Théorie statistique de l'apprentissage

Olivier CATONI*

Discipline inventée par Vladimir Vapnik ou évolution de l'inférence statistique traditionnelle vers l'analyse de données complexes, la théorie statistique de l'apprentissage se trouve au carrefour de différentes approches, qui touchent aux statistiques, bien entendu, mais aussi à la théorie de l'information ou à la mécanique statistique.

Préliminaires

Le développement des moyens informatiques, des télécommunications et des capteurs électroniques de toutes natures provoque la production et le stockage de données de plus en plus abondantes et de plus en plus complexes. L'exploitation de ces données par des moyens humains devient en conséquence de moins en moins facile, créant un besoin urgent de mettre au point des méthodes automatiques d'analyse permettant de confier à une machine des fonctions de perception, de discrimination et de reconnaissance qui étaient jusque là l'apanage du cerveau humain (ou animal pour bon nombre d'entre elles). Cette recherche s'est avérée plus malaisée que les premiers espoirs de la cybernétique ne l'auraient laissé espérer dans l'immédiat après guerre. En effet, les processus de perception et de traitement réalisés par le cerveau sont mal connus, et leur caractère très largement inconscient ne permet pas de les appréhender par une approche introspective. Nous illustrerons nos propos par deux applications emblématiques, parce qu'elles correspondent aux tâches de perception que le cerveau effectue à jet continu dans la vie courante : la reconnaissance de la parole et la classification d'images (plus généralement l'interprétation de scènes visuelles). Des progrès dans ces domaines en feraient faire à une multitude d'applications dont il serait vain de tenter de faire le tour, comprenant bien entendu la robotique, la navigation assistée, mais aussi le diagnostic médical, la conduite automatisée de processus industriels, et d'une manière plus indirecte (parce que les problèmes d'apprentissage se heurtent à des difficultés génériques indépendantes de l'application envisagée), à l'analyse du génome, la constitution et l'interrogation de bases de données, l'analyse de la langue naturelle, etc.

Nous ne parlerons pas des capteurs, on trouve dans le commerce des caméscopes qui permettent d'enregistrer sur un ordinateur du son et des images de très bonne qualité. Nous ne parlerons pas des recherches faites pour comprendre le fonctionnement du cerveau, parce que nous ne sommes pas certain qu'un ordinateur pourrait le reproduire efficacement : en effet, un cerveau possède un très grand nombre de neurones très fortement interconnectés qui traitent en parallèle des informations à l'aide de processus électrochimiques très lents. Ceci contraste nettement avec l'organisation d'un ordinateur, qui possède une unité centrale (ou un faible nombre d'entre elles) qui traite à une vitesse très élevée des signaux électromagnétiques mais n'accède qu'à une seule donnée à la fois parmi celles qui sont rangées dans une mémoire par ailleurs immobile. De même qu'un muscle et un moteur réalisent en gros la même fonction (fournir du travail mécanique) par des moyens très différents, de même, rien ne permet de penser que la reproduction par une machine de certaines fonctions perceptives du cerveau doive utiliser des algorithmes qui

* *CNRS, Laboratoire de Probabilités, UMR 7599, Université Paris 6, case 188, 4, place Jussieu, F-75252 Paris Cedex 05, catoni@ccr.jussieu.fr*
http://www.proba.jussieu.fr/users/catoni/homepage/newpage.html

auraient une analogie quelconque avec le fonctionnement interne de celui-ci (notre point de vue paraîtra délibérément polémique à certains, nous espérons que cette prise de position marquée aura au moins le mérite de susciter réactions et réflexions sur la question).

