Apprentissage Statistique de la Classification : Partie II.

Benoît Cadre

ENS Cachan - Antenne de Bretagne

benoit.cadre@bretagne.ens-cachan.fr

Dans ce deuxième exposé consacré à l'apprentissage statistique de la classification, on se fixe pour objectifs de présenter quelques-uns des outils mathématiques utilisés afin de valider les méthodes d'apprentissage, et d'expliquer pourquoi, selon le contexte, certaines méthodes d'apprentissage fonctionnent alors que d'autres ne fonctionnent pas.

Pour de plus amples informations sur la théorie de l'apprentissage statistique, on pourra consulter les ouvrages de Vapnik (1995), de Devroye, Györfi et Lugosi (1996), ou de Hastie, Tibshirani et Friedman (2001).

1. La theorie de Vapnik et Chervonenkis

1.1 Les enjeux

Comme souvent en statistique, cette histoire débute avec des *observations* $x_1, \cdots, x_n$ de $\mathbb{R}^d$. La modélisation consiste à considérer que ce sont des *réalisations* de n variables aléatoires $X_1, \cdots, X_n$, supposées indépendantes et de même loi μ, appelée *mesure théorique*.

Le théorème de Varadarajan affirme que la *mesure empirique* μ_n définie par

$$\mu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i},$$

où δ_x est la mesure de Dirac en x, est, pour les grandes valeurs de n, une bonne approximation de μ, au sens où la probabilité que μ_n converge vers μ au sens de la topologie de la convergence étroite (lorsque $n \to \infty$) vaut 1. Ce résultat, considéré comme l'un des piliers de la statistique, n'est cependant pas suffisant en pratique, notamment car il ne donne aucune information sur l'erreur commise en utilisant la mesure empirique plutôt que la mesure théorique. En clair, pour une famille donnée $\mathcal{A}$ de boréliens, il s'agit de dégager des propriétés sur la quantité $\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|$. Pour quel type de classe $\mathcal{A}$ y-a-t'il convergence et, si la convergence est assurée, en quel sens et à quelle vitesse ? Les réponses à ces questions sont évidemment très liées à la richesse de la classe $\mathcal{A}$.

1.2 La dimension de Vapnik-Chervonenkis

Soient $z_1, \cdots, z_n \in \mathbb{R}^d$ et $N_{\mathcal{A}}(z_1, \cdots, z_n)$ le nombre de sous-ensembles de $\{z_1, \cdots, z_n\}$ obtenus en l'intersectant avec les éléments de $\mathcal{A}$:

$$N_{\mathcal{A}}(z_1, \cdots, z_n) = \text{card}\{\{z_1, \cdots, z_n\} \cap A \, : \, A \in \mathcal{A}\}.$$

Ainsi, on a toujours $N_{\mathcal{A}}(z_1, \cdots, z_n) \leq 2^n$ et, en cas d'égalité, on dit que $\mathcal{A}$ *pulvérise* $\{z_1, \cdots, z_n\}$. C'est le cas d'une classe $\mathcal{A}$ très riche, en fait trop riche ...

Le *coefficient de pulvérisation* de $\mathcal{A}$ à l'ordre $n \geq 1$, noté $S_{\mathcal{A}}(n)$, est alors défini par :

$$S_{\mathcal{A}}(n) = \max_{z_1, \cdots, z_n} N_{\mathcal{A}}(z_1, \cdots, z_n).$$

A titre d'exemples, si $\mathcal{A}$ est l'ensemble des intervalles de $\mathbb{R}$, $S_{\mathcal{A}}(n) = 1 + n(n+1)/2$, et si $\mathcal{A}$ est l'ensemble des convexes du plan : $S_{\mathcal{A}}(n) = 2^n$.

Dans le premier exemple, le coefficient de pulvérisation de $\mathcal{A}$ passe en dessous de la courbe maximale dès $n = 3$, alors que dans le deuxième exemple, le coefficient de pulvérisation reste sur la courbe maximale. De manière générale, on peut observer le fait suivant : si, pour un $n_0 \geq 1$, $S_{\mathcal{A}}(n_0) < 2^{n_0}$ alors $S_{\mathcal{A}}(n) < 2^n$ pour tout $n \geq n_0$. Cette observation permet souvent de calculer facilement la *dimension de Vapnik-Chervonenkis* $V_{\mathcal{A}}$ de la classe $\mathcal{A}$, définie par :

$$V_{\mathcal{A}} = \max\{n \geq 1 \, : \, S_{\mathcal{A}}(n) = 2^n\}.$$

1.3 L'inegalite de Vapnik-Chervonenkis

A la suite des travaux pionniers de Vapnik, bon nombre de statisticiens ont orienté leurs recherches vers des *inégalités de concentration* de la mesure empirique. Voici l'une des inégalités de *concentration moyenne* obtenue :

$$\mathbb{E} \sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \leq c\sqrt{\frac{V_{\mathcal{A}}}{n}},$$

pour une certaine constante universelle $c > 0$.

2. Apprentissage statistique de la classification supervisee en dimension finie

2.1 Le probleme de la classification supervisee

Soient $x_1, \cdots, x_n$ des observations de $\mathbb{R}^d$, par exemples des caractéristiques médicales de n patients. Un médecin -*l'expert*- rend, pour chaque patient i, le diagnostic -ou *label*- $y_i = 1$ si le patient est atteint par une maladie donnée et, dans le cas contraire $y_i = 0$. L'objectif est *d'apprendre* une règle de classification qui donne les associations $x_1 \to y_1 \cdots, x_n \to y_n$, en vue de l'appliquer à un $n+1$-ème patient.

