Agrégation externe de mathématiques, session 2008
Épreuve de modélisation, option A : Probabilités et Statistiques

(*public 2008*)

Résumé : En classification supervisée, un label est associé à une observation. Ce label, obtenu par exemple à l'aide d'un avis d'expert, est censé représenter la nature de l'observation. En récoltant plusieurs observations, on peut ainsi réunir un échantillon de données -appelé échantillon d'apprentissage- constitué des observations et de leurs labels. Le but de l'apprentissage en classification supervisée est alors de construire une méthode (ou règle) de classification automatique pour une nouvelle observation, celle-ci se passant d'une nouvelle consultation d'un expert. Le texte est consacré à la modélisation de ce type de problématique et à la construction de 2 règles de classification.

Mots clefs : Échantillon, espérance conditionnelle, convergence.

► *Il est rappelé que le jury n'exige pas une compréhension exhaustive du texte. Vous êtes laissé(e) libre d'organiser votre discussion comme vous l'entendez. Des suggestions de développement, largement indépendantes les unes des autres, vous sont proposées en fin de texte. Vous n'êtes pas tenu(e) de les suivre. Il vous est conseillé de mettre en lumière vos connaissances à partir du fil conducteur constitué par le texte. Le jury appréciera que la discussion soit accompagnée d'exemples traités sur ordinateur.*

1. Introduction - Modélisation du phénomène

A l'interprétation de certaines caractéristiques financières d'un nombre donné d'entreprises, un analyste financier considèrera, dans le cadre d'une classification binaire, que chacune de ces entreprises est de type 1 ou 0, selon qu'elle est viable ou non. En se basant sur cet échantillon de données -appelé échantillon d'apprentissage- le but final de l'apprentissage en classification est de construire un procédé de classification automatique pour une nouvelle entreprise, ce procédé se passant ainsi d'une nouvelle consultation de l'analyste financier. Le propos de ce texte est de modéliser ce type de situation et de construire 2 règles d'apprentissage.

A chaque observation $x \in \mathbf{R}^d$, on associe un label y, représentant la nature de l'observation. Seul le cas d'une classification binaire est considéré, *i.e.*, y prend la valeur 0 ou 1. En l'absence d'un avis d'expert, une fonction $g : \mathbf{R}^d \to \{0,1\}$ est introduite de sorte que $g(x)$ représente la décision à prendre pour l'observation x : une telle fonction g s'appelle règle de décision et cette règle commet une erreur lorsque $g(x) \neq y$.

Un cadre probabiliste est adopté pour modéliser la situation : soit (X,Y) un vecteur aléatoire à valeurs dans $\mathbf{R}^d \times \{0,1\}$. La probabilité d'erreur associée à la règle g est $L(g) = \mathbb{P}(g(X) \neq Y)$.

En général, la plus petite probabilité d'erreur commise dans l'ensemble des règles de décision est non nulle : par exemple, si $\eta(X) = 1/2$ presque sûrement, où η est la fonction telle que $\eta(X) = \mathbb{P}(Y = 1|X) = \mathbb{E}[Y|X]$, alors pour toute règle g, $L(g) = 1/2$. Cependant, dans cet exemple caricatural, il n'est de toutes façons pas possible de trancher entre le label 0 ou le label 1. Par contre, il est intuitivement clair que la plus petite probabilité d'erreur est d'autant plus proche de 0 que $\eta(X)$ est proche de 0 ou 1 presque sûrement, ce qui se produit notamment lorsque Y est seulement fonction de la variable aléatoire X. Le choix de la règle de décision est alors crucial : si $\eta(X) = 1$ presque sûrement, $L(g) = 1$ pour la règle $g \equiv 0$, bien que la plus petite probabilité d'erreur soit nulle.

