

**COURS D'APPRENTISSAGE STATISTIQUE
(SYLVAIN ARLOT ET FRANCIS BACH)
MASTÈRE M2 PROBABILITÉS-STATISTIQUE 2011**

COURS 4 - MÉTHODES À NOYAUX

NOTES DE COURS PRISES PAR LUCIE MONTUELLE ET LINE LE GOFF

1. INTRODUCTION

Données : $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$, $i \in \llbracket 1, n \rrbracket$

But : estimer $f : \mathcal{X} \rightarrow \mathbb{R}$ en minimisant le risque empirique : $\hat{\mathcal{R}}(f) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, f(x_i))$

Pour y arriver, il faut contraindre l'espace de fonctions. Il y a deux méthodes :

– $+\lambda\Omega(f)$

– contrainte : $\Omega(f) \leq B$

Les deux formulations sont à peu de choses près équivalentes. En effet, si on suppose tout convexe, alors on peut écrire le Lagrangien $\mathcal{L}(f, \lambda) = \hat{\mathcal{R}}(f) + \lambda(\Omega(f) - B)$.

Mais la première méthode est plus facile à calibrer que la deuxième. On va donc effectuer la première : minimiser $\hat{\mathcal{R}}(f) + \lambda\Omega(f) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, f(x_i)) + \lambda\Omega(f)$.

On peut prendre $\Omega(f)$ comme la norme dans L^2 ou dans L^1 . Ici on traite le cas L^2 .

2. ESPACES DE HILBERT À NOYAUX REPRODUISANTS (RKHS)

Soit \mathcal{X} un ensemble quelconque

But : Trouver un espace de fonctions où “tout se passe bien” (bornes générales d'apprentissage et optimisation facile).

Exemple 1. – $\mathcal{X} = \mathbb{R}^p$, $f(x) = w^\top x$, norme : $\sqrt{w^\top w}$

– $\mathcal{X} = \mathbb{R}^p$, $\mathcal{F} = L^2(\mathbb{R}^p)$, norme : L^2

Définition 1. Un espace vectoriel de fonctions de \mathcal{X} dans \mathbb{R} est un RKHS s'il est de Hilbert et si les formes linéaires $f \mapsto f(x)$ sont continues pour tout $x \in \mathcal{X}$

Proposition 2.1. Si \mathcal{F} est un RKHS alors il existe une unique fonction Φ de \mathcal{X} dans \mathcal{F} telle que pour tout $x \in \mathcal{X}$, pour tout $f \in \mathcal{F}$, $f(x) = \langle \Phi(x), f \rangle$. Φ est appelée *feature map* et \mathcal{F} *feature space*.

Proposition 2.2. Soit \mathcal{F} un RKHS de *feature map* Φ . Soit $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ tel que $\forall x, y \in \mathcal{X}$, $k(x, y) = \langle \Phi(x), \Phi(y) \rangle$, alors k est une fonction *symétrique définie positive*. On dit que k est un *noyau*.

Définition 2. $K = (k(x_i, x_j))_{i,j}$ est appelée la *matrice de noyau*.

Proposition 2.3. *Propriétés reproduisantes* : Pour tout $x, y \in \mathcal{X}, f \in \mathcal{F}$,

$$(1) \quad f(x) = \langle k(\cdot, x), f \rangle$$

$$(2) \quad k(x, y) = \langle k(\cdot, x), k(\cdot, y) \rangle$$

Démonstration. (1) appliqué à $k(\cdot, y) \Rightarrow$ (2)

$$f(x) = \langle f, \Phi(x) \rangle$$

$$\Phi(x) = k(\cdot, x) \Leftrightarrow \forall y, \Phi(x)(y) = k(x, y) \Leftrightarrow \forall y \langle \Phi(x), \Phi(y) \rangle = k(x, y) \quad \square$$

Remarque 1. $L^2(\mathbb{R}^n)$ n'est pas un RKHS (les formes linéaires d'évaluation n'y sont pas continues).

Remarque 2. Les formes linéaires sont bien continues : $|f(x)| \leq \sqrt{k(x, x)} \|f\|$

Théorème 2.4. *Théorème d'Aronszajn (1950) : k est un noyau défini positif si et seulement si il existe un espace de Hilbert \mathcal{F} , et $\Phi : \mathcal{X} \rightarrow \mathcal{F}$ tel que $\forall x, y, k(x, y) = \langle \Phi(x), \Phi(y) \rangle$.*

Démonstration. \Leftarrow trivial

\Rightarrow a) Soit \mathcal{F}_0 sous espace engendré par tous les vecteurs $k(\cdot, x)$, $x \in \mathcal{X}$.

b) On définit un produit scalaire sur \mathcal{F}_0 par : $\langle \sum_i \alpha_i k(\cdot, x_i), \sum_j \beta_j k(\cdot, y_j) \rangle = \sum_{i,j} \alpha_i \beta_j k(x_i, y_j)$.

