APPRENTISSAGE STATISTIQUE - MS ENSAI
CORRIGÉ DES EXERCICES. 2025-2026

Prédicteur de Bayes en classification binaire

On considère un n-échantillon de v.a.i.i.d. $\mathscr{D}_n = \{Z_1, ..., Z_n\}$ avec $Z_i = (X_i, Y_i)$. On suppose que les v.a. X sont issues d'un espace $\mathbb{X}$, que les v.a. Y sont issues d'un espace $\mathbb{Y}$ et l'on se donne une loi de probabilité (inconnue) $\mathbb{P}$ sur l'espace $\mathscr{E} = \mathbb{X} \times \mathbb{Y}$. $\mathbb{P}$ est la loi de (X, Y) et également la loi jointe commune des (X_i, Y_i).

Fonction de prédiction $g \in \mathscr{F} = \mathscr{F}(\mathbb{X}, \mathbb{Y})$. Fonction de perte $l : \mathbb{Y}^2 \longrightarrow \mathbb{R}_+$ telles que $l(Y, g(X))$. Le risque de g est

$$R(g) = R_{\mathbb{P}}(g) = \mathbb{E}\left[l(Y, g(X))\right] \tag{1}$$

Le prédicteur de Bayes est l'élément $g^\star$ de $\mathscr{F}$ qui minimise la perte $R(g)$.

$$l(Y, Y') = \mathbb{1}_{[Y \neq Y']} \tag{2}$$

$$\eta(x) = \mathbb{P}[Y = 1 | X = x] = \mathbb{E}[Y | X = x] \tag{3}$$

$$g^\star(x) = \mathbb{1}_{[\eta(x) > 1/2]} \tag{4}$$

1°.

$$\begin{aligned}
&\mathbb{P}[Y = 1 | X = x] \\
&= \mathbb{P}[Y = g(X) | X = x]\mathbb{1}_{[g(x)=1]} + \mathbb{P}[Y = 0 | X = x]\mathbb{1}_{[g(x)=0]} \\
&= \eta(x)\mathbb{1}_{[g(x)=1]} + (1 - \eta(x))\mathbb{1}_{[g(x)=0]}
\end{aligned}$$

2°.

$$\begin{aligned}
&\mathbb{P}[g(X) \neq Y | X = x] = 1 - \mathbb{P}[g(X) = Y | X = x] \\
&= 1 - \mathbb{P}[Y = 1, g(X) = 1 | X = x] - \mathbb{P}[Y = 0, g(X) = 0 | X = x] \\
&= 1 - \mathbb{1}_{[g(x)=1]}\mathbb{P}[Y = 1 | X = x] - \mathbb{1}_{[g(x)=0]}\mathbb{P}[Y = 0 | X = x] \\
&= 1 - \mathbb{1}_{[g(x)=1]}\eta(x) - \mathbb{1}_{[g(x)=0]}(1 - \eta(x))
\end{aligned}$$

Ainsi,

$$\begin{aligned}
&\mathbb{P}[g(X) \neq Y | X = x] - \mathbb{P}[g^\star(X) \neq Y | X = x] \\
&= \eta(x)\left(\mathbb{1}_{[g^\star(x)=1]} - \mathbb{1}_{[g(x)=1]}\right) + ... \\
&... + (1 - \eta(x))\left(\mathbb{1}_{[g^\star(x)=0]} - \mathbb{1}_{[g(x)=0]}\right) \\
&= (2\eta(x) - 1)\left(\mathbb{1}_{[g^\star(x)=1]} - \mathbb{1}_{[g(x)=1]}\right) \\
&\geq 0
\end{aligned}$$

par définition de $g^\star(x)$ et parce que $2\eta(x) - 1 > 0 \iff g^\star(x) = 1$. D'où l'inégalité demandée :

$$\mathbb{P}[Y \neq g^\star(X) | X = x] \leq \mathbb{P}[Y \neq g(X) | X = x] \tag{5}$$

3°.

$$\begin{aligned}
R(g) &= \mathbb{E}[\mathbb{1}_{[g(X) \neq Y]}] \\
&= \mathbb{P}[g(X) \neq Y] = \mathbb{E}\left[\left(g(X) - Y\right)^2\right].
\end{aligned}$$