Les règles de décision de faible probabilité d'erreur dépendent de la loi, inconnue le plus souvent, du vecteur aléatoire (X,Y). Cependant, des expériences successives peuvent fournir un échantillon d'apprentissage représenté par une suite $(X_1,Y_1),\cdots,(X_n,Y_n)$ de vecteurs aléatoires indépendants et de même loi que (X,Y). En n'utilisant que cet échantillon d'apprentissage, il s'agit de construire une règle de décision g_n permettant de classer une nouvelle observation, notée encore X, de label Y, tel que (X,Y) est indépendant de l'échantillon. Un tel procédé de construction s'appelle apprentissage supervisé et la pertinence de la règle de décision empirique g_n est mesurée par :

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

La règle empirique g_n est dite convergente si, lorsque $n \to \infty$ (ce qui sera implicite dans la suite) $L_n(g_n) \to \inf_g L(g)$ dans L_1, ou de manière équivalente, puisque $\inf_g L(g) \leq L_n(g_n)$, si $\mathbb{E}L_n(g_n) \to \inf_g L(g)$, l'infinimum étant pris sur l'ensemble des règles de décision.

2. La règle de décision de Bayes

La règle de Bayes g^* est une règle optimale au sens où elle minimise la probabilité d'erreur. Elle est définie par $g^*(x) = 1$ si $\eta(x) > 1/2$, et 0 sinon.

Proposition 1. *On a :* $L(g^*) = \mathbb{E}\min(\eta(X), 1-\eta(X))$. *De plus, pour toute règle de décision* g *:* $L(g^*) \leq L(g)$.

Preuve. La probabilité d'erreur conditionnelle de toute règle de décision g peut s'écrire

$$\mathbb{P}(g(X) \neq Y|X = x) = 1 - \Big(\mathbf{1}_{\{g(x)=1\}}\eta(x) + \mathbf{1}_{\{g(x)=0\}}(1-\eta(x))\Big). \tag{1}$$

On en déduit que $L(g) - L(g^*) \geq 0$ et que $L(g^*) = \mathbb{E}\min(\eta(X), 1-\eta(X))$ pour la règle de Bayes. □

On retrouve avec cette proposition la propriété que $L(g^*)$ est d'autant plus proche de 0 que $\eta(X)$ est proche de 0 ou 1 presque sûrement.

Les méthodes de classification sont souvent basées sur la règle de Bayes. Or, celle-ci fait intervenir la fonction η, qui est le plus souvent inconnue, mais que l'on peut estimer à l'aide de

l'échantillon $(X_1,Y_1),\cdots,(X_n,Y_n)$. On introduit alors une règle de décision du type $\tilde{g}(x)=1$ si $\tilde{\eta}(x)>1/2$ et 0 sinon, où $\tilde{\eta}$ est une fonction censée approcher η.

Proposition 2. *Pour la règle de décision $\tilde{g}$ définie ci-dessus, on a l'inégalité : $L(\tilde{g})-L(g^*)\leq 2\mathbb{E}|\tilde{\eta}(X)-\eta(X)|$.*

Preuve. Reprenant (1), on déduit que

$$L(\tilde{g})-L(g^*)=\mathbb{E}\Big[|2\eta(X)-1|\mathbf{1}_{\{\tilde{g}(X)\neq g^*(X)\}}\Big],$$

d'où le résultat, car la propriété $\tilde{g}(x)\neq g^*(x)$ entraîne que $|\eta(x)-\tilde{\eta}(x)|\geq|\eta(x)-1/2|$. □

3. Les méthodes d'apprentissage décidant selon la majorité

De nombreuses méthodes d'apprentissage sont basées sur le principe suivant : on affecte le label 0 (*resp.* 1) à la nouvelle observation X si, dans l'échantillon d'apprentissage, on compte plus d'observations de label 0 (*resp.* 1) que d'observations de label 1 (*resp.* 0) à proximité de X.

Poursuivant cette idée, on note, avec une numérotation arbitraire, $\mathscr{P}_n=\{A_{n1},A_{n2},\cdots\}$ une partition cubique de $\mathbf{R}^d$, les côtés de chaque cube étant de longueur $r_n>0$. Tout élément de $\mathscr{P}_n$ est donc du type :

$$[a_1,a_1+r_n[\times\cdots\times[a_d,a_d+r_n[,\ a_1,\cdots,a_d\in\mathbf{R}.$$

La règle d'apprentissage par histogramme $g_n^{(1)}$ est définie pour $x\in\mathbf{R}^d$ par

$$g_n^{(1)}(x)=\begin{cases}1 & \text{si } \sum_{i=1}^n\mathbf{1}_{\{Y_i=0,X_i\in A_n(x)\}}<\sum_{i=1}^n\mathbf{1}_{\{Y_i=1,X_i\in A_n(x)\}}\ ;\\ 0 & \text{sinon,}\end{cases}$$

où $A_n(x)$ désigne l'élément de $\mathscr{P}_n$ qui contient x. Le nombre d'observations de l'échantillon d'apprentissage appartenant au même élément de la partition $\mathscr{P}_n$ est noté $N_n(x)$, c'est-à-dire que $N_n(x)=\mathrm{Card}\{i:X_i\in A_n(x)\}$. Le lemme suivant est admis.