Mises à part les tentatives d'analogie avec le fonctionnement cérébral tel que peuvent le décrire les neurosciences, les chercheurs ont pendant très longtemps essayé de reconnaître des sons ou des images en utilisant la méthode qui fait le succès de la physique depuis Descartes (ou peut-être depuis l'antiquité, nous ne nous prononcerons pas sur ces délicates questions d'histoire des sciences) : établir un modèle du phénomène observé, ici le son ou l'image numérisés, en estimer les paramètres à partir de mesures expérimentales, confronter les prédictions fournies par le modèle (concernant la nature des sons ou des images analysés) avec l'expérience. Cette approche n'a jamais fonctionné correctement pour traiter les problèmes qui nous intéressent ici, sauf dans des situations très simples : personne n'est capable de donner un modèle satisfaisant du bruit que fait une phrase prononcée dans une ambiance sonore quelconque, par un interlocuteur quelconque. Personne n'est non plus capable de donner un modèle de ce qu'enregistre une caméra placée sur le capot d'une voiture, … nous ne sommes pas en présence de phénomènes que l'on puisse décrire à l'aide d'un nombre raisonnable de paramètres et d'équations, comme on le fait des oscillations d'un pendule ou des mouvements d'une planète.

C'est sur cet échec que s'est construite la légitimité des approches statistiques. Bien que la reconnaissance des formes présente toutes sortes de difficultés annexes nous parlerons ici plus spécifiquement de la classification supervisée. Nous supposerons donc disposer d'une base de données $X_1,\ldots,X_N \in \mathcal{X}$ déjà classées et nommerons $Y_1,\ldots,Y_N \in \mathcal{Y}$ les classes correspondantes. L'ensemble $\mathcal{X}$ désignera l'espace mesurable dans lequel les données sont représentées. Dans le cas de la reconnaissance de visages, qui a servi de banc d'essai à de nombreuses méthodes, $X_1,\ldots,X_N$ seront des imagettes centrées soit sur des visages remis à une échelle normalisée, soit sur des contre exemples de ce que l'on rencontre ailleurs dans les images à traiter. Les classes $Y_1,\ldots,Y_N$ prendront alors deux valeurs, correspondant à la présence ou à l'absence de visage. La question posée par l'apprentissage statistique est la suivante : supposons que $X_1,\ldots,X_N$ aient été tirées au hasard parmi une grande « population » d'imagettes, comment choisir une règle de classification qui commette le moins d'erreurs possibles sur la population totale en n'utilisant pour construire cette règle que les exemples observés $X_1,\ldots,X_N$ (et les classes $Y_1,\ldots,Y_N$, supposées fournies par un expert, ou plus généralement tout autre moyen extérieur) ?

Il est important de comprendre que cette approche possède des avantages spécifiques avant d'en commenter les aspects techniques :

– on ne modélise pas les données à analyser, mais seulement une « expérience statistique » qui s'apparente à un sondage. Le seul aléa supposé est celui introduit par le statisticien dans le choix des exemples, les propriétés d'indépendance et d'équidistribution de l'échantillon $X_1,\ldots,\ X_N$ peuvent donc être garanties de façon réaliste ;

– à défaut de modéliser les données à analyser, on doit par contre modéliser les règles de classification qui vont leur être appliquées : ces règles étant construites par le statisticien, et leur complexité étant limitée par la puissance de calcul dont il dispose, cette opération de modélisation est réalisable en pratique. Elle consiste à « structurer » (le terme est de V. Vapnik) l'ensemble des règles dont le statisticien va tester les performances en une réunion de familles « paramétriques », c'est-à-dire de familles d'algorithmes identiques à la valeur d'un certain nombre de paramètres numériques près ;

– l'approche statistique réserve la possibilité de sélectionner des règles de classification différentes pour traiter des jeux de données différents. En ajustant ainsi au plus près le choix de la méthode aux données que l'on souhaite effectivement analyser, on évite d'essayer de résoudre un problème plus compliqué (c'est-à-dire ici plus générique) que nécessaire.