On modélise le problème de la manière habituelle : *l'échantillon d'apprentissage* est une suite de variables aléatoires indépendantes et de même loi $(X_1, Y_1), \cdots, (X_n, Y_n)$, dont $(x_1, y_1), \cdots, (x_n, y_n)$ sont des réalisations.

Une règle de classification est une fonction $g : \mathbb{R}^d \to \{0, 1\}$. Le *risque théorique* de cette règle, c'est-à-dire la probabilité d'erreur de la règle, est $L(g) = \mathbb{P}(g(X) \neq Y)$, (X, Y) désignant un vecteur aléatoire de même loi que (X_1, Y_1). La fonction de risque admet un minimum : le *risque de Bayes*, noté $L^\star$, qui dépend de la loi -inconnue- du couple (X, Y).

L'enjeu est maintenant de construire, à partir de l'échantillon d'apprentissage, une règle de classification $\hat{g}$ dont le risque moyen, par exemple, est proche du risque de Bayes :

$$\mathbb{E}L(\hat{g}) \to L^\star, \text{ si } n \to \infty.$$

2.2 Des regles classiques

De telles règles d'apprentissage existent : la règle du noyau, des plus proches voisins, de l'histogramme, ... Elles sont toutes basées sur le principe très démocratique suivant : on affecte le label 1 (resp. 0) à la nouvelle observation si elle se situe à proximité d'une majorité d'observations de label 1 (resp. 0). Toute la difficulté réside dans le sens que l'on donne au terme "proximité" ...

Cependant, ces méthodes ont un inconvénient majeur et incontournable que l'on retrouve dans toute méthode de statistique non paramétrique, le *fléau de la dimension* : plus la dimension d de l'espace des observations est élevée, moins la vitesse de convergence de $\mathbb{E}L(\hat{g})$ vers $L^\star$ est rapide. Il faut donc envisager de nouvelles stratégies de classification.

2.3 Retour vers la theorie de Vapnik et Chervonenkis

L'idée introduite par Vapnik est la suivante : on considère un sous-ensemble $\mathcal{G}$ de l'ensemble des règles de décision, suffisamment riche pour que

$$\inf_{g \in \mathcal{G}} L(g) \approx L^\star,$$

mais pas trop riche non plus, afin de ne pas avoir à faire face au fléau de la dimension. La règle de classification empirique $\hat{g}$ est alors choisie parmi les règles qui minimisent le *risque empirique*

$$L_n(g) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{g(X_i) \neq Y_i\}}$$

dans la classe $\mathcal{G}$, i.e. $L_n(\hat{g}) \leq L_n(g), \quad \forall g \in \mathcal{G}$. On obtient avec l'inégalité de Vapnik-Chervonenkis décrite plus haut :

$$\mathbb{E}L(\hat{g}) - \inf_{g \in \mathcal{G}} L(g) \leq c\sqrt{\frac{V_{\mathcal{A}}}{n}},$$

$\mathcal{A}$ désignant une famille d'ensembles construite à partir de $\mathcal{G}$. Autrement dit, si la dimension VC de $\mathcal{A}$ est finie, la règle empirique $\hat{g}$ a, à un facteur proportionnel à $1/\sqrt{n}$ près, la même performance qu'une règle théorique optimale de la classe $\mathcal{G}$.

3. Apprentissage statistique de la classification supervisee en dimension infinie

Il est assez fréquent que les observations $x_1, \cdots, x_n$ de la section 2 soient des courbes, par exemple, en reprenant l'exemple médical de cette section, des électro-cardiogrammes, des courbes d'évolution de la température corporelle,... Les variables aléatoires $X_1, \cdots, X_n$ sont alors *fonctionnelles*, et l'espace des observations est de dimension infinie.

Dans un premier temps, il est tentant de reprendre la construction des règles classiques en dimension finie (règle du noyau, des plus proches voisins,...), et de l'adapter à ce contexte fonctionnel. Or, ces règles ne sont pas convergentes en général. L'une des raisons principales est que le théorème de différentiation de Lebesgue, énoncé en dimension finie, n'est plus vrai en dimension infinie.

Dans un second temps, on peut reprendre, au profit de ce nouveau contexte, la méthode utilisant la théorie de Vapnik et Chervonenkis, bien que celle-ci soit plutôt élaborée pour les espaces de dimension finie. Du point de vue de la modélisation, il est raisonnable de supposer que l'espace des observations est un espace de Hilbert séparable. Une méthode possible d'apprentissage est définie comme suit : on projette les courbes observées sur l'espace vectoriel engendré par les d premiers éléments d'une base hilbertienne puis on construit, comme dans la section 2, une règle de classification empirique à partir de l'échantillon projeté et enfin, on fait tendre d vers l'infini. En pratique, cette méthode, dont on peut analyser la performance, donne des résultats relativement satisfaisants. Cependant, son périmètre d'utilisation n'est pas totalement connu à l'heure actuelle ...

Références

Devroye, L., Györfi, L. et Lugosi, G. (1996) *A Probabilistic Theory of Pattern Recognition*, Springer.
Hastie, T., Tibshirani, T. et Friedman, J. (2001) *Elements of Statistical Learning*, Springer.
Vapnik, V.N., (1995) *The Nature of Statistical Learning Theory*, Springer.