Lemme 1. *Si $nr_n^d\to\infty$, alors $N_n(X)\to\infty$ en probabilité.*

Théorème 1. *Si $r_n\to 0$ et $nr_n^d\to\infty$, alors la règle $g_n^{(1)}$ est convergente.*

Preuve. D'après la proposition 2, il suffit de montrer que $\mathbb{E}|\hat{\eta}_n(X)-\eta(X)|\to 0$, où

$$\hat{\eta}_n(x)=\frac{1}{N_n(x)}\sum_{i=1}^n Y_i\mathbf{1}_{\{X_i\in A_n(x)\}},$$

si $x\in\mathbf{R}^d$ est tel que $N_n(x)\neq 0$. Conditionnellement à $\mathbf{1}_{\{X_1\in A_n(x)\}},\cdots,\mathbf{1}_{\{X_n\in A_n(x)\}}$, la variable aléatoire $N_n(x)\hat{\eta}_n(x)$ suit une loi binomiale de paramètres $(N_n(x),\overline{\eta}_n(x))$, où on a noté $\overline{\eta}_n(x)=\mathbb{E}[Y|X\in A_n(x)]=\mathbb{E}[\eta(X)|X\in A_n(x)]$. On en déduit du lemme 1 que $\mathbb{E}|\hat{\eta}_n(X)-\overline{\eta}_n(X)|\to 0$,

car pour chaque $s > 0$,

$$\begin{aligned}\mathbb{E}|\hat{\eta}_n(X) - \overline{\eta}_n(X)| &\leq \mathbb{E}\Big(\frac{1}{\sqrt{N_n(X)}}\mathbf{1}_{\{N_n(X)>0\}}\Big) + \mathbb{P}(N_n(X) = 0) \\ &\leq \mathbb{P}(N_n(X) \leq s) + \frac{1}{\sqrt{s}} + \mathbb{P}(N_n(X) = 0).\end{aligned}$$

Il reste à montrer que $\mathbb{E}|\overline{\eta}_n(X) - \eta(X)| \to 0$. Soit $\varepsilon > 0$. Comme η est bornée, il existe une fonction continue à support compact η_ε telle que $\mathbb{E}|\eta_\varepsilon(X) - \eta(X)| \leq \varepsilon$. En adoptant la notation $\overline{\eta}_{\varepsilon,n}(x) = \mathbb{E}[\eta_\varepsilon(X)|X \in A_n(x)]$, on remarque que

$$\mathbb{E}|\overline{\eta}_n(X) - \overline{\eta}_{\varepsilon,n}(X)| \leq \mathbb{E}|\eta(X) - \eta_\varepsilon(X)| \leq \varepsilon.$$

Or, comme on a aussi $\mathbb{E}|\overline{\eta}_{\varepsilon,n}(X) - \eta_\varepsilon(X)| \leq \varepsilon$ si r_n est assez petit, le théorème est démontré. □

Cette méthode d'apprentissage par histogramme présente néanmoins un inconvénient majeur, dû au fait qu'elle peut affecter des labels différents à des observations pourtant proches. Le processus de décision peut aussi être amélioré en affectant plus de poids aux points les plus proches de chaque élément de l'échantillon. La règle du noyau est élaborée dans ce sens : soit $g_n^{(2)}$ la règle définie par

$$g_n^{(2)}(x) = \begin{cases} 1 & \text{si } \sum_{i=1}^n \mathbf{1}_{\{Y_i=0\}} K\big(\frac{x-X_i}{h_n}\big) < \sum_{i=1}^n \mathbf{1}_{\{Y_i=1\}} K\big(\frac{x-X_i}{h_n}\big) ; \\ 0 & \text{sinon,} \end{cases}$$