Il faut cependant avoir à l'esprit le fait que la constitution de la base de données sur laquelle va porter la méthode d'apprentissage statistique évoquée ci-dessus pose aussi des questions délicates. En particulier l'extraction et la normalisation des données, que ce soit dans la phase d'apprentissage (c'est-à-dire de sélection de la méthode) ou dans la phase de reconnaissance (c'est-à-dire au moment où on va appliquer la méthode sur des données brutes), est une étape cruciale pour la réussite de l'opération. Elle présente des obstacles tout aussi fondamentaux que l'exploitation de la base de données elle-même, dont les moindres ne sont pas les problèmes dits de « segmentation ». Dans le cas des visages, la segmentation consiste à positionner et à dimensionner convenablement l'imagette autour des zones pouvant contenir un visage. Ce n'est pas trop compliqué parce qu'un visage est un objet assez rigide qui s'inscrit de façon satisfaisante dans un cadre rectangulaire et qu'il y a « peu » de rectangles dans une image. La reconnaissance

Figure 1 – *Les éléments de bords sont des caractéristiques intermédiaire souvent employées en analyse d'images. Ce sont des éléments plus structurés et plus géométriques que les pixels. Ils possèdent une intensité (que l'on peut seuiller) et une orientation. On peut ensuite les regrouper par paquets suivant leurs positions et orientations relatives, et compter le nombre d'apparition de ces configurations dans les images (dans l'illustration ci-dessus, on a regroupé les orientations en trois classes, rouge, verte ou bleue). On obtient ainsi une famille très riche de mesures, à partir de laquelle on peut construire une famille encore plus riche de règles de classification (par exemple en séparant des groupes de mesures par des hyperplans). Les techniques décrites dans cette présentation ont pour but de sélectionner parmi toutes ces règles possibles, une règle dont le taux d'erreur à un niveau de confiance donné soit garanti par une inégalité mathématiquement prouvée.*

d'objets beaucoup plus déformables (un serpent, un chat …) ou partiellement occultés par d'autres poserait de vrais problèmes supplémentaires, qui restent à ce jour largement ouverts. De même, dans le domaine de la reconnaissance de la parole, il existe un saut très important entre la reconnaissance de mots isolés et la reconnaissance d'un discours continu dans lequel les mots s'enchaînent les uns aux autres sans silences permettant d'en identifier facilement les frontières.

L'approche PAC-Bayésienne

Ces remarques faites, venons-en à l'apprentissage d'une règle de classification à partir d'exemples classés $(X_1,Y_1),\ldots,(X_N,Y_N) \in (\mathcal{X} \times \mathcal{Y})$, formant une suite de couples de variables aléatoires indépendants identiquement distribués (i.i.d. en abrégé). Notons $\mathbb{P}$ la loi jointe inconnue de cette suite. Soit $\{f_\theta : \mathcal{X} \to \mathcal{Y};\ \theta \in \Theta\}$ l'ensemble de toutes les règles de classification qui seront envisagées pour classer les données. Comme expliqué dans les préliminaires, Θ se décomposera le plus souvent en une réunion de sous-ensembles de « dimensions » différentes. Un critère naturel, mais malheureusement inaccessible, pour juger de la qualité de la règle f_θ est fourni par son taux d'erreur moyen $R(\theta) = \mathbb{P}\big[f_\theta(X_1) \neq Y_1\big]$ (où l'indice 1 peut être remplacé par n'importe quel autre, l'échantillon étant supposé i.i.d.). Néanmoins ce taux d'erreur moyen est l'espérance d'une variable aléatoire observable, le taux d'erreur empirique (c'est-à-dire constaté sur l'échantillon observé) $r(\theta) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\big[f_\theta(X_i) \neq Y_i\big]$. Si nous ne considérions qu'une seule règle f_θ, le lien entre $r(\theta)$ et son espérance serait simple, $r(\theta)$ étant une moyenne de variables de Bernoulli i.i.d. Malheureusement, nous voulons considérer la question bien plus délicate des relations du processus $\theta \mapsto r(\theta)$, où θ varie dans un très grand ensemble Θ, avec le taux d'erreur moyen minimum $\inf_{\theta\in\Theta} R(\theta)$ et la valeur (ou les valeurs) du paramètre θ où il est atteint, soit $\arg\min_{\theta\in\Theta} R(\theta)$. Un phénomène bien mis en valeur par V. Vapnik dès le début de la théorie est celui du « sur-apprentissage » qui peut se décrire qualitativement ainsi : si Θ est « trop grand », les deux minima $\inf_{\theta} R(\theta)$ et $\inf_{\theta\in\Theta} r(\theta)$ peuvent n'avoir aucun lien entre eux, ni les valeurs du paramètre θ pour lesquelles ils sont atteints. Dans ce cas, on obtiendra un meilleur résultat en considérant la relation entre $\inf_{\theta\in\Theta_1} r(\theta)$ et $\inf_{\theta\in\Theta} R(\theta)$, où Θ_1 est un sous-ensemble de Θ de taille convenable. En restreignant Θ à Θ_1, on fait apparaître deux phénomènes antagonistes : un phénomène favorable de « réduction de la variance » qui va faire que $\arg\min_{\theta\in\Theta_1} r(\theta)$ va se rapprocher de plus en plus de $\arg\min_{\theta\in\Theta_1} R(\theta)$, et un phénomène défavorable de biais, qui va faire