Cela revient à définir le produit scalaire sur chacun des éléments générateurs par $\langle k(\cdot, x), k(\cdot, y) \rangle = k(x, y)$.

C'est bien un produit scalaire, car :

– k est bilinéaire, symétrique sur \mathcal{F}_0 .

– Soit $f \in \mathcal{F}_0$, $f = \sum_{i=1}^n \alpha_i k(\cdot, x_i)$ et $\|f\|_{\mathcal{F}_0}^2 = \sum \alpha_i \alpha_j k(x_i, x_j) \geq 0$

– montrer que $\|f\|_{\mathcal{F}_0} = 0 \Rightarrow f = 0$. $\forall x \in \mathcal{X}$, $f(x) = \langle f, k(\cdot, x) \rangle$ et $|f(x)| \leq \|f\|_{\mathcal{F}_0} k(x, x)^{\frac{1}{2}}$

Il ne reste plus qu'à montrer que \mathcal{F}_0 est préhilbertien. Indication : Soit \mathcal{F} le complété de \mathcal{F}_0 (i.e. $\mathcal{F} = \{\text{limites de suites de Cauchy de } \mathcal{F}_0\}$).

□

2.1. Exemples.

Remarque 3. Donc un noyau défini positif fournit :

- un espace de fonctions de \mathcal{X} dans \mathbb{R}
- une norme sur cet espace
- un feature map $\Phi : \mathcal{X} \rightarrow \mathcal{F}$, $k(x, y) = \langle \Phi(x), \Phi(y) \rangle$

Exemple 2. Noyau linéaire : $\mathcal{X} = \mathbb{R}^p$, $k(x, y) = x^t y$, \mathcal{F} est l'espace des fonctions linéaires auquel on associe la norme $\|x \mapsto w^t x\|^2 = w^t w$. $\Phi = id$

Exemple 3. Noyau polynômial : $\mathcal{X} = \mathbb{R}^p$, $k(x, y) = (x^t y)^r$

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

$$\Phi(x) = \left\{ \binom{r}{\alpha_1, \dots, \alpha_p}^{\frac{1}{2}} x_1^{\alpha_1} \dots x_p^{\alpha_p} \right\}$$

$$k(x, y) = \langle \Phi(x), \Phi(y) \rangle$$

$$\mathcal{F} = \{\text{fonctions polynomiales homogènes de degré } p\}$$

Remarque 4. $\dim \mathcal{F} \simeq p^r$ c'est un grand espace. Le noyau nous permet de manipuler un grand espace sans en payer le coût.

Exemple 4. Noyau invariant par translation : $\mathcal{X} = \mathbb{R}^p$, $k(x, y) = q(x - y)$ avec $q : \mathbb{R}^p \rightarrow \mathbb{R}$,

Théorème 2.5. *Théorème de Bôchner : k est défini positif $\Leftrightarrow q$ est la transformée de Fourier d'une mesure de Borel finie positive $\Leftrightarrow q \in L^1$ et sa transformée de Fourier est positive.*

Démonstration. (partielle) Soit $x_1, \dots, x_n \in \mathbb{R}^p$, soit $\alpha_1, \dots, \alpha_n \in \mathbb{R}$,

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

□

Proposition 2.6. *Soit $q \in L^1$ tel que $\hat{q} \in L^1$ et $\forall w \in \mathbb{R}^p$, $\hat{q}(w) \geq 0$, alors $\|f\|_{\mathcal{F}}^2 = \frac{1}{(2\pi)^d} \int \frac{|\hat{f}(w)|^2}{\hat{q}(w)} dw$*

Démonstration. Soit $\langle f, g \rangle = \frac{1}{(2\pi)^d} \int \frac{\hat{f}(w)\overline{\hat{g}(w)}}{\hat{q}(w)} dw$,

$$\begin{aligned} f(x) &= \langle f, k(\cdot, x) \rangle \\ &= \frac{1}{(2\pi)^d} \int \frac{\hat{f}(w)\hat{g}(w) \exp^{-iw^\top y}}{\hat{q}(w)} dw \\ &= \hat{f}(x) = f(x) \end{aligned}$$

□

Exemple 5. Noyau exponentiel : $\mathcal{X} = \mathbb{R}$ et $\hat{q}(w) = \frac{1}{1+w^2}$, $q(x) = e^{-\frac{|x|}{2}}$ alors $\|f\|_{\mathcal{F}}^2 = \frac{1}{2\pi} \int |\hat{f}(w)|^2 (1+w^2) dw = \|f\|_{L^2}^2 + \|f'\|_{L^2}^2$, si $f' \in L^2$. On retrouve la norme de Sobolev.

