Leçon de mathématiques: no free lunch theorems

Erwan Scornet, encadré par Sylvain Arlot

11 mars 2011

Cet exposé s'intéresse à l'apprentissage statistique, c'est-à-dire à la capacité que l'on a de prédire un phénomène lorsque l'on possède un assez grand nombre de données relatives à ce phénomène. Plus précisément, on voudrait réussir à contrôler l'erreur entre la prédiction et la réalité indépendamment du phénomène étudié (fréquence des tremblements de terre, résultats électoraux, efficacité d'un médicaments sur un panel ...). Pour cela, on rappelle dans la première partie, les définitions qui nous serviront à traiter ces phénomènes puis on montrera que le contrôle de l'erreur ne peut pas être universel.

Table des matières

1 Généralités

Pour prédire un phénomène (par exemple, le temps qu'il fera la semaine prochaine), il convient d'avoir recueilli au préalable un assez grand nombre de données (le temps qu'il a fait le mois dernier).

Définition On appelle *observation* un ensemble de données notée dans la suite X_i à valeurs dans $\mathbb{R}^d$.

Cependant, on ne peut pas prédire exactement la température qu'il fera le lendemain mais on peut essayer de déterminer la classe à laquelle appartient la température (par exemple déterminer si la température est négative ou positive).

Définition On appelle *classe* d'une observation, notée y, la nature inconnue de l'observation à valeur dans un ensemble fini $\{1, ..., M\}$.

Il faut maintenant préciser ce qu'on entend par prédiction : on cherche une fonction qui associe à une observation sa classe .

Définition On appelle *classifieur* une application g telle que

$$\begin{aligned} g : \mathbb{R}^d &\mapsto \{1, ..., M\} \\ X &\mapsto g(X) \ . \end{aligned}$$

$g(X)$ est la prédiction de la classe de X par le classifieur g.

Dans la pratique, un classifieur se construit grâce aux observations : la prédiction faite dépend de ce qu'on a observé précédemment. Remarquons qu'il est nécessaire de posséder un échantillon du type $(X_1, Y_1, ..., X_n, Y_n)$ où X_i est l'observation (température mesurée) et Y_i sa classe (négative ou positive). En effet, posséder l'échantillon $(X_1, ..., X_n)$ n'est pas suffisant pour pouvoir extrapoler la classe Y_{n+1} d'une nouvelle observation X_{n+1}.

Définition (classifieur empirique) Dans la pratique, un *classifieur* g_n se construit mesurablement par rapport à $(X_1, Y_1..., X_n, Y_n)$:

$$\begin{aligned} g_n : \mathbb{R}^d \times (\mathbb{R}^d \times \{1, ..., M\})^n &\mapsto \{1, ..., M\} \\ (X, \times (X_1, Y_1, ..., X_n, Y_n)) &\mapsto g(X; X_1, Y_1, ..., X_n, Y_n) \end{aligned}$$

On considère, dans la suite, que les couples (X_i, Y_i) sont des variables aléatoires indépendantes identiquement distribuées (v.a. i.i.d.).

Cependant, une prédiction n'est pas parfaite et il convient de mesurer l'erreur entre la prédiction et la réalité.

Définition (erreur d'un classifieur) Pour une distribution (X, Y) donnée et pour un classifieur g, on note $L(g)$ l'erreur d'un classifieur définie par :

$$L(g) = \mathbb{P}\{g(X) \neq Y\} \ .$$

Propriété (classifieur de Bayes, erreur de Bayes) Pour une distribution (X, Y) donnée, il existe un meilleur classifieur appelé *classifieur de Bayes* noté g^* vérifiant :

$$g^* = \underset{g:\mathbb{R}^d \mapsto \{1,...,M\}}{\arg\min} \mathbb{P}\{g(X) \neq Y\} \ .$$

On appelle *erreur de Bayes* l'erreur du classifieur de Bayes notée $L^* = L(g^*)$.

Définition (erreur d'un classifieur empirique) L'erreur d'un classifieur empirique est définie de la même manière par :

$$L_n = L(g_n) = \mathbb{P}\{g_n(X; X_1, Y_1, ..., X_n, Y_n) \neq Y | X_1, Y_1, ..., X_n, Y_n\} \ .$$

Le classifieur de Bayes étant le meilleur, on cherchera à construire des classifieurs s'approchant le plus possible de ce dernier.