La dernière égalité provient du fait que d'une part, Y et $g(X)$ valent 0 ou 1 uniquement et d'autre part l'indicatrice de $[g(X) \neq Y]$ et $\left(g(X) - Y\right)^2$ prennent les mêmes valeurs (0 ou 1) quelque soit les valeurs prises par $g(X)$ et Y (0 ou 1).

$$\begin{aligned}
\mathbb{1}_{[g(X) \neq Y]} &= \left(g(X) - Y\right)^2 \\
&= Y^2 + g(X)^2 - 2Yg(X) = Y + g(X)(1 - 2Y)
\end{aligned}$$

(car $Y = Y^2$). Ainsi,

$$\begin{aligned}
R(g) &= \mathbb{E}[Y] + \mathbb{E}[g(X)(1 - 2Y)] \\
&= \mathbb{E}[\mathbb{E}[Y|X]] + \mathbb{E}[g(X)(1 - 2\mathbb{E}[Y|X])] \\
&= \mathbb{E}[\eta(X)] + \mathbb{E}[g(X)(1 - 2\eta(X))]
\end{aligned}$$

Maintenant, si $g = g^\star$,

$$R^\star = R(g^\star) = \mathbb{E}[\eta(X)] + \mathbb{E}[\mathbb{1}_{[\eta(X) > 1/2]}(1 - 2\eta(X))]$$

Si $\eta(X) > 1/2$ l'expression devient

$$\begin{aligned}
R^\star &= \mathbb{E}\left[\eta(X)(1 - 2\mathbb{1}_{[\eta(X) > 1/2]}) + \mathbb{1}_{[\eta(X) > 1/2]}\right] \\
&= \mathbb{E}[1 - \eta(X)]
\end{aligned}$$

et de même, si $\eta(X) < 1/2$,

$$R^\star = \mathbb{E}\left[\eta(X)\right]$$

et donc

$$R^\star = \mathbb{E}[\eta(X) \wedge (1 - \eta(X))]$$

Pour la seconde égalité, rappelons que

$$a \wedge b = (a + b - |a - b|)/2$$

de sorte que

$$R^\star = \mathbb{E}\left[\frac{1}{2}\left(1 - |2\eta(X) - 1|\right)\right] = \frac{1}{2}\left(1 - \mathbb{E}\left[|2\eta(X) - 1|\right]\right)$$

avec $x \wedge y = \inf(x, y)$.

4°.

$$\begin{aligned}
&\mathbb{E}\left[(f(X) - Y)^2 | X = x\right] = \mathbb{E}\left[(f(x) - \eta(x) + \eta(x) - Y)^2 | X = x\right] \\
&= (f(x) - \eta(x))^2 + 2(f(x) - \eta(x))\mathbb{E}[\eta(x) - Y | X = x] + \mathbb{E}[(\eta(X) - Y)^2 | X = x] \\
&= (f(x) - \eta(x))^2 + \mathbb{E}[(\eta(X) - Y)^2 | X = x]
\end{aligned}$$

en passant à l'espérance des deux côtés de l'égalité (par linéarité de l'espérance), il vient :

$$\mathbb{E}\left[(f(X) - Y)^2\right] = (f(x) - \eta(x))^2 + \mathbb{E}[(\eta(X) - Y)^2|].$$

Comme le terme $(f(x) - \eta(x))^2$ est positif, on en déduit l'inégalité recherchée :

$$\mathbb{E}\left[(\eta(X) - Y)^2\right] \leq \mathbb{E}\left[(f(X) - Y)^2\right]$$

Quelque soit la fonction f de $\mathbb{X}$ dans $\mathbb{R}$, $\eta(X)$ minimise donc bien l'erreur quadratique lorsque $f(X)$ prédit Y.

5°. Si X suit une loi uniforme sur $[0, 4c]$, alors

$$\eta(x) = \mathbb{E}[Y | X = x]$$

et

$$R^\star = \mathbb{E}[\eta(X) \wedge (1-\eta(X))] = \mathbb{E}\left[\frac{c \wedge X}{c+X}\right]$$
$$= \frac{1}{4c}\int_0^{4c} \frac{c \wedge x}{c+x} dx = \frac{1}{4}\ln\left(\frac{5e}{4}\right) \simeq 0,3$$