que $\inf_{\theta\in\Theta_1} R(\theta)$ va s'éloigner de plus en plus de $\inf_{\theta\in\Theta} R(\theta)$. Pour trouver le meilleur compromis entre ces deux phénomènes, on est conduit à considérer toute une famille de sous-ensembles $(\Theta_j)_{j\in J}$ de Θ. La situation se complique donc : au lieu de chercher le meilleur paramètre θ, nous en sommes maintenant à chercher le meilleur sous-ensemble de paramètres où chercher le meilleur paramètre ! et là encore, une étape de modélisation s'impose : de même qu'il est désavantageux de chercher le meilleur θ dans un ensemble Θ trop grand, de même il est désavantageux de chercher le meilleur Θ_j dans un ensemble de parties de Θ trop grand (en particulier on voit tout de suite que cela n'avance à rien de pousser cette démarche jusqu'à l'extrême en considérant tous les singletons de Θ). Le choix d'une famille $(\Theta_j)_{j\in J}$ de sous-modèles de Θ a été baptisé par V. Vapnik « minimisation structurelle du risque ». La théorie des processus empiriques permet de quantifier les phénomènes que nous venons de décrire en faisant des hypothèses sur la structure du processus $\theta \mapsto r(\theta)$ pour une métrique qui majore les covariances, dans le domaine de la classification on considère souvent $D(\theta,\theta') = \mathbb{P}\big[f_\theta(X_1) \neq f_{\theta'}(X_1)\big]$ qui majore la variance de $\sqrt{N}\big[r(\theta) - r(\theta')\big]$. C'est la voie la plus « classique » d'approche de l'apprentissage statistique. Elle est néanmoins semée d'embûches, la moindre n'étant pas que la distance $D(\theta,\theta')$ n'est en pratique pas plus connue que le reste, et que des hypothèses portant sur le contrôle de l'entropie métrique des sous-espaces (Θ_j, D) sont difficiles à vérifier.

Nous présenterons ici une approche alternative, qui contrôle les quantités évoquées ci-dessus par des moyens détournés. Elle est née dans la communauté du « machine learning », sous l'impulsion séminale de D. McAllester qui l'a baptisée et en a prouvé les premiers résultats. Notons au passage que ce nom de baptême, « Probably Approximately Correct Bayesian theorems », est à comprendre dans une perspective purement historique : la façon de poser le problème que nous venons de décrire n'a rien de Bayésien, de plus l'approche inventée par D. McAllester fournit certes des inégalités de déviations (vérifiées avec probabilité $1-\epsilon$, d'où le préfixe « PAC »), mais peut aussi fournir directement des inégalités en espérance, comme nous allons l'évoquer ; en un mot son contenu n'a rien à voir avec son nom !

