Apprentissage statistique

Chapitre 3 : Problème de régression et de classification

Lucie Le Briquer

19 février 2018

Table des matières

1 Régression, $\mathbb{Y} = \mathbb{R}$

Comme dit au chapitre précédent, on est plutôt intéressés par la condition de $Y|X$, mais cela est trop compliqué à apprendre en général.

Soit P un noyau de Markov, régulier, associé à la loi conditionnelle de $Y|X$.

Définition 1 (fonction de régression)

On appelle fonction de régression associée à P :

$$\eta(x) = \int_{\mathbb{Y}} yP(x, dy)$$

Remarque. On a que $\eta(X) = \mathbb{E}[Y|X]$ $\mathbb{P}-$p.s.

En régression, le coût le plus fréquemment utilisé est le coût quadratique/des moindres carrés :

$$(y, y') \longmapsto (y - y')^2$$

Remarque. Si on note $\varepsilon = Y - \eta(X)$, $Y = \eta(X) + \varepsilon$ (ε correspond au bruit), on a $\mathbb{E}[\varepsilon] = 0$ par définition de η.

Cadre des statistiques non-paramétriques.

Propriété 1

$\mathbb{P} \in \mathcal{P} = \{\tilde{\mathbb{P}} \mid \tilde{\mathbb{E}}[Y^2] < +\infty\}$.

1. La fonction de régression est un prédicteur de Bayes pour $\mathbb{P}$ et le coût quadratique.
2. Soit f un autre prédicteur de Bayes, $f = \eta$ $\mathbb{P}_X-$p.s.
3. Le risque de Bayes est :
$$\mathbb{E}[(Y - \eta(X))^2]$$
4. L'excès de risque d'un prédicteur f est donné par :
$$e(f^*, f) = \mathbb{E}[(f(X) - \eta(X))^2]$$

Preuve.

1. Soit $f \in \mathcal{F}(\mathbb{X}, \mathbb{Y})$.

$$\begin{aligned} R_{\mathbb{P}}(f) &= \mathbb{E}[(Y - f(X))^2] \\ &= \mathbb{E}[(Y \pm \eta(X) - f(X))^2] \\ &= \mathbb{E}[(Y - \eta(X))^2] + \mathbb{E}[(\eta(X) - f(X))^2] + 2\mathbb{E}[(Y - \eta(X))\underbrace{(\eta(X) - f(X))}_{\in \mathcal{L}^2(\Omega, \sigma(X), \mathbb{P})}] \end{aligned}$$

Or par définition de l'espérance conditionnelle dans $\mathcal{L}^2$,

$$\eta(X) = \text{proj}^{\perp}_{\mathcal{L}^2(\Omega, \sigma(X), \mathbb{P})}(Y)$$

Ainsi,

$$R_{\mathbb{P}}(f) = \mathbb{E}[(Y - \eta(X))^2] + \mathbb{E}[(\eta(X) - f(X))^2] \underset{*}{\geqslant} \mathbb{E}[(Y - \eta(X))^2] = R_{\mathbb{P}}(\eta)$$

Comme c'est valable $\forall f \in \mathcal{F}(\mathbb{X}, \mathbb{Y})$, on obtient $R^*_{\mathbb{P}} = R_{\mathbb{P}}(\eta)$, et η est un classifieur de Bayes.

2. Si f est aussi un prédicteur de Bayes, alors tous les calculs pour $(*)$ sont des égalités et donc :
$$\mathbb{E}[(f(X) - \eta(X))^2] = 0$$
3. L'inégalité $(*)$ implique le résultat.
4. Trivial.

□

Autres fonctions de coût :

1. Coût seuillé :
$$c(y, y') = \mathbb{1}_{\{|y-y'|\geqslant\alpha\}} \qquad \text{avec } \alpha > 0$$
2. Coût quadratique seuillé :
$$c(y, y') = \min(|y - y'|^2, \alpha^2)$$
3. Coût de Huber :
$$\psi(u) = \begin{cases} u^2 & \text{si } u \leqslant \alpha \\ 2\alpha u - \alpha^2 & \text{si } u \geqslant \alpha \end{cases}$$
4. Coût $\mathcal{L}^p$: $\psi(u) = |u|^p$
5. $$c(y, y') = \mathbb{1}_{]-\infty,-\alpha]}(y - y')$$
$\Rightarrow c = 1$ si $y - y' - \alpha \Rightarrow y + \alpha \leqslant y'$

On observe que les fonctions de coût considérées sont de la forme $c(y, y') = \psi(|y - y'|)$ avec $\psi \uparrow$ et $\psi(0) = 0$.