Exemple 6. Noyau gaussien : $\mathcal{X} = \mathbb{R}$, $q(x) = e^{-\frac{x^2}{2}} = \hat{q}(x)$ alors

$$\begin{aligned} \|f\|_{\mathcal{F}}^2 &= \int |\hat{f}(w)|^2 e^{\frac{w^2}{2}} dw \\ &= \int |\hat{f}(w)|^2 \sum_k \frac{w^{2k}}{2^k k!} dw \\ &= \sum_k \frac{1}{2^k k!} \|f^{(k)}\|_{L^2}^2 \text{ (modulo de bonnes hypothèses)} \end{aligned}$$

$k(x, y) = e^{-\frac{\|x-y\|_2^2}{2\sigma^2}}$ mais Φ n'est pas explicite

Proposition 2.7. Pour le noyau gaussien, \mathcal{F} est dense dans $L^2(\mathbb{R}^p)$.

Proposition 2.8. Soient k_1 et k_2 deux noyaux définis sur le même espace, définis positifs. Alors $k_1 + k_2$ et $k_1 k_2$ sont des noyaux définis positifs.

Pour $k_1 + k_2$, $\mathcal{F} = \mathcal{F}_1 + \mathcal{F}_2$ et $\Phi(x) = \begin{pmatrix} \Phi_1(x) \\ \Phi_2(x) \end{pmatrix}$.

Pour $k_1 k_2$, $\mathcal{F} = \mathcal{F}_1 \otimes \mathcal{F}_2$ et $K = K_1 \otimes K_2$ avec \circ le produit d'Hadamard.

Exemple 7. Noyau sur données non vectorielles : $\mathcal{X} = \{\text{ensemble des séquences d'éléments d'un alphabet donné}\}$. Voir la page de Jean-Philippe Vert (<http://cbio.ensmp.fr/~jvert/teaching/>).

2.2. Noyau de Mercer.

Définition 3. k est un noyau de Mercer, si $\forall (x, y) \in \mathcal{X} \times \mathcal{X}$,

$$\begin{aligned} k(x, y) &= \sum_{n=1}^{\infty} \lambda_n \Phi_n(x) \Phi_n(y) \\ &= \langle (\lambda_n^{\frac{1}{2}} \Phi_n(x))_n, (\lambda_n^{\frac{1}{2}} \Phi_n(y))_n \rangle \end{aligned} \quad \text{avec } \lambda_n \geq 0, \sum \lambda_n < \infty$$

2.3. Lien avec les splines. En fait les splines du monde des statistiques et les RKHS de celui de l'informatique sont à peu de choses près les mêmes notions.

Prenons $\mathcal{F} = \{f : [0, 1] \rightarrow \mathbb{R} / f(0) = 0, f \text{ continue, } f \text{ dérivable pp}\}$. \mathcal{F} est un RKHS associé à $k(x, y) = \min(x, y)$, noyau sur $[0, 1]$. De plus $\|f\|_{\mathcal{F}}^2 = \int_0^1 f'(t) dt$.

3. THÉORÈME DU REPRÉSENTANT

Théorème 3.1. Théorème du représentant (1971) :

Soit \mathcal{F} un RKHS, soit $(x_1, \dots, x_n) \in \mathcal{X}^n$, soit $\Psi : \mathbb{R}^{n+1} \rightarrow \mathbb{R}$ strictement croissante par rapport à sa dernière variable,

alors le minimum de $\Psi(f(x_1), \dots, f(x_n), \|f\|_{\mathcal{F}}^2)$ est atteint pour $f = \sum_{i=1}^n \alpha_i k(\cdot, x_i)$ avec $\alpha \in \mathbb{R}^n$.