Le premier ingrédient de l'approche PAC-Bayésienne consiste à lisser l'étape de minimisation structurelle du risque : au lieu de considérer une famille de sous-modèles, $(\Theta_j)_{j\in J}$, nous allons considérer une mesure de probabilités π sur Θ. Cette mesure n'a pas d'interprétation probabiliste ! Elle peut être vue comme un moyen de spécifier partiellement une représentation des éléments de Θ en choisissant la longueur du code associé (du moins dans le cas où Θ est fini ou dénombrable, mais on peut toujours se ramener à ce cas en pratique en tronquant la représentation des variables réelles). Elle peut aussi être vue comme une sorte de pénalisation *a priori* des différentes parties de Θ.

Le deuxième ingrédient consiste à remplacer le contrôle des fluctuations du processus $\theta \mapsto r(\theta)$ par le contrôle d'une quantité bien connue des physiciens, l'énergie libre, plus sobrement pour les mathématiciens la transformée de Laplace, à savoir $\frac{1}{\lambda}\log\Big[\iint \exp\big[-\lambda r(\theta,\omega)\big]\pi(d\theta)\mathbb{P}(d\omega)\Big]$. En jouant sur le paramètre λ, on va pouvoir se rapprocher plus ou moins de $\inf_\theta r(\theta)$, et donc réaliser quelque chose qui ressemble au compromis sur la taille du modèle recherché par la minimisation structurelle du risque dont nous avons parlé plus haut. En utilisant un peu d'analyse convexe, on pourra alors contrôler le comportement de toutes les « lois *a posteriori* » $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$ possibles, c'est-à-dire de toutes les lois de probabilités sur les paramètres qui dépendent de l'échantillon observé (et sont de ce fait des mesures aléatoires, on supposera plus précisément sans le dire dans ce qui suit que ce sont des probabilités conditionnelles régulières). L'espace probabilisable Ω désigne ici celui sur lequel les variables aléatoires représentant l'échantillon sont construites, on peut en particulier choisir la représentation dite canonique de l'aléa dans laquelle $\Omega = (\mathcal{X}\times\mathcal{Y})^N$ et $(X_i,Y_i)_{i=1}^N(\omega) = \omega$). On pourra en effet utiliser l'identité remarquable :

$$\log\Big\{\int \exp\big[-\lambda r(\theta)\big]\pi(d\theta)\Big\} = \sup_{\rho\in\mathcal{M}_+^1(\Theta)} \lambda\rho\big[r(\theta)\big] - \mathcal{K}(\rho,\pi),$$

où $\mathcal{M}_+^1(\Theta)$ désigne l'ensemble des mesures de probabilités sur Θ et où $\mathcal{K}(\rho,\pi) = \int \log\left(\frac{\rho}{\pi}\right) d\rho$ désigne l'entropie relative de la loi ρ par rapport à la loi *a priori* π (quand ρ n'est pas absolument continue par rapport à π, on pose

$\mathcal{K}(\rho,\pi) = \infty$ par convention). On obtient ainsi facilement les bornes en déviations et en moyenne

$$\mathbb{P}\left\{ \sup_{\rho\in\mathcal{M}_+^1(\Theta)} \int R(\theta)\rho(d\theta) - \frac{\int r(\theta)\rho(d\theta) + \dfrac{\mathcal{K}(\rho,\pi) - \log(\epsilon)}{\lambda}}{1 - \dfrac{\lambda}{2N}} \leqslant 0 \right\} \geqslant 1-\epsilon,\ \lambda < 2N,$$

$$\text{et} \quad \iint R(\theta)\rho(\omega,d\theta)\mathbb{P}(d\omega) \leqslant \int \left[\frac{\int r(\theta,\omega)\rho(\omega,d\theta) + \dfrac{\mathcal{K}[\rho(\omega),\pi]}{\lambda}}{1 - \dfrac{\lambda}{2N}} \right] \mathbb{P}(d\omega),\ \ \rho : \Omega \to \mathcal{M}_+^1(\Theta), \lambda < 2N.$$