Remarque. Pour le coût de Huber et celui $\mathcal{L}^1$, les valeurs aberrantes dans l'ensemble d'apprentissage ont beaucoup moins d'influence sur le risque (donc sur les règles d'apprentissage à partir d'ERM=minimisation de risque empirique).

2 Règle d'apprentissage en régression

2.1 Règle d'apprentissage par partition pour la régression

$\mathbb{Y} = \mathbb{R}$. On suppose que l'on dispose d'une partition $\mathcal{P}$ de $\mathbb{X}$ qui est de cardinal dénombrable et constitué de sous-ensembles de $\mathbb{X}$ mesurables.

Pour tout $x \in \mathbb{X}$, on note $P(x)$ l'unique élément de la partition qui contient x, appelé cellule associée à x, et $\forall A \subset \mathbb{X}$ on note :

$$N_A(x_{1\ldots n}) = \text{Card}\{x_i \ : \ x_i \in A\}$$

$(x_1, \ldots x_n) \in \mathbb{X}$.

La règle de régression par partition associée à P est :

$$\hat{f}\big((x_i,y_i)_{1\leqslant i\leqslant n},x\big)=\begin{cases}\frac{1}{N_{P(x)}(x_{1\dots n})}\sum_{i=1}^n y_i\mathbb{1}_{P(x)}(x_i) & \text{si } N_{P(x)}(x_{1\dots n})\neq 0\\ 0 & \text{sinon}\end{cases}$$

pour $(x_i,y_i)_{1\leqslant i\leqslant n}\in(\mathbb{X}\times\mathbb{Y})^n$ et $x\in\mathbb{X}$.

Exemple. (de règle par partition)
Considérons la partition cubique associée à un pas $h>0$:

$$\mathcal{P}=\left\{\prod_{i=1}^d[hk_i,h(k_i+1)[\ \mid\ (k_1,\dots,k_d)\in\mathbb{R}^d\right\}$$

dans le cas $\mathbb{X}=\mathbb{R}^d$. Si $\mathbb{X}\subset\mathbb{R}^d$:

$$\mathcal{P}=\left\{X\cap\prod_{i=1}^d[hk_i,h(k_i+1)[\ \mid\ (k_1,\dots,k_d)\in\mathbb{R}^d\right\}$$

Remarque. On peut considérer et on étudiera des partitions variant avec le nombre d'observations. Par exemple dans le cas de la régression par partition cubique, on va prendre une suite de pas $h_n\xrightarrow[n\to+\infty]{}0$.

2.2 Classification supervisée

$\mathbb{Y}=\{0,1\}$

Remarque. (sur la classification multiclasse)

$$\mathbb{Y}=\{1,\dots,M\}$$

- $i\in\mathbb{Y}$, la première stratégie est de considérer la décomposition binaire de i, cela construit un arbre binaire. et ensuite on effectue une suite de classification binaire.
  $\to$ problème d'accumulation de l'erreur d'apprentissage
- On considère M problèmes binaires $Y=i$ v.s. $Y\neq i$ auxquels on associe un pseudo-classifieur i.e. des fonctions $f_i\colon\mathbb{X}\longrightarrow\mathbb{R}$. Pour la prédiction on utilise alors pour $x\in\mathbb{X}$:

  $$Y=\operatorname{argmax}_i\{f_i(x)\}$$

  Pour simplifier les calculs, on considère dès fois $\mathbb{Y}=\{\pm1\}$ (on a une bijection classique entre les deux : $x\mapsto 2x-1$).
  ⚠ Les résultats théoriques dépendent de la convention choisie.

Dans le cas de la classification binaire on choisit le coût $0-1$ ou de Hamming :

$$c(y,y')=\mathbb{1}_{\Delta_{\mathbb{Y}}}(y,y')=\mathbb{1}_{\{y\neq y'\}}$$

Propriété 2

Soit $\mathbb{Y} = \{0, 1\}$, $\mathbb{P} \in \mathcal{P}$, et c le coût $0-1$.