Démonstration. soit $f \in \mathcal{F}$, soit $\mathcal{F}_D = \{\sum \alpha_i k(\cdot, x_i) / \alpha \in \mathbb{R}^n\}$, soit $f_D \in \mathcal{F}_D$ et $f_{\perp} \in \mathcal{F}_D^{\perp}$ tel que $f = f_D + f_{\perp}$,

alors $\forall i, f(x_i) = f_D(x_i) + f_\perp(x_i)$ avec $f_\perp(x_i) = \langle f_\perp, k(\cdot, x_i) \rangle = 0$

D'après le théorème de Pythagore, on a : $\|f\|_{\mathcal{F}}^2 = \|f_D\|_{\mathcal{F}}^2 + \|f_\perp\|_{\mathcal{F}}^2$. Par Conséquent, on a :

$$\begin{aligned} \Psi(f(x_1), \dots, f(x_n), \|f\|_{\mathcal{F}}^2) &= \Psi(f_D(x_1), \dots, f_D(x_n), \|f_D\|_{\mathcal{F}}^2 + \|f_\perp\|_{\mathcal{F}}^2) \\ &\geq \Psi(f_D(x_1), \dots, f_D(x_n), \|f_D\|_{\mathcal{F}}^2) \end{aligned}$$

Donc

$$\inf_{f \in \mathcal{F}} \Psi(f(x_1), \dots, f(x_n), \|f\|_{\mathcal{F}}^2) = \inf_{f \in \mathcal{F}_D} \Psi(f(x_1), \dots, f(x_n), \|f\|_{\mathcal{F}}^2)$$

□

Corollaire 3.2. $\min_{f \in \mathcal{F}} \frac{1}{n} \sum \ell(y_i, f(x_i)) + \frac{\lambda}{2} \|f\|_{\mathcal{F}}^2$ est atteint en $f = \sum_{i=1}^n \alpha_i k(\cdot, x_i)$.

Remarque 5. Il est important de remarquer qu'il n'y a aucune hypothèse sur ℓ (pas de convexité).

Écrivons : $\forall j \in [1, n], f(x_j) = \sum_{i=1}^n \alpha_i k(x_i, x_j) = (K\alpha)_j$ où K est la matrice de noyau et $\|f\|^2 = \alpha^\top K\alpha$. On peut alors réécrire :

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum \ell(y_i, f(x_i)) + \frac{\lambda}{2} \|f\|_{\mathcal{F}}^2 = \min_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum \ell(y_i, (K\alpha)_i) + \frac{\lambda}{2} \alpha^\top K\alpha$$

L'astuce du noyau permet donc de :

- remplacer \mathcal{F} 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).

3.1. Théorème du représentant convexe. Posons : $H(f) = \frac{1}{n} \sum \ell(y_i, \langle f, \Phi(x_i) \rangle) + \frac{\lambda}{2} \langle f, f \rangle$ avec $\forall y, u \rightarrow \ell(y, u)$ convexe.

On veut : $\min_{f \in \mathcal{F}, u \in \mathbb{R}^n} \frac{1}{n} \sum \ell(y_i, u_i) + \frac{\lambda}{2} \langle f, f \rangle$ tel que : $\forall i, u_i = \langle f, \Phi(x_i) \rangle$.

Le lagrangien associé à ce problème est :

$$\mathcal{L}(f, u, \alpha) = \frac{1}{n} \sum_i \ell(y_i, u_i) + \frac{\lambda}{2} \langle f, f \rangle + \sum_i \lambda \alpha_i (u_i - \langle f, \Phi(x_i) \rangle)$$

Le problème dual est alors :

$$\inf_{f, u} \mathcal{L}(f, u, \alpha) = \frac{1}{n} \sum_i \underbrace{\inf_{u_i} \{ \ell(y_i, u_i) + n\lambda \alpha_i u_i \}}_{\text{concave en } \alpha} - \frac{\lambda}{2} \alpha^\top K\alpha$$

gradient/ f : $\lambda f - \lambda \sum \alpha_i \Phi(x_i) = 0 \Leftrightarrow f = \sum \alpha_i k(\cdot, x_i)$

Nous avons vu désormais deux problèmes d'optimisation :

- **problème dual (D)** : $\max_{\alpha \in \mathbb{R}^n} J(\alpha) - \frac{\lambda}{2} \alpha^\top K\alpha$
- **problème primal + représentant (P)** : $\min_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum \ell(y_i, (K\alpha)_i) + \frac{\lambda}{2} \alpha^\top K\alpha$

Proposition 3.3. Si α est optimal pour (D), alors α est optimal pour (P).