6°. Soit F la fonction de répartition de X. On a :

$$\begin{aligned} F(x) &= \mathbb{P}[X \le x] \\ &= \mathbb{P}[X \le x | Y=0]\mathbb{P}[Y=0] + \mathbb{P}[X \le x | Y=1]\mathbb{P}[Y=1] \\ &= 2x(1-p) + xp = x(2-p) \text{ si } x \in [0,1/2] \\ &= (1-p) + xp \text{ si } x \in [1/2,1] \end{aligned}$$

On en déduit la densité de X en dérivant F :

$$\begin{aligned} f(x) &= 2-p \text{ si } x \in [0,1/2] \\ &= p \text{ si } x \in [1/2,1] \end{aligned}$$

puis

$$\begin{aligned} \mathbb{E}\left[Y\mathbb{1}_{[X\le x]}\right] &= \mathbb{E}\left[\mathbb{1}_{[X\le x]\cap[Y=1]}\right] \\ &= \mathbb{P}\left([X \le x] \cap [Y=1]\right) \\ &= \mathbb{P}\left([X \le x] | [Y=1]\right)\mathbb{P}[Y=1] \\ &= px \end{aligned}$$

Nous allons maintenant calculer $\eta(x)$, en identifiant deux expressions différentes de $\mathbb{E}\left[Y\mathbb{1}_{[X\le x]}\right]$. Rappelons que

$$\eta(x) = \mathbb{E}[Y|X=x]$$

Par ailleurs, en utilisant la loi jointe de (X, Y) on a également

$$\begin{aligned} \mathbb{E}\left[Y\mathbb{1}_{[X\le x]}\right] &= \int\int y\mathbb{1}_{[X\le x]}\mathbb{P}_{(X,Y)}(dx,dy) \\ &= \int\int y\mathbb{1}_{[X\le x]}\mathbb{P}_{(Y/X)}(x,dy)\mathbb{P}_X(dx) \\ &= \int\left(\int y\,\mathbb{P}_{(Y/X)}(x,dy)\right)\mathbb{1}_{[X\le x]}\mathbb{P}_X(x) \\ &= \int_0^x \mathbb{E}[Y|X=u] f_X(u)du \\ &= \int_0^x \eta^\star(u) f_X(u)du = px = \int_0^x p\,du \end{aligned}$$

Si $x < 1/2$, alors $f_X(u) = 2-p$ et par identification avec l'égalité précédente, on a $\eta^\star(x) = p/(2-p)$. De la même façon, si $x > 1/2$, $f_X(u) = p$ et par identification on a alors $\eta^\star(x) = 1$

Finalement, la loi conditionnelle de Y sachant $[X = x]$ est une loi de Bernoulli :

$$\mathbb{P}[Y=1|X=x] = \mathbb{E}[Y|X=x] = \eta^\star(x) = 1 \text{ ou } p/(2-p)$$
$$\mathbb{P}[Y=0|X=x] = 1 - \mathbb{P}[Y=1|X=x] = 0 \text{ ou } 1-p/(2-p)$$

et le prédicteur de Bayes s'exprime de la façon suivante :

$$g^\star(x) = \mathbb{1}_{[\eta^\star(x)>1/2]}$$

La fonction $\phi(p) = p/(2-p)$ est croissante de 0 à 1 lorsque p varie de 0 à 1 et l'on voit facilement que $\phi(p) > 1/2$ dès que $p > 2/3$. Ainsi,

$$g^\star(x) = \begin{cases} \begin{cases} 0 \text{ si } p < 1/3 \text{ et } x \le 1/2 \\ 1 \text{ si } p > 1/3 \text{ et } x \le 1/2 \end{cases} \\ 1 \text{ si } x \ge 1/2 \end{cases}$$

7°. On utilise la formule de Bayes

$$\eta(x) = \mathbb{P}[Y=1|X=x] = \frac{\mathbb{P}[X=x|Y=1]\mathbb{P}[Y=1]}{\mathbb{P}[X=x]}$$

avec

$$\begin{aligned} \mathbb{P}[X=x] &= \mathbb{P}[X=x|Y=1]\mathbb{P}[Y=1] + \mathbb{P}[X=x|Y=0]\mathbb{P}[Y=0] \\ &= \sum_{k=0}^{x} \frac{(2\theta)^k}{k!}e^{-2\theta}p + \sum_{k=0}^{x}\frac{\theta^k}{k!}e^{-\theta}(1-p) \\ &= \sum_{k=0}^{x}\frac{\theta^k}{k!}e^{-\theta} \times \left(e^{-\theta}2^x p + (1-p)\right) \end{aligned}$$

en remplaçant dans l'expression initiale de $\eta(x)$, on obtient le résultat.