1. Le classifieur f^* défini pour tout $x \in \mathbb{X}$ par :

$$f^*(x) = \mathbb{1}_{\{\eta(x) > \frac{1}{2}\}}$$

est un classifieur de Bayes pour c et $\mathbb{P}$.

2. Si f est un autre classifieur de Bayes pour c et $\mathbb{P}$, alors :

$$f(x) = f^*(x) \text{ pour } x \notin \left\{\eta(x) = \frac{1}{2}\right\}$$

3. Le risque de Bayes est :

$$R^*_{\mathbb{P}} = \mathbb{E}[\min(\eta(X), 1 - \eta(X))]$$

4. L'excés de risque pour $f \in \mathcal{F}(\mathbb{X}, \mathbb{Y})$ est :

$$e(f, f^*) = \mathbb{E}\big[|2\eta(X) - 1|\mathbb{1}_{\{f(X) \neq f^*(X)\}}\big]$$

Preuve.

1. Soit $f \in \mathcal{F}(\mathbb{X}, \mathbb{Y})$.

$$\begin{aligned} R^{0-1}_{\mathbb{P}}(f) &= \mathbb{E}[\mathbb{1}_{\{f(X) \neq Y\}}] \\ &= \mathbb{E}\left[\mathbb{1}_{\{f(X)=1\}}\mathbb{1}_{\{Y=1\}} + \mathbb{1}_{\{f(X)=1\}}\mathbb{1}_{\{Y=0\}}\right] \\ &= \mathbb{E}\Big[\mathbb{E}\left[\mathbb{1}_{\{f(X)=1\}}\mathbb{1}_{\{Y=1\}} + \mathbb{1}_{\{f(X)=1\}}\mathbb{1}_{\{Y=0\}} \mid X\right]\Big] \\ &= \mathbb{E}\Big[\mathbb{1}_{\{f(X)=1\}}\mathbb{E}\left[\mathbb{1}_{\{Y=1\}} \mid X\right] + \mathbb{1}_{\{f(X)=1\}}\mathbb{E}\left[\mathbb{1}_{\{Y=0\}} \mid X\right]\Big] \end{aligned}$$

Or,

$$\mathbb{E}[\mathbb{1}_{\{Y=1\}} \mid X] = \mathbb{P}(Y = 1|X) = \mathbb{E}[Y|X] = \eta(X)$$

Ainsi,

$$\begin{aligned} R_{\mathbb{P}}(f) &= \mathbb{E}\Big[\eta(X)\mathbb{1}_{\{f(X)=0\}} + (1 - \eta(X))\mathbb{1}_{\{f(X)=1\}}\Big] \\ &\geqslant \mathbb{E}[\min(\eta(X), 1 - \eta(X))] \qquad (*) \end{aligned}$$

Remarque. $\forall x,\ \eta(x) \leqslant 1 - \eta(x) \ \Leftrightarrow\ \eta(x) \leqslant 1 - \eta(x) \ \Leftrightarrow\ \eta(x) \leqslant \frac{1}{2}$

Conclusion la borne est atteinte pour $f = f^*$.

2. Soit f un classifieur de Bayes. $(*)$ devient une égalité, alors :

$$\begin{aligned} \mathbb{E}\big[\eta(X)\mathbb{1}_{\{f(X)=0\}} + (1 - \eta(X))\mathbb{1}_{\{f(X)=1\}}\big] &= \mathbb{E}[\min(\eta(X), 1 - \eta(X))] \\ &= \mathbb{E}[\eta(X)\mathbb{1}_{\{f^*(X)=0\}} + (1 - \eta(X))\mathbb{1}_{\{f^*(X)=1\}}] \end{aligned}$$

Alors,

$$\mathbb{E}\Big[\eta(X)\big[\mathbb{1}_{\{f(X)=0\}}-\mathbb{1}_{\{f^*(X)=0\}}\big]+(1-\eta(X))\big[\mathbb{1}_{\{f(X)=1\}}-\mathbb{1}_{\{f^*(X)=1\}}\big]\Big]=\mathbb{E}[Z]=0$$

La variable aléatoire Z est positive car le raisonnement précédent est valable à X fixé déterministe. Ainsi $Z=0$ $\mathbb{P}-$p.s.
$\Rightarrow \mathbb{1}_{\{f(X)=0\}}=\mathbb{1}_{\{f^*(X)=0\}}$ si $\eta(X)\neq\frac{1}{2}$

