1_{\{\widehat{f}(X) \neq f^\star(X)\}} \,\middle|\, D_n\right]$$

$$
\begin{aligned}
&\leqslant 2\mathbb{E}\left[\,|\widehat{\eta}(X)-\eta^*(X)|\;|\;D_n\,\right]\\
&\leqslant 2\sqrt{\mathbb{E}\left[\,\left(\widehat{\eta}(X)-\eta^*(X)\right)^2\;\middle|\;D_n\right]}\\
&= 2\sqrt{\mathcal{R}^{reg}\left(\widehat{\eta}(D_n)\right)-\mathcal{R}^{reg}\left(\eta^*\right)}\;.
\end{aligned}
$$

En intégrant par rapport à D_n, on obtient

$$
\begin{aligned}
\mathbb{E}\left[\mathcal{R}^{0-1}\left(\widehat{f}(D_n)\right)-\mathcal{R}^{0-1}\left(f^\star\right)\right] &\leqslant 2\mathbb{E}\left[\,|\widehat{\eta}(X)-\eta^*(X)|\,\right]\\
&\leqslant 2\sqrt{\mathbb{E}\left[\,\mathcal{R}^{reg}\left(\widehat{\eta}(D_n)\right)-\mathcal{R}^{reg}\left(\eta^*\right)\right]}
\end{aligned}
\tag{3}
$$

grâce à l'inégalité de Jensen et la concavité de $\sqrt{\cdot}$. En particulier, si $\widehat{\eta}$ est un estimateur consistant de η^* au sens du risque des moindre carrés, alors l'estimateur plug-in associé est consistant au sens du risque 0–1.

1.2 Condition suffisante de consistance en classification pour les estimateur par partition

Théorème 1. *On se place en classification binaire ($\mathcal{Y}=\{0,1\}$) et l'on note $\mathcal{R}^{0-1}(\cdot)$ le risque associé à la perte 0–1. Soit $A_n=\{A_{1,n},\ldots,A_{k,n},\ldots\}_{n\in\mathbb{N}}$ une suite de partitions finies ou dénombrables de $\mathcal{X}=\mathbb{R}^d$. Soit $\widehat{f}_{A_n}$ l'algorithme par partition associé, donné par la Définition 1. Pour tout $E\subset\mathbb{R}^d$, on définit son diamètre (éventuellement infini) $\operatorname{diam}(E)=\sup_{x,y\in E}\{\|x-y\|\}$ où $\|\cdot\|$ désigne la norme euclidienne dans $\mathbb{R}^d$. Si*

1. $\operatorname{diam}(A_n(X))\to 0$ *en probabilité, et*
2. $N_{A_n}(X)\to+\infty$ *en probabilité,*

alors, l' algorithme de classification $(\widehat{f}_{A_n})_{n\in\mathbb{N}}$ est consistant :

$$
\mathbb{E}\left[\mathcal{R}^{0-1}\left(\widehat{f}_{A_n}(D_n)\right)\right]-\mathcal{R}^{0-1}\left(f^\star\right)\xrightarrow[n\to+\infty]{}0\;. \tag{4}
$$

La démonstration du Théorème 1 repose notamment sur le lemme suivant.

Lemme 1. *Soit $N\geqslant 1$, $p\in[0,1]$ et Z une variable de loi binomiale de paramètres N et p. Alors,*

$$
\mathbb{E}\left|\frac{Z-Np}{N}\right|\leqslant\frac{1}{2\sqrt{N}}\;. \tag{5}
$$

Démonstration du Théorème 1. Notons $\widehat{\eta}_n=\widehat{\eta}_{A_n}$. Au vu de (3), il suffit de montrer que

$$
\mathbb{E}\,|\widehat{\eta}_n(X;D_n)-\eta^*(X)|\xrightarrow[n\to+\infty]{}0\;.
$$

Introduisons la fonction $\overline{\eta}_n:\mathcal{X}\mapsto\mathbb{R}$ définie par

$$
\forall x\in\mathcal{X}\,,\quad \overline{\eta}_n(x)=\mathbb{E}\left[\,\eta^*(X)\,|\;X\in A_n(x)\right]\;.
$$

Alors, d'après l'inégalité triangulaire,

$$
\mathbb{E}\,|\widehat{\eta}_n(X;D_n)-\eta^*(X)|\leqslant\mathbb{E}\,|\widehat{\eta}_n(X;D_n)-\overline{\eta}_n(X)|+\mathbb{E}\,|\overline{\eta}_n(X)-\eta^*(X)|\;,
$$

et il suffit de montrer que chacun de ces deux termes tend vers zéro.

