

Apprentissage: cours 2 (1ère partie)

Règles par partition

Sylvain Arlot

23 février 2012

4 Règles par partition en classification et régression

Référence : chapitre 6 de [DGL96]

4.1 Définition

Définition 4 (Règle par partition). Soit $A = \{A_1, \dots, A_k, \dots\}$ une partition finie ou dénombrable de \mathcal{X} et $D_n = (X_i, Y_i)_{1 \leq i \leq n}$ un échantillon. Pour tout $x \in \mathcal{X}$, on note $A(x)$ l'élément de la partition qui contient x et $N_A(x) := \text{card} \{i \in \{1, \dots, n\} \text{ t.q. } X_i \in A(x)\}$. La règle d'apprentissage par partition associée est définie comme suit.

– en régression :

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

– en classification :

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

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

4.2 Rappel : lien entre régression et classification pour les règles plug-in

La règle de classification par partition est une règle de type “plug-in” : on apprend d'abord la fonction de régression par $\hat{\eta}$, puis on considère la règle de classification donnée par $\hat{f} = \mathbb{1}_{\hat{\eta} \geq 1/2}$. Comme vu lors du premier cours (Théorème 1), on a alors un lien entre le risque quadratique de $\hat{\eta}$ en régression et le risque 0-1 de \hat{f} en classification :

$$\begin{aligned} \mathcal{R}^{0-1}(\hat{f}(D_n)) - \mathcal{R}^{0-1}(f^*) &= \mathbb{E} \left[|2\eta^*(X) - 1| \mathbb{1}_{\hat{f}(X) \neq f^*(X)} \mid D_n \right] \\ &\leq 2\mathbb{E} [|\hat{\eta}(X) - \eta^*(X)| \mid D_n] \end{aligned}$$

$$\begin{aligned}
&\leq 2\sqrt{\mathbb{E} \left[(\hat{\eta}(X) - \eta^*(X))^2 \mid D_n \right]} \\
&= 2\sqrt{\mathcal{R}^{reg}(\hat{\eta}(D_n)) - \mathcal{R}^{reg}(\eta^*)} .
\end{aligned}$$

En intégrant par rapport à D_n , on obtient

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

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

4.3 Consistance en classification

Théorème 3. *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 $A_n = \{A_{1,n}, \dots, A_{k,n}, \dots\}_{n \in \mathbb{N}}$ une suite de partitions finies ou dénombrables de $\mathcal{X} = \mathbb{R}^d$. Soit \hat{f}_{A_n} la règle par partition associée, donnée par la Définition 4. Pour tout $E \subset \mathbb{R}^d$, on définit son diamètre (éventuellement infini) $\text{diam}(E) = \sup_{x,y \in E} \{\|x - y\|\}$ où $\|\cdot\|$ désigne la norme euclidienne dans \mathbb{R}^d . Si*

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

alors, la règle de classification $(\hat{f}_{A_n})_{n \in \mathbb{N}}$ est consistante :

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

La démonstration du Théorème 3 repose notamment sur le lemme suivant.

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

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

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

$$\mathbb{E} |\hat{\eta}_n(X; D_n) - \eta^*(X)| \xrightarrow{n \rightarrow +\infty} 0 .$$

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

$$\forall x \in \mathcal{X}, \quad \bar{\eta}_n(x) = \mathbb{E}[\eta^*(X) \mid X \in A_n(x)] .$$

Alors, d'après l'inégalité triangulaire,

$$\mathbb{E} |\hat{\eta}_n(X; D_n) - \eta^*(X)| \leq \mathbb{E} |\hat{\eta}_n(X; D_n) - \bar{\eta}_n(X)| + \mathbb{E} |\bar{\eta}_n(X) - \eta^*(X)| ,$$

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

Contrôle du premier terme $\mathbb{E} |\widehat{\eta}_n(X; D_n) - \bar{\eta}_n(X)|$ 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 $\mathbb{P}(Y_i = 1 \mid X_i \in A(x)) = \bar{\eta}_n(x)$, donc c'est une variable binomiale de paramètres $(N_{A_n}(x), \bar{\eta}_n(x))$. Ainsi,

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

En intégrant, on en déduit

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

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

$$\limsup_{n \rightarrow +\infty} \mathbb{E} |\widehat{\eta}_n(X; D_n) - \bar{\eta}_n(X)| \leq \frac{1}{2\sqrt{k}}$$

pour tout $k > 0$, et donc

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

Contrôle du deuxième terme $\mathbb{E} |\bar{\eta}_n(X) - \eta^*(X)|$ 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 $\mathbb{E} |\eta^*(X) - g(X)| \leq \varepsilon$. Quitte à remplacer g par $\min\{1, \max\{0, g\}\}$, on peut supposer que g est à valeurs dans $[0, 1]$. On définit alors $\bar{g}_n : \mathcal{X} \mapsto \mathbb{R}$ par