Définition (règle et consistance) On appelle *règle* une suite de classifieur notée $(g_n)_{n\in\mathbb{N}}$. On dira qu'une règle (g_n) est consistante si :

$$\lim_{n\leftrightarrow+\infty} \mathbb{E}(L_n) = L^* \ .$$

Autrement dit, plus on possède de données, meilleure est la prédiction.

Problème La consistance ne donne pas d'information sur la vitesse de convergence de L_n. Or, sans estimation, cette convergence est inutile : on ne dispose jamais d'une quantité infinie de donnée. Il est assez facile d'obtenir des inégalités de convergence en posant des restrictions sur la loi de (X, Y). Cependant, cela suppose de posséder des informations supplémentaires sur les données.

Afin de gagner en généralité, on aimerait avoir une estimation de la vitesse de convergence de L_n vers L^* qui ne dépend pas de la loi de (X, Y), c'est-à-dire une estimation universelle. Or, on va montre que cela n'est pas possible.

2 Un classifieur universel

2.1 La convergence universelle

Le théorème suivant montre que quel que soit la règle (g_n) considérée, pour chaque $n \in \mathbb{N}$, il existe une distribution (X, Y) qui rende le classifieur g_n extrêmement mauvais.

Théorème 1 Soit $\varepsilon > 0$. Pour tout entier n et pour toute règle (g_n), il existe une distribution (X, Y) avec une erreur de Bayes $L^* = 0$ telle que

$$\mathbb{E}(L_n) \geq \frac{1}{2} - \varepsilon \ . \tag{1}$$

Preuve On va construire une famille $\mathcal{F}$ de distribution (X, Y) tel que l'erreur de Bayes associé à chaque élément soit nulle. Pour tout classifieur, on montrera qu'au moins un élément de cette famille le rend mauvais ce qui conclura le théorème.

Soit $K \in \mathbb{N}$ que l'on choisira plus tard. Pour tout $b = 0, b_1 b_2 ... \in [0, 1]$ (où les b_i correspondent à l'écriture de b en binaire), on considère la distribution suivante :

– X est choisi uniformément sur $\{1, ..., K\}$.

– Y est déterminé par $Y = b_X$.

On a ainsi construit une famille $\mathcal{F}$ contenant 2^K éléments, pour laquelle l'erreur de Bayes associée à chaque élément est nulle (car $Y = f(X)$).

On note $R_n(b) = \mathbb{E}L_n$ l'erreur moyenne prise sur les observations $D_n = ((X_1, Y_1), ..., (X_n, Y_n))$.

Méthode probabiliste Soit $\mathcal{A}$ un ensemble et $f : \mathcal{A} \mapsto \mathbb{R}$. Pour montrer qu'il existe $b \in \mathcal{A}$ tel que $f(b) > \frac{1}{2}$, on introduit une variable aléatoire B à valeur dans $\mathcal{A}$ et on montre que :

$$\mathbb{E}\{f(B)\} > \frac{1}{2} \ . \tag{2}$$

Suite de la preuve Ici, on choisit B uniformément dans $[0, 1]$, indépendant de $X, X_1, ..., X_n$ et on calcul :

$$\begin{aligned}
\sup_{b \in [0,1]} R_n(b) &\geq \mathbb{E}\{R_n(B)\} \\
&= \mathbb{E}\{\mathbb{E}\{L_n(B)|B\}\} \\
&= \mathbb{E}\{L_n(B)\} \\
&= \mathbb{E}\{\mathbb{P}\{g_n(X, D_n) \neq B_X | D_n\}\} \\
&= \mathbb{E}\{\mathbb{P}\{g_n(X, X_1, B_{X_1}, ..., X_n, B_{X_n}) \neq B_X | X_1, B_{X_1}, ..., X_n, B_{X_n}\}\} \ .
\end{aligned}$$