Contrôle du premier terme $\mathbb{E}\,|\widehat{\eta}_n(X;D_n)-\overline{\eta}_n(X)|$ Pour tout $x\in\mathcal{X}$, conditionnellement à $N_{A_n}(x)$,

$$N_{A_n}(x)\widehat{\eta}_n(x;D_n)=\sum_{i\,/\,X_i\in A(x)}Y_i$$

est la somme de $N_{A_n}(x)$ variables de Bernoulli de paramètre $\mathbb{P}\,(\,Y_i=1\,|\;X_i\in A(x)\,)=\overline{\eta}_n(x)$, donc c'est une variable binomiale de paramètres $(N_{A_n}(x),\overline{\eta}_n(x))$. Ainsi,

$$\begin{aligned}
&\mathbb{E}\,[\,|\widehat{\eta}_n(X;D_n)-\overline{\eta}_n(X)|\,|\;X,N_{A_n}(X)\,]\\
&\leqslant 1_{\{N_{A_n}(X)=0\}}+\frac{1_{\{N_{A_n}(X)>0\}}}{N_{A_n}(X)}\mathbb{E}\,[\,|N_{A_n}(X)\widehat{\eta}_n(X;D_n)-N_{A_n}(X)\overline{\eta}_n(X)|\,|\;X,N_{A_n}(X)\,]\\
&\leqslant 1_{\{N_{A_n}(X)=0\}}+\frac{1_{\{N_{A_n}(X)>0\}}}{2\sqrt{N_{A_n}(X)}}\qquad\text{d'après le Lemme 1.}
\end{aligned}$$

En intégrant, on en déduit

$$\begin{aligned}
\mathbb{E}\,|\widehat{\eta}_n(X;D_n)-\overline{\eta}_n(X)|&\leqslant\mathbb{P}\,(\,N_{A_n}(X)=0\,)+\mathbb{E}\left[\frac{1_{\{N_{A_n}(X)>0\}}}{2\sqrt{N_{A_n}(X)}}\right]\\
&\leqslant\mathbb{P}\,(\,N_{A_n}(X)=0\,)+\frac{\mathbb{P}\,(\,N_{A_n}(X)\leqslant k\,)}{2}+\frac{1}{2\sqrt{k}}
\end{aligned}$$

pour tout $k>0$. Comme $N_{A_n}(X)\to+\infty$ en probabilité, ceci implique

$$\limsup_{n\to+\infty}\mathbb{E}\,|\widehat{\eta}_n(X;D_n)-\overline{\eta}_n(X)|\leqslant\frac{1}{2\sqrt{k}}$$

pour tout $k>0$, et donc

$$\lim_{n\to+\infty}\mathbb{E}\,|\widehat{\eta}_n(X;D_n)-\overline{\eta}_n(X)|=0\ .$$

Contrôle du deuxième terme $\mathbb{E}\,|\overline{\eta}_n(X)-\eta^*(X)|$ Soit $\varepsilon>0$ quelconque. Comme l'ensemble des fonctions continues à support compact est dense dans $L_1(P_X)$ (où P_X désigne la loi commune des X_i), on peut trouver une fonction g continue à support compact (donc uniformément continue) telle que $\mathbb{E}\,|\eta^*(X)-g(X)|\leqslant\varepsilon$. Quitte à remplacer g par $\min\{\,1,\max\{\,0,g\,\}\,\}$, on peut supposer que g est à valeurs dans $[0,1]$. On définit alors $\overline{g}_n:\mathcal{X}\mapsto\mathbb{R}$ par

$$\forall x\in\mathcal{X}\,,\quad\overline{g}_n(x)=\mathbb{E}\,[\,g(X)\,|\;X\in A_n(x)\,]$$

et l'inégalité triangulaire donne

$$\mathbb{E}\,|\overline{\eta}_n(X)-\eta^*(X)|\leqslant\mathbb{E}\,|\overline{\eta}_n(X)-\overline{g}_n(X)|+\mathbb{E}\,|\overline{g}_n(X)-g(X)|+\mathbb{E}\,|g(X)-\eta^*(X)|\ .$$