En effet, supposons que $\mathbb{1}_{\{f(X)=0\}}\neq\mathbb{1}_{\{f^*(X)=0\}}$. Considérons l'évènement $\{f(X)=0\}\cap\{f^*(X)=1\}$. $Z=\eta(X)-(1-\eta(X))=2\eta(X)-1$. Mais sur

$$\left\{\eta(X)\neq\frac{1}{2}\right\}\cap\{f^*(X)=1\}\subset\left\{\eta(X)>\frac{1}{2}\right\}$$

Z est strictement positive.

3. Déduit de la preuve de (1).

4. Soit $f\in\mathcal{F}(\mathbb{X},\mathbb{Y})$.

$$\begin{aligned}
R_{\mathbb{P}}(f)-R^*_{\mathbb{P}}&=\mathbb{E}[Z]\\
&=\mathbb{E}\Big[\eta(X)\big[\mathbb{1}_{\{\eta(X)>\frac{1}{2}\}}+\mathbb{1}_{\{\eta(X)\leqslant\frac{1}{2}\}}\big]\big(\mathbb{1}_{\{f(X)=0\}}-\mathbb{1}_{\{f^*(X)=0\}}\big)\\
&\quad+(1-\eta(X))\big[\mathbb{1}_{\{\eta(X)>\frac{1}{2}\}}+\mathbb{1}_{\{\eta(X)\leqslant\frac{1}{2}\}}\big]\big(\mathbb{1}_{\{f(X)=1\}}-\mathbb{1}_{\{f^*(X)=1\}}\big)\Big]\\
&=\mathbb{E}\Big[\mathbb{1}_{\{\eta(X)>\frac{1}{2}\}}\big[\eta(X)\mathbb{1}_{\{f(X)=0\}}-(1-\eta(X))\mathbb{1}_{\{f^*(X)=1\}}\mathbb{1}_{\{f(X)=0\}}\big]\\
&\quad+\mathbb{1}_{\{\eta(X)\leqslant\frac{1}{2}\}}\big[(1-\eta(X))\mathbb{1}_{\{f(X)=1\}}-\eta(X)\mathbb{1}_{\{f^*(X)=0\}}\mathbb{1}_{\{f(X)=1\}}\big]\Big]\\
&=\mathbb{E}\Big[\mathbb{1}_{\{f^*(X)=1\}}|2\eta(X)-1|\times\mathbb{1}_{\{f(X)=0\}}+\mathbb{1}_{\{f^*(X)=0\}}|2\eta(X)-1|\times\mathbb{1}_{\{f^*(X)=1\}}\Big]\\
&=\mathbb{E}\Big[|2\eta(X)-1|\mathbb{1}_{\{f(X)\neq f^*(X)\}}\Big]
\end{aligned}$$

□

Exemple. (de règle de classification)

1. Règle de classification par partition. Soit $\mathcal{P}$ une partition de $\mathbb{X}$ dénombrable et mesurable. On considère la règle de régression associée :

$$f((x_i,y_i)_{1\leqslant i\leqslant n},x)=\frac{1}{N_{P(x)}(x_{1,n})}\sum_{=1}^{n}y_i\mathbb{1}_{P(x)}(x_i)$$

La règle de classification associée est :

$$\hat{f}^{0-1}((x_i,y_i),x)=\mathbb{1}_{\{\hat{f}((x_i,y_i),x)>\frac{1}{2}\}}$$

On peut définir, comme dans le cas de la régression, la règle de classification par partition cubique.

2. Règles des k plus proches voisins. $\mathbb{X} = \mathbb{R}^d$ et $\mathcal{X} = \mathcal{B}(\mathbb{R}^d)$ et d'une norme $\|.\|$. Soient $(x_i, y_i)_{1 \leqslant i \leqslant n}$ et $x \in \mathbb{X}$. On définit par récurrence n fonctions $i_1, \ldots, i_n \colon \mathbb{X} \longrightarrow \{1, \ldots, n\}$ par :

$$i_1(x) = \min \left\{ i \mid \|x - x_i\| = \min_j \|x - x_j\| \right\}$$

$$i_2(x) = \min \left\{ i \in \{1, \ldots, n\} \backslash \{i_1(x)\} \mid \|x - x_i\| = \min_{j \neq \{i_1(x)\}} \|x - x_j\| \right\}$$