Or B_X est indépendant de $X, X_1, B_{X_1}, ..., X_n, B_{X_n}$ donc de $g_n(X, X_1, B_{X_1}, ..., X_n, B_{X_n})$ si $X \neq X_i$ pour tout $i \leq n$. De plus, si

$$\begin{cases} \text{Y} & \text{est indépendant de X} \\ \text{X} & \text{suit une loi de bernoulli non biaisé sur } \{0,1\} \end{cases}$$

alors $\mathbb{P}\{Y = X\} = \mathbb{P}\{Y \neq X\} = \frac{1}{2}$

On a donc :

$$
\begin{aligned}
& \mathbb{E}\{\mathbb{P}\{g_n(X, X_1, B_{X_1}, ..., X_n, B_{X_n}) \neq B_X | X_1, B_{X_1}, ..., X_n, B_{X_n}\}\} \\
\geq\ & \frac{1}{2}\mathbb{E}\{\mathbb{P}\{X \neq X_1, X \neq X_2, ..., X \neq X_n | X_1, B_{X_1}, ..., X_n, B_{X_n}\}\} \\
=\ & \frac{1}{2}\mathbb{E}\left\{\prod_{i=1}^{n} \mathbb{P}\left\{X \neq X_i | X_1, B_{X_1}, ..., X_n, B_{X_n}\right\}\right\} \\
& \textit{(car les événements } X \neq X_i \textit{ sont indépendants)} \\
=\ & \frac{1}{2}\mathbb{E}\left\{\prod_{i=1}^{n} \mathbb{P}\left\{X \neq X_i | X_i\right\}\right\} \\
& \textit{(car les } X_i \textit{ suivent la même loi)} \\
=\ & \frac{1}{2}\mathbb{E}\left\{(1 - \frac{1}{K})^n\right\} \\
& \textit{(car les } X_i \textit{ sont uniformément répartis sur } \{1, ..., K\}\ ) \\
=\ & \frac{1}{2}(1 - \frac{1}{K})^n \\
& \rightarrow_{K \to +\infty} \frac{1}{2}\ .
\end{aligned}
$$

Ce qui conclut la preuve.

2.2 Un taux de convergence universel

On vient de montrer qu'on peut trouver une distribution $\mathcal{L}_n = (X, Y)_n$ telle que l'estimateur $L_{n,\mathcal{L}_n}$ soit très mauvais. Cependant, on peut toujours espérer que tous les classifieurs consistant convergent avec la même vitesse. Le théorème 1 n'interdit pas la proposition suivante :

$$\forall \mathcal{L}_{(X,Y)},\ \exists c > 0 \text{ telle que, pour toute règle } g_n,\ |\mathbb{E}L_n - L^*| \leq \frac{c}{n} \text{ pour tout } n$$

Cependant, elle est fausse. On ne peut pas contrôler universellement la vitesse de convergence d'un classifieur consistant :

Théorème 2 Soit $(a_n) \in {\mathbb{R}_+}^{\mathbb{N}}$ telle que :
- $\lim_{n \to +\infty} a_n = 0$
- $\frac{1}{16} \geq a_1 \geq a_2 \geq ...$

Pour toute règle (g_n), il existe une distribution (X, Y) avec $L^* = 0$ telle que

$$\mathbb{E}L_n \geq a_n \quad \text{pour tout n}\ . \tag{3}$$

Preuve Pour la démonstration, on admet le lemme technique suivant (c.f. démonstration en annexe) :

Lemme Pour toute suite (a_n) positive, décroissante vers 0 avec $\frac{1}{16} \geq a_1$, on peut trouver une distribution de probabilité $(p_1, ...)$ telle que

1. $p_1 \geq p_2 \geq ...$
2. $\sum_{i=n+1}^{+\infty} p_i \geq \max(8a_n, 32np_{n+1})$

Comme précédemment, on note $b = 0, b_1 b_2 ...$ l'écriture en binaire de b pour $b \in [0,1]$. De même, pour une variable aléatoire B à valeur dans $[0,1]$, on note $B = 0, B_1 B_2 ...$ son écriture binaire (où les B_i sont des variables aléatoires à valeurs dans $\{0,1\}$).

Comme pour le théorème 1 , on va définir une famille $\mathcal{F}$ de loi sur (X,Y) et on va montrer qu'au moins un élément de cette famille vérifie l'inégalité du théorème 2. Soit p_i une distribution de probabilité dont on précisera plus tard les propriétés.