Cas particulier

Prenons : $H(f) = \frac{1}{2n} \sum (y_i, \langle f, \Phi(x_i) \rangle)^2 + \frac{\lambda}{2} \langle f, f \rangle$. Le problème devient :

$$\min_{f \in \mathcal{F}, u \in \mathbb{R}^n} \frac{1}{n} \sum \frac{1}{2} (y_i - u_i)^2 + \frac{\lambda}{2} \langle f, f \rangle$$

et le lagrangien

$$\inf_{f, u} \mathcal{L}(f, u, \alpha) = \frac{1}{n} \sum_i \inf_{u_i} \{ \frac{1}{2} (y_i - u_i)^2 + n\lambda \alpha_i u_i \} - \frac{\lambda}{2} \alpha^\top K\alpha$$

(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** : $\min_{\alpha \in \mathbb{R}^n} \frac{1}{2n} \|y - K\alpha\|_2^2 + \frac{\lambda}{2} \alpha^\top K\alpha$

Méthode du noyau : Commençons par optimiser par rapport à α
 gradient 1 / α : $-\lambda K\alpha - \frac{\lambda}{n}(n\lambda\alpha - y) = 0 \Leftrightarrow (\lambda K + n\lambda^2)\alpha = \lambda y \Leftrightarrow \alpha = (K + n\lambda I)^{-1}y$ 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 prédiction est unique : $K\alpha = K(K + n\lambda I)^{-1}y$.

Méthode directe : Maintenant optimisons par rapport à f et faisons l'hypothèse que : $\mathcal{F} = \mathbb{R}^p$.
 gradient de H / f : $\frac{1}{n} \sum_i (\langle \Phi(x_i), f \rangle - y_i) \Phi(x_i) + \lambda f = 0 \Leftrightarrow (\frac{1}{n} \sum_i \Phi(x_i) \Phi(x_i)^\top + \lambda I) f = \frac{1}{n} \sum y_i \Phi(x_i)$
 alors avec $\Phi \in \mathbb{R}^{n \times p}$, $f = (\frac{1}{n} \Phi^\top \Phi + \lambda I)^{-1} \frac{1}{n} \Phi^\top y \Leftrightarrow \Phi f = \Phi (\frac{1}{n} \Phi^\top \Phi + \lambda I)^{-1} \frac{1}{n} \Phi^\top y$.

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

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

$n \times n$ $p \times p$

Ce résultat n'est autre que le lemme suivant :

Lemma 3.4. : *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. ANALYSE STATISTIQUE

Cadre classique d'apprentissage statistique : (X_i, Y_i) i.i.d. de loi $P(X, Y)$,
 soit \hat{f} un estimateur, soit $\mathcal{R}(f) = E(\ell(Y, f(X)))$ le risque, $\hat{\mathcal{R}}(f) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, f(x_i))$ le risque empirique et $\mathcal{R}^* = \inf_f (E(\ell(Y, f(X))))$ le risque idéal, soit \mathcal{G} un espace de fonction et $g \in \mathcal{G}$,

$$\mathcal{R}(\hat{f}) - \mathcal{R}^* = \underbrace{\mathcal{R}(\hat{f}) - \hat{\mathcal{R}}(\hat{f})}_{\text{I}} + \underbrace{\hat{\mathcal{R}}(\hat{f}) - \hat{\mathcal{R}}(g)}_{\text{II}} + \underbrace{\hat{\mathcal{R}}(g) - \mathcal{R}(g)}_{\text{III}} + \underbrace{\mathcal{R}(g) - \mathcal{R}^*}_{\text{IV}}$$

II ≤ 0

I et III $\leq \sup_{g \in \mathcal{G}} |\mathcal{R}(g) - \hat{\mathcal{R}}(g)|$

Pour le IV, on minimise par rapport à \mathcal{G} . On obtient :

$$\mathcal{R}(\hat{f}) - \mathcal{R}^* \leq \underbrace{2 \sup_{g \in \mathcal{G}} |\mathcal{R}(g) - \hat{\mathcal{R}}(g)|}_{\text{estimation}} + \underbrace{\inf_{g \in \mathcal{G}} (\mathcal{R}(g) - \mathcal{R}^*)}_{\text{approximation}}$$

Hypothèse : $\mathcal{G} = \{f / \|f\|_{\mathcal{F}}^2 \leq B\}$.

Définition 4. k est un noyau universel, si \mathcal{F} est dense dans $L^2(X)$. Dans ce cas, quand B tend vers $+\infty$, l'erreur d'approximation tend vers zéro.

On suppose que ℓ est lipschitzienne par rapport à sa deuxième variable et borné par M . On a alors (voir [1] pour une preuve utilisant les complexités de Rademacher) :

$$\sup_{\|f\| \leq B} |\mathcal{R}(f) - \hat{\mathcal{R}}(f)| \leq 3M \sqrt{\frac{\log(\frac{2}{\delta})}{2n}} + 2 \frac{LB}{\sqrt{n}}$$

avec probabilité supérieure à $1 - \delta$.

RÉFÉRENCES

[1] Stéphane Boucheron, Olivier Bousquet, Gábor Lugosi, Theory of Classification : a Survey of Some Recent Advances, ESAIM, Volume 9, Juin 2005, pages 323-375