Minimisation de l'erreur stochastique

1°. $g^\star$ est par définition le minimiseur de R (sur $\mathcal{G}$), d'où la première inégalité. De la même façon, $\widehat{g}_n$ est par définition le minimiseur de R_n, d'où la seconde inégalité.

2°. Il est clair que

$$R(\widehat{g}_n) - R_n(\widehat{g}_n) \le \sup_{g\in\mathcal{G}} |R(g) - R_n(g)|$$

puisque $\widehat{g}_n$ est une fonction de $\mathcal{G}$. Et de la même façon,

$$R_n(g^\star) - R(g^\star) \le \sup_{g\in\mathcal{G}} |R(g) - R_n(g)|$$

puisque $g^\star$ est une également une fonction de $\mathcal{G}$.

3°. D'après la question 1°, $R(\widehat{g}_n) - R(g^\star) \ge 0$. Par ailleurs,

$$R(\widehat{g}_n) - R(g^\star) = R(\widehat{g}_n) - R_n(\widehat{g}_n) + R_n(\widehat{g}_n) - R_n(g^\star) + R_n(g^\star) - R(g^\star)$$

La différence du centre $R_n(\widehat{g}_n) - R_n(g^\star)$ est négative par définirion de $\widehat{g}_n$. Donc

$$\begin{aligned} R(\widehat{g}_n) - R(g^\star) &\le R(\widehat{g}_n) - R_n(\widehat{g}_n) + R_n(g^\star) - R(g^\star) \\ &\le 2\sup_{g\in\mathcal{G}} |R(g) - R_n(g)| \end{aligned}$$

d'après la question précédente.

Inégalité oracle pour un dictionnaire fini

1°. L'interprétation a été vue en cours : l'écart de risque augmente avec M, mais à M fixé, il tend vers 0 lorsque n tend vers l'infini à la vitesse $1/\sqrt{n}$. L'inégalité ne fait absolument aucune hypothèse sur la loi $\mathbb{P}$. Le terme majorant dépend de M de façon logarithmique. Donc on peut prendre M assez grand (modèle riche) sans sur-apprentissage (par rapport à n). L'inégalité représente un contrôle de l'erreur stochastique pour un dictionnaire fini, qui sert souvent dans les inégalités d'oracle ; on peut comparer les performances de l'estimateur à celles du « meilleur » élément du dictionnaire, avec un terme de pénalisation dépendant de n et $\ln \mathrm{M}$.

Cette inégalité est utilisée partout où un estimateur est choisi parmi un nombre fini de modèles. Elle est la base des méthodes de sélection de modèle, agrégation, apprentissage statistique, PAC-learning, etc.

2°. Une des versions de l'inégalité de Hoeffding est la suivante : si $(\mathrm{V}_i)_i$ est une suite de v.a.i.i.d. à valeurs presque sûrement dans $[a,b]$, alors $\forall\, t>0$,

$$\mathbb{P}\left[\left|\overline{\mathrm{V}}_n - \mathbb{E}[\overline{\mathrm{V}}_n]\right| \geq t\right] \leq 2\exp\left(-\frac{2nt^2}{(b-a)^2}\right)$$

la démonstration (que nous ne ferons pas maintenant) est très intéressante.

3°. On part de l'inégalité sur l'erreur stochastique : on sait que

$$\mathrm{R}(\widehat{g}_n) - \mathrm{R}(g^\star) \leq 2\sup_{g\in\mathscr{G}} |\mathrm{R}(g) - \mathrm{R}_n(g)|$$

Donc,

$$\begin{aligned}
&\mathbb{P}\left[\mathrm{R}(\widehat{g}_n) - \mathrm{R}(g^\star) \geq 2t\right] \leq \mathbb{P}\left[2\sup_{g\in\mathscr{G}} |\mathrm{R}(g) - \mathrm{R}_n(g)| \geq 2t\right] \\
&= \mathbb{P}\left[2\max_{m=1}^{\mathrm{M}} |\mathrm{R}(g_m) - \mathrm{R}_n(g_m)| \geq 2t\right] \\
&= \mathbb{P}\left(\bigcup_{m=1}^{\mathrm{M}} \left[|\mathrm{R}(g_m) - \mathrm{R}_n(g_m)| \geq t\right]\right) \\
&\leq \sum_{m=1}^{\mathrm{M}} \mathbb{P}\left[|\mathrm{R}(g_m) - \mathrm{R}_n(g_m)| \geq t\right] = \bullet
\end{aligned}$$