(cela revient à définir pour un échantillon (x_i, y_i) la suite des indices i tels que $d(x_i, x)$ est croissante)

On définit alors :

$$\hat{\eta}((x_i, y_i)_{1 \leqslant i \leqslant n}, x) = \frac{1}{k} \sum_{j=1}^{k} y_{i_j}(x)$$

On définit par la méthode de classification $k-$NN (k plus proches voisins) pour tout $(x_i, y_i)_{1 \leqslant i \leqslant n}$, x :

$$\hat{f}^{0-1}((x_i, y_i)_{1 \leqslant i \leqslant n}, x) = \mathbb{1}_{\{\hat{\eta}((x_i, y_i), x) > \frac{1}{2}\}}$$

3. Règle plug-in. Soit $\hat{\eta}$ une règle d'apprentissage pour le problème de régression.

Définition 2 (règle plug-in)

On appelle règle plug-in associée à $\hat{\eta}$, la règle d'apprentissage pour le problème de classification définie par :

$$\hat{f}^{\hat{\eta}}((x_i, y_i)_{1 \leqslant i \leqslant n}, x) = \mathbb{1}_{\left\{\hat{\eta}((x_i, y_i)_{1 \leqslant i \leqslant n}, x) > \frac{1}{2}\right\}}$$

Proposition 1

Soit $\hat{\eta}$ une règle de régression et $\hat{f}^{\hat{\eta}}$ la règle de classification plug-in associée. Pour tout $D_n = (x_i, y_i)_{1 \leqslant i \leqslant n}$:

$$R_{\mathbb{P}}(D_n, \hat{f}^{\hat{\eta}}) - R^*_{\mathbb{P}} \leqslant 2\mathbb{E}\left[|\hat{\eta}(X) - \eta(X)| \;\middle|\; D_n\right]$$

Preuve.

D'après la proposition pour le risque $0-1$:

$$\begin{aligned}
e(\hat{f}_n^{\hat{\eta}}, f^*) &= \mathbb{E}\big[|2\eta(X)-1|\mathbb{1}_{\{\hat{f}_n^{\hat{\eta}}(X)\neq f^*(X)\}} \mid D_n\big] \\
&= \mathbb{E}\Big[(2\eta(X)-1)\mathbb{1}_{\{\hat{f}_n(X)=0\}}\mathbb{1}_{\{f^*(X)=1\}} \\
&\quad + (1-2\eta(X))\mathbb{1}_{\{\hat{f}_n(X)=1\}}\mathbb{1}_{\{f^*(X)=0\}} \mid D_n\Big] \\
&= \mathbb{E}\Big[2(\eta(X)-\frac{1}{2})\mathbb{1}_{\{\hat{\eta}_n(X)\leqslant\frac{1}{2}\}}\mathbb{1}_{\{\eta(X)>\frac{1}{2}\}} \\
&\quad + 2(\frac{1}{2}-\eta(X))\mathbb{1}_{\{\hat{\eta}_n(X)<\frac{1}{2}\}}\mathbb{1}_{\{\eta(X)\leqslant\frac{1}{2}\}} \mid D_n\Big] \\
&\leqslant \mathbb{E}\Big[2(\eta(X)-\hat{\eta}_n(X))\mathbb{1}_{\{\hat{\eta}_n(X)\leqslant\frac{1}{2}\}}\mathbb{1}_{\{\eta(X)>\frac{1}{2}\}} \\
&\quad + 2(\hat{\eta}_n(X)-\eta(X))\mathbb{1}_{\{\hat{\eta}_n(X)>\frac{1}{2}\}}\mathbb{1}_{\{\eta(X)\leqslant\frac{1}{2}\}} \mid D_n\Big] \\
&\leqslant 2\mathbb{E}\big[|\hat{\eta}_n(X)-\eta(X)|\big]
\end{aligned}$$

□

Exercice 1

Soit $\omega_0-\omega_1 \geqslant 0$, $\omega_0+\omega_1>0$. On associe le coût asymétrique :

$$c_\omega(y,y') = \omega_{y'}\mathbb{1}_{\{y\neq y'\}} = \omega_0\mathbb{1}_{\{y=1,y'=0\}} + \omega_1\mathbb{1}_{\{y=0,y'=1\}}$$

Proposition 2

Soit $\mathbb{Y}=\{0,1\}$, $\mathbb{P}\in\mathcal{P}$ et c un coût asymétrique.