Commençons par définir X :

$$\begin{cases} \mathbb{P}\{X = i\} = p_i & \text{pour tout } \ i \geq 1 \\ p_1 \geq p_2 \geq ... \geq 0 & \text{pour tout } \ i \geq 1 \end{cases}$$

On définit $Y = B_X$ où B est une variable aléatoire uniformément distribué sur $[0,1]$. Définissons :

$$\begin{cases} \Delta_n = ((X_1, B_{X_1}), ..., (X_n, B_{X_n})) \\ G_{n,i} = g_n(i, \Delta_n) \\ L_n(B) = \mathbb{P}\{g_n(X, \Delta_n) \neq Y | B, \Delta_n\} \end{cases}$$

On a, avec les notations précédentes,

$$L_n(B) = \sum_{i=1}^{+\infty} p_i \mathbb{1}_{\{G_{n,i} \neq B_i\}} \ . \tag{4}$$

De plus,

$$\begin{aligned} \sup_b \inf_n \mathbb{E}\left\{\frac{L_n(b)}{a_n}\right\} &\geq \sup_b \mathbb{E}\left\{\inf_n \frac{L_n(b)}{a_n}\right\} \\ &\geq \mathbb{E}\left\{\inf_n \frac{L_n(b)}{a_n}\right\} \\ &= \mathbb{E}\left\{\mathbb{E}\left\{\inf_n \frac{L_n(B)}{a_n} | \Delta_n\right\}\right\} \ . \end{aligned} \tag{5}$$

En s'intéressant uniquement à l'espérance conditionnelle qui apparaît dans (5), on a :

$$\begin{aligned} \mathbb{E}\left\{\inf_n \frac{L_n(B)}{a_n} | \Delta_n\right\} &\geq \mathbb{P}\{\cap_{n=1}^{+\infty} \{L_n(B) \geq a_n\} | \Delta_n\} \\ &\geq 1 - \sum_{n=1}^{+\infty} \mathbb{P}\{L_n(B) < a_n | \Delta_n\} \\ &= 1 - \sum_{n=1}^{+\infty} \mathbb{P}\{L_n(B) < a_n | \Delta_n\} \ . \end{aligned} \tag{6}$$

On calcule :

$$\begin{aligned}
\mathbb{P}\{L_n(B) < a_n|\Delta_n\} &\leq \mathbb{P}\left\{\sum_{i\notin\{X_1,\dots,X_n\}} p_i \mathbb{1}_{\{G_{n,i}\neq B_i\}} < a_n|\Delta_n\right\} \\
&= \mathbb{P}\left\{\sum_{i\notin\{X_1,\dots,X_n\}} p_i \mathbb{1}_{\{B_i=1\}} < a_n|\Delta_n\right\} \\
&\textit{(par indépendance des } B_i \textit{ et mesurabilité des } G_{n,i} \textit{ par rapport à } \Delta_n\textit{)} \\
&\leq \mathbb{P}\left\{\sum_{i=n+1}^{+\infty} p_i \mathbb{1}_{\{B_i=1\}} < a_n\right\} \\
&\textit{(par décroissance des } p_i\textit{)} \\
&= \mathbb{P}\left\{\sum_{i=n+1}^{+\infty} p_i B_i < a_n\right\} . \qquad (7)
\end{aligned}$$

Inégalité de Chernoff D'après l'inégalité de Markov, pour toute variable aléatoire X et pour tout $s > 0, \varepsilon \geq 0$, on a :

$$\mathbb{P}\{X \geq \varepsilon\} = \mathbb{P}\{e^{sX} \geq e^{s\varepsilon}\} \leq \frac{\mathbb{E}\{e^{sX}\}}{e^{s\varepsilon}} .$$

edEn appliquant cette méthode à (7), on obtient :