Le troisième terme est majoré par ε par définition de g. Le premier terme est inférieur au troisième (donc à ε) car

$$\begin{aligned}
\mathbb{E}\,|\overline{\eta}_n(X)-\overline{g}_n(X)|&=\mathbb{E}\,|\mathbb{E}\,[\,\overline{\eta}_n(X)-\overline{g}_n(X)\,|\;A_n(X)\,]|\quad\text{car }\overline{\eta}_n(X)-\overline{g}_n(X)\text{ ne dépend que de }A_n(X)\\
&=\mathbb{E}\,|\mathbb{E}\,[\,\eta^*(X)-g(X)\,|\;A_n(X)\,]|\\
&\leqslant\mathbb{E}\,[\,\mathbb{E}\,[\,|\eta^*(X)-g(X)|\,|\;A_n(X)\,]\,]\\
&=\mathbb{E}\,|\eta^*(X)-g(X)|\ .
\end{aligned}$$

Enfin, pour le deuxième terme, remarquons que g étant uniformément continue, il existe $\theta>0$ tel que g varie d'au plus ε sur tout ensemble de diamètre inférieur à θ. Comme $g(X)\in[0,1]$ p.s., on en déduit que

$$\mathbb{E}\,|\overline{g}_n(X)-g(X)|\leqslant\varepsilon+\mathbb{P}\,(\,\mathrm{diam}(A_n(X))>\theta\,)\leqslant 2\varepsilon$$

pour n assez grand, puisque $\mathrm{diam}(A_n(X))\to 0$ en probabilité. Comme $\varepsilon>0$ peut être choisi arbitrairement petit, on a bien montré que $\mathbb{E}\,|\overline{\eta}_n(X)-\eta^*(X)|\to 0$ quand $n\to+\infty$, d'où le résultat. □

1.3 Consistance universelle pour les partitions régulières en classification

Pour l'instant, le Théorème 1 nous donne une condition suffisante de consistance, mais sans préciser si cette condition peut être satisfaite sans connaître a priori la loi P des données. L'exemple suivant montre qu'un bon choix de partition de $\mathbb{R}^d$ permet d'avoir la consistance universelle.

Définition 2. Pour tout $h > 0$, on définit $A^r(h)$ la *partition régulière de pas* $h > 0$ de $\mathbb{R}^d$ en cubes de taille h définis par la grille régulière $h\mathbb{Z}^d$. Autrement dit,

$$A^r(h) = \left\{ \prod_{i=1}^{d} [k_i h \, , \, (k_i + 1)h[\, / \, k_1, \ldots, k_d \in \mathbb{Z} \right\} \ .$$

Théorème 2. *Soit $(h_n)_{n\in\mathbb{N}}$ une suite de réels strictement positifs, $(A^r(h_n))_{n\in\mathbb{N}}$ la suite de partitions régulières de $\mathbb{R}^d$ associée (donnée par la Définition 2) et $\left(\widehat{f}_{A^r(h_n)}\right)_{n\in\mathbb{N}}$ le classifieur par partition associé. Alors, si*

$$h_n \to 0 \qquad \text{et} \qquad n h_n^d \to +\infty$$

quand $n \to +\infty$, l'algorithme $\left(\widehat{f}_{A^r(h_n)}\right)_{n\in\mathbb{N}}$ est universellement consistant : pour toute loi P des données,

$$\mathbb{E}\left[\mathcal{R}^{0-1}\left(\widehat{f}_{A^r(h_n)}(D_n) \right) \right] - \mathcal{R}^{0-1}\left(f^\star \right) \xrightarrow[n\to+\infty]{} 0 \ . \tag{6}$$

La démonstration du Théorème 2 repose notamment sur le lemme suivant.

Lemme 2. *Soit $N \geqslant 1$, $p \in]0, 1]$ et Z une variable de loi binomiale de paramètres N et p. Alors,*

$$\mathbb{P}\left(Z \leqslant \frac{Np}{2} \right) \leqslant \frac{4}{Np} \ . \tag{7}$$

Démonstration du Théorème 2. Il suffit de vérifier que les deux hypothèses du Théorème 1 sont satisfaites quelle que soit la loi P. Comme $\mathrm{diam}(A^r(h_n)) = \sqrt{d} h_n^d$, la condition sur le diamètre est satisfaite dès lors que $h_n \to 0$. Pour la deuxième condition, fixons $M \in]0, +\infty[$ et majorons $\mathbb{P}\left(N_{A^r(h_n)}(X) \geqslant M \right)$.