Mais

$$\mathrm{R}_n(g_m) = \frac{1}{n}\sum_{i=1}^{n} l(\mathrm{Y}_i, g_m(\mathrm{X}_i))$$

Posons alors $\mathrm{V}_i = h_m(\mathrm{X}_i,\mathrm{Y}_i) = l(\mathrm{Y}_i, g_m(\mathrm{Y}_i)) \in [0,1]$. Ces va sont i.i.d. car les $(\mathrm{X}_i,\mathrm{Y}_i)$ sont iid (et g_m est une fonction quelconque de $\mathscr{G}$ qui ne dépend pas de l'échantillon). Par ailleurs, l'espérance de $\mathrm{R}_n(g_m)$ est égale à $\mathrm{R}(g_m)$. On peut alors poser $a=0$ et $b=1$ et appliquer l'inégalité de Hoeffding :

$$\begin{aligned}
\bullet &\leq \sum_{m=1}^{\mathrm{M}} 2\exp\left(-\frac{2nt^2}{(1-0)^2}\right) \\
&= 2\mathrm{M}e^{-2nt^2}
\end{aligned}$$

Pour avoir un majorant égal à ϵ, il suffit alors de poser

$$\begin{aligned}
&\epsilon = 2\mathrm{M}e^{-2nt^2} \\
&\iff t = \frac{1}{2}\sqrt{\frac{2}{n}\ln\left(\frac{2\mathrm{M}}{\epsilon}\right)}.
\end{aligned}$$

Finalement, pour obtenir l'inégalité demandée, il suffit de passer à l'évènement complémentaire.

Arbres de décision

1°. C'est le premier arbre qui a le moins de risque de sur-apprendre, car il possède le moins de feuilles. C'est donc le modèle le moins riche.

2°. $\mathscr{G}$ est la réunion des $\mathscr{G}_i$, donc $\widehat{g}_n \in \mathscr{G}$. Supposons que $\widehat{g}_n \in \mathscr{G}_1$ $(i=1)$. Alors le minimum est atteint sur $\mathscr{G}_1 \subset \mathscr{G}$, donc il coïncide avec celui de $\mathscr{G}_1$. On a alors

$$\begin{aligned}
&\mathrm{R}_n(\widehat{g}_1) \leq \mathrm{R}_n(\widehat{g}_2) \text{ et } \mathrm{R}_n(\widehat{g}_1) \leq \mathrm{R}_n(\widehat{g}_3) \\
\Rightarrow &\mathrm{R}_n(\widehat{g}_1) = \mathrm{R}_n(\widehat{g}_n) \in \operatorname{argmin}_{g=\widehat{g}_1,\widehat{g}_2,\widehat{g}_3} \mathrm{R}_n(g)
\end{aligned}$$

La démonstration est la même si l'on suppose $i=2,3$.

3°. Pour chaque observation, on détermine à quelle feuille elle appartient. Notons-les $\mathrm{F}_1,\ldots,\mathrm{F}_4$ de gauche à droite. On a, dans l'ordre des cases,

$$\mathrm{F}_2 - \mathrm{F}_3 - \mathrm{F}_2 - \mathrm{F}_3 - \mathrm{F}_3 - \mathrm{F}_3 - \mathrm{F}_2 - \mathrm{F}_2 - \mathrm{F}_2 - \mathrm{F}_3$$

avec 0 observations dans F_1 et F_4. On définit $\widehat{g}_1$ sur F_1 et F_4 par tirage à pile ou face équitable. F_2 contient 5 observations de labels : $0,1,1,0,1$ donc $\widehat{g}_1 = 1$ est majoritaire. F_3 contient 5 observations de labels : $0,0,1,1,0$ donc $\widehat{g}_1 = 0$ est majoritaire sur cette feuille. Ainsi,

$$\widehat{g}_1(x) = \mathbb{1}_{\mathrm{F}_2}(x) + \epsilon\mathbb{1}_{\mathrm{F}_1}(x)$$

où ϵ est le résultat d'un tirage à pile ou face.