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

et l'inégalité triangulaire donne

$$\mathbb{E} |\bar{\eta}_n(X) - \eta^*(X)| \leq \mathbb{E} |\bar{\eta}_n(X) - \bar{g}_n(X)| + \mathbb{E} |\bar{g}_n(X) - g(X)| + \mathbb{E} |g(X) - \eta^*(X)| .$$

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

$$\mathbb{E} |\bar{\eta}_n(X) - \bar{g}_n(X)| = \mathbb{E} |\mathbb{E} [\bar{\eta}_n(X) - \bar{g}_n(X) \mid A_n(X)]| \quad \text{car } \bar{\eta}_n(X) - \bar{g}_n(X) \text{ ne dépend que de } A_n(X)$$

$$\begin{aligned}
&= \mathbb{E} |\mathbb{E} [\eta^*(X) - g(X) \mid A_n(X)]| \\
&\leq \mathbb{E} \mathbb{E} [|\eta^*(X) - g(X)| \mid A_n(X)] \\
&= \mathbb{E} |\eta^*(X) - g(X)| .
\end{aligned}$$

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

$$\mathbb{E} |\bar{g}_n(X) - g(X)| \leq \varepsilon + \mathbb{P}(\text{diam}(A_n(X)) > \theta) \leq 2\varepsilon$$

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

4.4 Consistance universelle pour les partitions régulières en classification

Pour l'instant, le Théorème 3 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 5. 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,

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

Théorème 4. 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 5) et $(\hat{f}_{A^r(h_n)})_{n \in \mathbb{N}}$ la règle de classification par partition associée. Alors, si

$$h_n \rightarrow 0 \quad \text{et} \quad nh_n^d \rightarrow +\infty$$

quand $n \rightarrow +\infty$, la règle $(\hat{f}_{A^r(h_n)})_{n \in \mathbb{N}}$ est universellement consistante : pour toute loi P des données,

$$\mathbb{E} \left[\mathcal{R}^{0-1} \left(\hat{f}_{A^r(h_n)}(D_n) \right) \right] - \mathcal{R}^{0-1}(f^*) \xrightarrow{n \rightarrow +\infty} 0 . \quad (6)$$

La démonstration du Théorème 4 repose notamment sur le lemme suivant.

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

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

Démonstration du Théorème 4. Il suffit de vérifier que les deux hypothèses du Théorème 3 sont satisfaites quelle que soit la loi P . Comme $\text{diam}(A^r(h_n)) = \sqrt{d}h_n^d$, la condition sur le diamètre est satisfaite dès lors que $h_n \rightarrow 0$. Pour la deuxième condition, fixons $M \in]0, +\infty[$ et majorons $\mathbb{P}(N_{A^r(h_n)}(X) \geq M)$.

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 $k_1, \dots, k_d \in \mathbb{Z}$, où $c_1, c_2 > 0$ sont des constantes. On a alors

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

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

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

$$\mathbb{P}(X \in B, N_{A^r(h_n)}(X) \leq M) \leq \frac{2M}{n} .$$

(ii) Soit $\mathbb{P}(X \in B) \geq \frac{2M}{n}$, et alors on va appliquer le Lemme 2, en remarquant que sachant $X \in B$, $Z = N_{A^r(h_n)}(X)$ suit une loi binomiale de paramètres n et $\mathbb{P}(X \in B)$. Ainsi,

$$\begin{aligned} \mathbb{P}(X \in B, N_{A^r(h_n)}(X) \leq M) &= \mathbb{P}(X \in B) \mathbb{P}(N_{A^r(h_n)}(X) \leq M \mid X \in B) \\ &\leq \mathbb{P}(X \in B) \mathbb{P}\left(N_{A^r(h_n)}(X) \leq \frac{n\mathbb{P}(X \in B)}{2} \mid X \in B\right) \\ &\leq \mathbb{P}(X \in B) \frac{4}{n\mathbb{P}(X \in B)} = \frac{4}{n} . \end{aligned}$$

Au final, on a donc

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

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

4.5 Partition et minimisation du risque empirique

Il est intéressant de noter que les règles par partition peuvent aussi s'interpréter comme des règles 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, la règle par partition associée à A minimise le risque empirique sur l'ensemble de prédicteurs*

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

des fonctions constantes sur chaque élément de la partition.

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

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

4.6 Exercice

Exercice 3 (À faire pour la semaine prochaine : m'envoyer votre code source par mail (langage au choix, octave ou matlab conseillés)).

Implémenter la règle de classification par partition régulière de pas $h > 0$ dans $[0, 1]^2$. La faire tourner avec $h = 1/5$ sur un échantillon de taille $n = 1000$ généré comme suit : X_i uniforme sur $[0, 1]^2$ et

$$\eta^*((a, b)) = \sqrt{2} \sqrt{\left(a - \frac{1}{2}\right)^2 + \left(b - \frac{1}{2}\right)^2} .$$

Sur un échantillon, évaluer son risque (avec la perte 0-1) et tracer ce risque en fonction de $h = 1/k$ pour $k \in \{1, \dots, 20\}$.

Références

- [DGL96] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1996.