1. Le classifieur définit par :
$$f_w^*(x) = \mathbb{1}_{\{\eta(x)\geqslant\frac{\omega_1}{\omega_0+\omega_1}\}}$$
est un classifieur de Bayes pour c_ω.
2. Si f est un autre classifieur de Bayes pour c_ω alors $f(X)=\eta(X)$ p.s. sur $\{\eta(X)\neq\frac{\omega_0}{\omega_0+\omega_1}\}$
3. Le risque de Bayes est :
$$R_{\mathbb{P}}^* = \mathbb{E}[\min(\omega_0\eta(X), \omega_1(1-\eta(X)))]$$
4. L'excès de risque pour $f\in\mathcal{F}(\mathbb{X},\mathbb{Y})$ est :
$$(\omega_0+\omega_1)\mathbb{E}\left[\left|\eta(X)-\frac{\omega_1}{\omega_1+\omega_0}\right|\mathbb{1}_{f(X)\neq f_\omega^*(X)}\right]$$

Correction.

- Soit f un classifieur.

$$\begin{aligned}
R_{\mathbb{P}}^{c_\omega}(f) &= \mathbb{E}[c_\omega(Y, f(X))] = \mathbb{E}\big[\omega_0 \mathbb{1}_{Y=1}\mathbb{1}_{f(X)=0} + \omega_1 \mathbb{1}_{Y=0}\mathbb{1}_{f(X)=1}\big] \\
&= \mathbb{E}\big[\omega_0 Y(1-f(X)) + \omega_1(1-Y)f(X)\big] \\
&= \mathbb{E}\Big[\mathbb{E}\big[\omega_0 Y(1-f(X)) + \omega_1(1-Y)f(X) \mid X\big]\Big] \qquad \eta(X) = \mathbb{E}[Y|X] \\
&= \mathbb{E}\big[\omega_0\eta(X)(1-f(X)) + \omega_1(1-\eta(X))f(X)\big] \\
&\geqslant \mathbb{E}\big[\min(\omega_0\eta(X)(1-f(X)), \omega_1(1-\eta(X))f(X)\big]
\end{aligned}$$

On a égalité si $f(X) = 1 \;\Leftrightarrow\; \omega_1(1-\eta(X)) \leqslant \omega_0\eta(X) \;\Leftrightarrow\; \eta(X) \geqslant \frac{\omega_1}{\omega_0+\omega_1}$. Donc

$$f(x) = \mathbb{1}_{\eta(x) \geqslant \frac{\omega_1}{\omega_0+\omega_1}}$$

est un classifieur de Bayes.

- Excès de risque :

$$\begin{aligned}
\rho(f, f^*) &= \mathbb{E}\big[\omega_0\eta(X)(f^*(X) - f(X)) + \omega_1(1-\eta(X))(f(X) - f^*(X))\big] \\
&= \mathbb{E}\big[\mathbb{1}_{f(X)=0}\mathbb{1}_{f^*(X)=1}(\omega_0\eta(X) - \omega_1(1-\eta(X))) \\
&\quad - \mathbb{1}_{f(X)=1}\mathbb{1}_{f^*(X)=0}(\omega_0\eta(X) - \omega_1(1-\eta(X)))\big]
\end{aligned}$$

Sur $f^*(X) = 1$, $\omega_0\eta(X) - \omega_1(1-\eta(X)) \geqslant 0$, sur $f^*(X) = 0$, $\omega_0\eta(X) - \omega_1(1-\eta(X)) \leqslant 0$. D'où :

$$\rho(f, f^*) = (\omega_0 + \omega_1)\mathbb{E}\left[\left|\eta(X) - \frac{\omega_1}{\omega_0+\omega_1}\right| \mathbb{1}_{f(X) \neq f^*(X)}\right]$$

Si f est un classifieur de Bayes, alors $\rho(f, f^*) = 0$. Donc :

$$f(X) = f^*(X) \text{ sur } \eta(X) \neq \frac{\omega_1}{\omega_0+\omega_1}$$