Adaboost

Soit $h_0 \in \mathscr{F}(\mathbb{X},\mathbb{R})$, $\phi(x) = e^x$ la perte boosting et A_n le ϕ-risque. On cherche le couple $(\widehat{\alpha},\widehat{h})$ qui minimise ce ϕ-risque :

$$(\widehat{\alpha},\widehat{h}) = \epsilon \operatorname*{arg\,min}_{\alpha\geq 0, h\in\mathscr{H}} \mathrm{A}_n(h_0 + \alpha h) \tag{6}$$

L'expression à minimiser s'écrit

$$\begin{aligned}
\mathrm{A}_n(h_0+\alpha h) &= \frac{1}{n}\sum_{i=1}^{n} \exp\left(-\mathrm{Y}_i h_0(\mathrm{X}_i) - \alpha \mathrm{Y}_i h(\mathrm{X}_i)\right) &(7)\\
&= \frac{1}{n}\sum_{i=1}^{n} w_i e^{-\alpha \mathrm{Y}_i h(\mathrm{X}_i)} &(8)\\
&= \frac{1}{n}\sum_{i=1}^{n} w_i e^{-\alpha \mathrm{Y}_i h(\mathrm{X}_i)} \left[\mathbb{1}_{[\mathrm{Y}_i = h(\mathrm{X}_i)]} + \mathbb{1}_{[\mathrm{Y}_i \neq h(\mathrm{X}_i)]}\right] &(9)
\end{aligned}$$

en posant (pour alléger les formules et sans perte de généralité) :

$$w_i = \phi(-Y_i h_0(X_i)) / \sum_{i=1}^{n} \phi(-Y_i h_0(X_i)) \qquad (10)$$

qui ne dépend que de l'échantillon de données et de h_0 (autrement dit on normalise les poids w_i).

Puisque Y_i et $h(X_i)$ sont à valeurs binaires (± 1),
$Y_i = h(X_i) \iff Y_i h(X_i) = 1$ et de même,
$Y_i \neq h(X_i) \iff Y_i h(X_i) = -1$. Ainsi,

$$\begin{aligned}(\widehat{\alpha}, \widehat{h}) &= \underset{\alpha \geq 0, h \in \mathscr{H}}{\arg\min} \left(e^{-\alpha} \frac{1}{n} \sum_{i=1}^{n} w_i \mathbb{1}_{[Y_i = h(X_i)]} + e^{\alpha} \frac{1}{n} \sum_{i=1}^{n} w_i \mathbb{1}_{[Y_i \neq h(X_i)]} \right) \\ &= \underset{\alpha > 0}{\arg\min} \left((e^{\alpha} - e^{-\alpha}) \frac{1}{n} \sum_{i=1}^{n} w_i \mathbb{1}_{[Y_i \neq h(X_i)]} + \frac{1}{n} e^{-\alpha} \right).\end{aligned}$$

$\forall \alpha \geq 0$, le minimum est atteint quand

$$\widehat{h} \in \underset{h}{\arg\min} \sum_{i} w_i \mathbb{1}_{[Y_i \neq h(X_i)]}. \qquad (11)$$

La forme de l'expression permet de minimiser séparemment h et α. $\widehat{h}$ étant choisi pour minimiser h, alors

$$\widehat{\alpha} \in \underset{\alpha > 0}{\arg\min} \left((e^{\alpha} + e^{-\alpha}) \epsilon + e^{-\alpha} \right) = \underset{\alpha}{\arg\min} G(\alpha), \qquad (12)$$

avec $\epsilon = \sum_i w_i \mathbb{1}_{[Y_i \neq h(X_i)]}$.

$G'(\alpha) = (e^{\alpha} - e^{-\alpha})\epsilon - e^{-\alpha}$ et $G''(\alpha) = e^{\alpha}\epsilon + e^{-\alpha}(1-\epsilon)$. Ainsi, G est convexe et sa dérivée s'annule en

$$\alpha = \frac{1}{2} \ln \left(\frac{1-\epsilon}{\epsilon} \right) \qquad (13)$$

que l'on choisit dans l'étape 3 de l'algorithme.