$$\begin{aligned}
\mathbb{P}\{\sum_{i=n+1}^{+\infty} p_i B_i < a_n\} &\leq \mathbb{E}\left\{\exp\left(s\left(a_n - \sum_{i=n+1}^{+\infty} p_i B_i\right)\right)\right\} \\
&= e^{sa_n}\Pi_{i=n+1}^{+\infty}\frac{1}{2}\left(1+e^{-sp_i}\right) \\
&\textit{(car } B_i \textit{ est une bernoulli de paramètre } \frac{1}{2}\textit{)} \\
&\leq e^{sa_n}\Pi_{i=n+1}^{+\infty}\left(1-\left(\frac{sp_i}{2}+\frac{s^2p_i^2}{4}\right)\right) \\
&\textit{(car } e^{-x} \leq 1-x+\frac{x^2}{2} \textit{ pour } x \geq 0\textit{)} \\
&\leq e^{sa_n}\exp\left(\sum_{i=n+1}^{+\infty}\left(-\frac{sp_i}{2}+\frac{s^2p_i^2}{4}\right)\right) \\
&\textit{(car } 1-x \leq e^{-x}\textit{)} \\
&\leq e^{sa_n}\exp\left(-\frac{s\sum}{2}+\frac{s^2p_{n+1}\sum}{4}\right) \qquad (8) \\
&\textit{(par décroissance des } p_i\textit{, en posant } \sum = \sum_{i=n+1}^{+\infty} p_i\textit{)} \\
&= \exp\left(-\frac{1}{4}\frac{(2a_n-\sum)^2}{\sum p_{n+1}}\right) \\
&\textit{(en évaluant (8) en } s = \frac{\sum -2a_n}{p_{n+1}\sum} > 0 \textit{ car } \sum > 2a_n \textit{)} \\
&\leq \exp\left(-\frac{1}{16}\frac{\sum}{p_{n+1}}\right) \\
&\textit{(car } \sum \geq 4a_n\textit{)} \\
&\leq e^{-2n} \\
&\textit{(en supposant } \sum \geq 32p_{n+1}n\textit{)} \qquad (9)
\end{aligned}$$

En reprenant les inégalités (5), (6) et (7), on obtient :

$$\begin{aligned}
\sup_b \inf_n \mathbb{E}\left\{\frac{L_n(b)}{a_n}\right\} &\geq \mathbb{E}\left\{\mathbb{E}\left\{\inf_n \frac{L_n(b)}{a_n}|\Delta_n\right\}\right\} \\
&\geq 1-\sum_{n=1}^{+\infty}\mathbb{P}\left\{L_n(B) < a_n|\Delta_n\right\} \\
\text{avec } \mathbb{P}&\left\{\sum_{i=n+1}^{+\infty} p_i B_i < a_n\right\} \leq e^{-2n}
\end{aligned}$$

D'où,

$$\begin{aligned}\sup_b \inf_n \mathbb{E}\left\{\frac{L_n(b)}{a_n}\right\} &\geq 1 - \sum_{n=1}^{+\infty} e^{-2n} \\ &= \frac{e^2-2}{e^2-1} \\ &\geq \frac{1}{2}\,.\end{aligned}$$

On a donc obtenu sous réserve d'avoir

$$\left\{\begin{array}{l}\sum \geq 4a_n \\ \sum \geq 32 p_{n+1} n\end{array}\right. \tag{10}$$

l'existence d'un $b \in [0,1]$ tel que

$$L_n(b) \geq \frac{a_n}{2} \tag{11}$$

Posons $b_n = \frac{a_n}{2}$. Si b_n vérifie les hypothèses du lemme alors, il existe une distribution p_i vérifiant :

$$\left\{\begin{array}{l}\sum \geq 8b_n \\ \sum \geq 32 p_{n+1} n\end{array}\right.$$

Donc (10) est vérifiée. En transposant (11) à la suite b_n, on obtient que, quelle que soit la suite b_n vérifiant les hypothèses du lemme,

$$L_n(b) \geq b_n\ ,$$

ce qui termine la preuve.

3 Convergence universelle vers l'erreur de Bayes

On a vu qu'on recherchait des classifieurs qui minimisent l'erreur ou tout du moins qui tendent vers le classifieur de Bayes qui lui minimise l'erreur. Cependant si le classifieur de Bayes est mauvais (i.e. si L^* est trop éloigné de 0), trouver une règle consistante ne sert à rien (son erreur sera celle du classifieur de Bayes, c'est-à-dire trop grande). Une question qu'on peut alors se poser est : comment estimer l'erreur de Bayes ? Peut-on contrôler universellement la vitesse de convergence d'un tel estimateur ?