Soit S une boule centrée à l'origine. Elle a une intersection non-vide avec au plus $c_1 + c_2 h_n^{-d}$ cellules de la forme $\prod_{i=1}^d [k_i h_n \, , \, (k_i + 1)h_n[$ avec $k_1, \ldots, k_d \in \mathbb{Z}$, où $c_1, c_2 > 0$ sont des constantes. On a alors

$$\mathbb{P}\left(N_{A^r(h_n)}(X) \leqslant M \right) \leqslant \sum_{B \in A^r(h_n) \, / \, B \cap S \neq \emptyset} \mathbb{P}\left(X \in B \, , \, N_B \leqslant M \right) + \mathbb{P}\left(X \notin S \right) \ .$$

Considérons une cellule $B \in A^r(h_n)$ quelconque telle que $\mathbb{P}\left(X \in B \right) > 0$ et majorons $\mathbb{P}\left(X \in B \, , \, N_B \leqslant M \right)$. Deux cas sont possibles :

(i) Soit $\mathbb{P}\left(X \in B \right) \leqslant \frac{2M}{n}$, et alors

$$\mathbb{P}\left(X \in B \, , \, N_B \leqslant M \right) \leqslant \frac{2M}{n} \ .$$

(ii) Soit $\mathbb{P}\left(X \in B \right) \geqslant \frac{2M}{n}$, et alors on va appliquer le Lemme 2, en remarquant que sachant $X \in B$, $Z = N_B$ suit une loi binomiale de paramètres n et $\mathbb{P}\left(X \in B \right)$. Ainsi,

$$\begin{aligned} \mathbb{P}\left(X \in B \, , \, N_B \leqslant M \right) &= \mathbb{P}\left(X \in B \right) \mathbb{P}\left(N_B \leqslant M \mid \ X \in B \right) \\ &\leqslant \mathbb{P}\left(X \in B \right) \mathbb{P}\left(N_B \leqslant \frac{n\mathbb{P}\left(X \in B \right)}{2} \,\middle|\, X \in B \right) \\ &\leqslant \mathbb{P}\left(X \in B \right) \frac{4}{n\mathbb{P}\left(X \in B \right)} = \frac{4}{n} \ . \end{aligned}$$

Au final, on a donc

$$\begin{aligned}\mathbb{P}\left(N_{A^r(h_n)}(X) \geqslant M\right) &\leqslant \operatorname{card}\{\, B \in A^r(h_n) \,/\, B \cap S \neq \emptyset \,\} \, \frac{\max\{4, 2M\}}{n} + \mathbb{P}(X \notin S) \\ &\leqslant \left(c_1 + c_2 h_n^{-d}\right) \frac{\max\{4, 2M\}}{n} + \mathbb{P}(X \notin S)\end{aligned}$$

et lorsque n tend vers l'infini, par l'hypothèse $nh_n^d \to +\infty$, ce majorant tend vers $\mathbb{P}(X \notin S)$ qui peut être rendue arbitrairement petite. Ainsi, pour tout $M > 0$, $\mathbb{P}\left(N_{A^r(h_n)}(X) \geqslant M\right) \to 0$ lorsque $n \to +\infty$, c'est-à-dire, $N_{A^r(h_n)}(X) \to +\infty$ en probabilité. □

1.4 Partition et minimisation du risque empirique

Il est intéressant de noter que les prédicteur par partition peuvent aussi s'interpréter comme des prédicteurs minimisant le risque empirique (définies au premier cours).

Proposition 3. *Soit A une partition finie ou dénombrable de $\mathcal{X}$. En régression avec le risque quadratique, le prédicteur par partition associée à A minimise le risque empirique sur l'ensemble de prédicteurs*

$$S_A^r = \{\, f : \mathcal{X} \mapsto \mathbb{R} \text{ mesurable } / \forall k \geqslant 1\,,\ f \text{ est constante sur } A_k \,\}$$

des fonctions constantes sur chaque élément de la partition.

De même, en classification binaire avec le risque 0–1, le prédicteur par partition minimise le risque empirique sur l'ensemble de prédicteurs

$$S_A^c = \{\, f : \mathcal{X} \mapsto \{0, 1\} \text{ mesurable } / \forall k \geqslant 1\,,\ f \text{ est constante sur } A_k \,\} \quad .$$

Théorème de Stone

Le théorème de Stone que nous ne démontrerons pas généralise l'argument que nous avons vu pour les algorithme par partition à un ensemble assez général de méthodes par moyennage local. Il s'applique aux plus proches voisins, aux méthodes d'histogrammes et au prédicteurs de Nadaraya-Watson.

Théorème 3. *(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) $\exists D > 0, \quad \forall n \in \mathbb{N}, \quad \sum_{i=1}^{n} |W_i(X)| \leqslant D \quad \mathbb{P}$-*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.

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.