Il est important de noter que la perte Boosting $l_b(x) = e^x$ est convexe et que $l_b(x) \geq l_{0-1}(x)$. Adaboost minimise cette perte en minimisant l'expression $\sum_i w_i \mathbb{1}_{[\bullet]}$.

Bagging

On note (X, Y) une variable aléatoire suivant la loi théorique inconnue, et

$$f(x) = \mathbb{E}[Y \mid X = x]$$

la fonction de régression. Pour un estimateur h évalué en un point x, on rappelle la décomposition biais–variance :

$$\mathbb{E}\left[(h(x) - Y)^2\right] = \underbrace{\left(\mathbb{E}[h(x)] - f(x)\right)^2}_{\text{biais}^2} + \underbrace{\mathbb{V}(h(x))}_{\text{variance}} + \sigma^2(x).$$

1°. Cas de M échantillons indépendants.

On dispose de M échantillons indépendants $\mathscr{D}_{n,1}, \ldots, \mathscr{D}_{n,M}$, tous de taille n. Chaque échantillon sert à entraîner un estimateur de régression au sens des moindres carrés : $h_1, \ldots, h_M$. On définit la moyenne empirique

$$\widehat{h}_M(x) = \frac{1}{M} \sum_{m=1}^{M} h_m(x).$$

Comme les h_m sont i.i.d.,

$$\mathbb{E}[\widehat{h}_M(x)] = \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}[h_m(x)] = \mathbb{E}[h_1(x)],$$

d'où

$$\mathbb{B}(\widehat{h}_M(x)) = \mathbb{B}(h_1(x)).$$

Le biais de $\mathbb{B}(\widehat{h}_M)$ est identique à celui d'un estimateur individuel.
Variance. Par indépendance :

$$\begin{aligned}\mathbb{V}(\widehat{h}_M(x)) &= \mathbb{V}\left(\frac{1}{M} \sum_{m=1}^{M} h_m(x) \right) \\ &= \frac{1}{M^2} \sum_{m=1}^{M} \mathbb{V}(h_m(x)) \\ &= \frac{1}{M} \mathbb{V}(h_1(x)).\end{aligned}$$

La variance de $\widehat{h}_M$ décroît comme $1/M$.

2°. Cas bootstrap : un seul échantillon

On ne dispose plus que d'un seul échantillon

$$\mathscr{D}_n = \{(X_i, Y_i)\}_{i=1}^{n}.$$

On construit M échantillons bootstrap $\mathscr{D}^*_{n,1}, \ldots, \mathscr{D}^*_{n,M}$ en tirant n observations *avec remise* dans $\mathscr{D}_n$ avec la loi uniforme. Pour chaque échantillon bootstrap, on calcule un estimateur h^*_m et on définit

$$h^*_M(x) = \frac{1}{M} \sum_{m=1}^{M} h^*_m(x).$$

Analyse conditionnelle : conditionnellement à $\mathscr{D}_n$, les échantillons bootstrap $\mathscr{D}^*_{n,m}$ sont i.i.d. suivant la loi empirique. Ainsi :

$$\mathbb{E}^*[h^*_M(x) \mid \mathscr{D}_n] = \mathbb{E}^*[h^*_1(x) \mid \mathscr{D}_n],$$

et

$$\mathbb{V}^*\left(h^*_M(x) \mid \mathscr{D}_n\right) = \frac{1}{M} \mathbb{V}^*\left(h^*_1(x) \mid \mathscr{D}_n\right),$$

où $\mathbb{E}^*$ et $\mathbb{V}^*$ désignent les espérances et variances prises uniquement par rapport au mécanisme de bootstrap.

Conclusion conditionnelle : pour un jeu de données fixé, moyenner les estimateurs bootstrap réduit la variance conditionnelle par un facteur $1/M$, sans modifier le biais conditionnel.

Comparaison avec le cas idéal : lorsque n est grand, la loi empirique est proche de la vraie loi : les propriétés biais–variance du bootstrap miment alors celles obtenues dans la question 1°. En pratique :

- le biais global de h^*_M reste proche de celui de h^*_1 ;
- la variance décroît avec M, ce qui réduit l'erreur de généralisation pour des estimateurs instables (par exemple les arbres de régression).

Nom de la méthode : la procédure décrite (tirages bootstrap + moyenne des estimateurs) est le bagging.