Là encore, comme précédemment, la réponse est négative. Quel que soit l'estimateur de l'erreur de Bayes que l'on considère, pour tout rang n , il existe une distribution (X, Y) telle que cet estimateur soit mauvais. C'est ce que montre le théorème suivant.

Théorème 3 Pour tout n, pour tout estimateur $\hat{L}_n$ de l'erreur de Bayes L^*, pour tout $\varepsilon > 0$, il existe une distribution (X, Y) telle que :

$$\mathbb{E}\{|\hat{L}_n - L^*|\} \geq \frac{1}{4} - \varepsilon \tag{12}$$

Preuve Soit $n \in \mathbb{N}$. On va construire une famille $\mathcal{F}$ de distributions et on va montrer que (12) est vraie pour au moins une d'entre elle. Soit $m \in \mathbb{N}$. On choisit $X_1, ..., X_n$ indépendamment et uniformément dans $\{1, ..., m\}$. Soit $B_0, ..., B_n$ des variables de Bernoulli i.i.d., de paramètre $\frac{1}{2}$, indépendants des X_i. On définit :

- le premier membre de la famille $\mathcal{F}$. On pose $Y_i = B_i$. On a alors $L^* = \frac{1}{2}$ (car les Y_i sont indépendants des X_j).
- les 2^m membres restant de $\mathcal{F}$ en posant $Y_i = a_{X_i}$ où les m paramètres $a_1, ..., a_m$ décrivent $\{0, 1\}^m$. Pour chaque membre de cette famille, $L^* = 0$ (cf théorème 1)

Remarquons que toute distribution avec X uniformément distribué sur $\{1, ..., m\}$ et telle que $L^* = 0$ sont dans $\mathcal{F}$ ($L^* = 0$ signifie que $Y = f(X)$ et 2^m est le cardinal de $\{0, 1\}^{\mathcal{F}}$). Comme dans les preuves des théorèmes précédents, on va incorporer de l'aléatoire dans cette famille. On choisit une distribution de la manière suivante :

- Si $B_0 = 0$ on choisit le premier membre de $\mathcal{F}$.
- Si $B_0 = 1$, on construit les Y_i de la manière suivante :

$$Y_i = \begin{cases} B_i \text{ si } X_i \notin \{X_1, ..., X_{i-1}\} \\ B_j \text{ si } j < i \text{ est le plus petit indice tel que } X_i = X_j \end{cases} \tag{13}$$

Remarquons que, si $B_0 = 1$ pour toute réalisation $b_1, ..., b_n$ de $B_0, ..., B_n$ la construction devient déterministe et donc $L^* = 0$. Donc la construction est dans $\mathcal{F}$ d'après (3). Soit $A = \{$ tous les X_i sont différents $\}$. Remarquons que sous A, $\hat{L}_n$ est une fonction de $X_1, ..., X_n, B_1, ..., B_n$ mais pas de

B_0. On a donc :

$$\begin{aligned}
\sup_{\mathcal{F}} \mathbb{E}\{|\hat{L}_n - L^*|\} &\geq \mathbb{E}\{|\hat{L}_n - L^*|\} \\
&\geq \mathbb{E}\{\mathbb{1}_A|\hat{L}_n - L^*|\} \\
&= \mathbb{E}\left\{\mathbb{1}_A\left(\mathbb{1}_{B_0=0}|\hat{L}_n - \frac{1}{2}| + \mathbb{1}_{B_0=1}|\hat{L}_n - 0|\right)\right\} \\
&= \mathbb{E}\left\{\frac{\mathbb{1}_A}{2}\left(|\hat{L}_n - \frac{1}{2}| + |\hat{L}_n - 0|\right)\right\} \\
&\geq \mathbb{E}\left\{\frac{\mathbb{1}_A}{4}\right\} \\
&= \frac{1}{4}\mathbb{P}\{A\} \qquad (14)
\end{aligned}$$

Or, si on choisit m assez grand, les X_i étant répartis uniformément sur $\{1, ..., m\}$, on peut rendre $\mathbb{P}\{A\}$ aussi proche de 1 que l'on veut. D'où

$$\sup_{\text{loi de } (X,Y)} \mathbb{E}\{|\hat{L}_n - L^*|\} \geq \frac{1}{4} \qquad (15)$$

Bibliographie

Cet article est très largement inspiré de :

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