où $K : \mathbf{R}^d \to \mathbf{R}$ (le noyau) est une fonction intégrable et $h_n > 0$. En pratique, le choix du noyau est essentiel en tant que facteur de pondération. Cependant, un noyau quelconque n'induira pas nécessairement une règle convergente. On introduit alors la classe des noyaux réguliers, qui formeront une famille raisonnable pour la théorie et la pratique. On dit qu'un noyau K est régulier si il est à valeurs positives et si il existe $b > 0$ et une boule euclidienne non vide B dans $\mathbf{R}^d$ centrée à l'origine tels que $K \geq b\mathbf{1}_B$ et

$$\int_{\mathbf{R}^d} \sup_{y \in B} K(x+y)\mathrm{d}x < \infty.$$

A titre d'exemples de noyaux réguliers, on peut citer ($\|.\|$ désigne la norme euclidienne sur $\mathbf{R}^d$) le noyau gaussien $K(x) = \exp(-\|x\|^2)$, le noyau de Cauchy $K(x) = (1+\|x\|^2)^{-1}$, le noyau naïf $K(x) = \mathbf{1}_{\{\|x\| \leq 1\}}$ et le noyau d'Epanechnikov $K(x) = (1-\|x\|^2)\mathbf{1}_{\{\|x\| \leq 1\}}$.

La preuve du théorème ci-dessous est admise.

Théorème 2. *Soit K un noyau régulier. Si $h_n \to 0$ et $nh_n^d \to \infty$, alors la règle $g_n^{(2)}$ est convergente.*

Suggestions pour le développement

► *Soulignons qu'il s'agit d'un menu à la carte et que vous pouvez choisir d'étudier certains points, pas tous, pas nécessairement dans l'ordre, et de façon plus ou moins fouillée. Vous pouvez aussi vous poser d'autres questions que celles indiquées plus bas. Il est très vivement souhaité que vos investigations comportent une partie traitée sur ordinateur et, si possible, des représentations graphiques de vos résultats.*

– *Etude de la modélisation.* Vous pouvez critiquer la modélisation, notamment le cadre probabiliste qui est adopté. Vous pouvez aussi préciser pourquoi, en général, il n'est pas raisonnable de considérer que Y est une fonction de la seule variable aléatoire X.
– *Développements mathématiques.* Vous pouvez détailler les preuves de certains résultats du texte. Par exemple, vous pouvez détailler les preuves des propositions 1 et 2, du théorème 1, mais aussi les calculs de $L(g)$ donnés en introduction.
– *Etude numérique.* Vous pouvez prendre pour loi de Y une loi de Bernoulli de paramètre p à fixer, et des lois à densité pour X sachant $Y = 0$ et X sachant $Y = 1$ (par exemple des densités dans $\mathbf{R}^2$). L'échantillon d'apprentissage (de taille raisonnable, par exemple $n = 50$) est alors simulé suivant la loi ainsi choisie (vous pouvez considérer que le nombre d'observations de label 1 est le plus petit entier le plus proche de np). Vous pouvez ensuite développer un ou plusieurs des points suivants :

  • Représentez graphiquement l'échantillon d'apprentissage en adoptant la convention suivante : un carré noir (*resp.* blanc) représente une observation de label 1 (*resp.* 0). Représentez sur le même graphique une nouvelle observation X de label Y. Suivant le principe de décision selon la majorité, quel label affecter à X ?

  • Pour la règle de l'histogramme et/ou la règle du noyau : estimez la plus petite probabilité d'erreur ; cette estimation est-elle sensible au choix du paramètre (r_n ou h_n, selon la méthode) ? ; au choix du noyau K (dans le cas de la règle du noyau) ? Vous pouvez choisir pour paramètre "optimal" un paramètre qui minimise, en r_n ou h_n selon la règle, la fonction

$$\sum_{i=1}^{n} \mathbf{1}_{\{g_n(X_i)\neq Y_i\}},$$

  g_n étant la règle de décision considérée. Discutez la pertinence de ce critère.

  • Comparez les décisions prises, pour une nouvelle observation, par les 2 méthodes d'apprentissage. Représentez graphiquement le résultat (en adoptant par exemple la méthode graphique décrite ci-dessus). Quelle méthode permet de retrouver le plus souvent le bon label ? L'une des méthodes fournit-elle des estimateurs de faible variance ? (simulez un nombre suffisant d'échantillons d'apprentissage).