(dont nous ne donnons pas les formes les plus précises par souci de simplicité). Les bornes en espérance sont moins intéressantes du point de vue théorique, mais donnent d'un point de vue pratique des constantes plus serrées et sont souvent plus faciles à lire, même si elles ne fournissent que des majorations « sans biais » de l'erreur de généralisation moyenne, qui pourraient s'avérer sans intérêt si une borne en déviation ne permettait de prouver que leurs fluctuations ne sont pas trop grandes.

En travaillant un peu plus, on peut optimiser le paramètre λ dans la première inégalité pour obtenir avec $\mathbb{P}$ probabilité au moins $1-\epsilon$, pour toute loi *a posteriori* $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$,

$$\int R(\theta)\rho(\omega,d\theta) \leqslant \left(1 + \frac{2\alpha d}{N}\right)^{-1} \left\{ \int r(\theta,\omega)\rho(\omega,d\theta) + \frac{\alpha d}{N} + \sqrt{\frac{2\alpha d \int r\rho(d\theta)\left[1 - \int r\rho(d\theta)\right]}{N} + \frac{\alpha^2 d^2}{N^2}} \right\},$$

où α est un paramètre réel positif supérieur à 1, que l'on peut prendre par exemple égal à $1 + \left[\log(N)\right]^{-1}$, et où $d = \mathcal{K}\left[\rho(\omega),\pi\right] + \log\left(\dfrac{\log(2\alpha N)}{\epsilon\log(\alpha)}\right)$ est un « terme de complexité ».

Ces inégalités fournissent une première majoration du taux d'erreur moyen d'une règle de classification tirée au hasard suivant la loi *a posteriori* $\rho(\omega,d\theta)$, par une borne qui a le mérite d'être observable, et le défaut d'être infinie pour les masses de Dirac (tout au moins quand π est une mesure diffuse). C'est le prix à payer, semble-t-il, pour obtenir des bornes sans faire d'hypothèses contraignantes sur la structure de (Θ, D). Sous des hypothèses de structure, on pourrait alors montrer que pour $\widehat{\theta}(\omega)$ et $\rho(\omega,d\theta)$ bien choisis $\int D\left[\theta,\widehat{\theta}(\omega)\right]\rho(\omega,d\theta)$ est petit et donc que $R\left[\widehat{\theta}(\omega)\right] \leqslant \int R(\theta)\rho(d\theta) + \int D\left(\theta,\widehat{\theta}\right)\rho(d\theta)$ l'est aussi.

Ces premiers théorèmes PAC Bayésiens peuvent être améliorés d'au moins deux façons : d'une part en jouant sur un choix spécifique de π relié à celui de ρ, menant à des bornes plus « locales », d'autre part en utilisant la structure des covariances du processus $\theta \mapsto r(\theta)$ au lieu d'utiliser la variance de $r(\theta)$. Cependant ces améliorations vont se faire au détriment de la valeur des constantes, si bien qu'elles n'en seront vraiment que pour des valeurs suffisamment grandes de la taille N de l'échantillon. Pour cette raison, les bornes les plus simples gardent tout leur intérêt, en dépit du fait qu'elles ne soient pas asymptotiquement optimales quand N tend vers l'infini.

Localisation

On voit facilement que le choix optimal de la loi *a priori* π dans la borne en espérance est $\pi = \int \rho(\omega)\mathbb{P}(d\omega)$. Malheureusement cette probabilité *a priori* sur les paramètres n'est pas observable (puisque $\mathbb{P}$ est inconnue). Notons qu'elle donne un renseignement intéressant sur le plan théorique : $\int \mathcal{K}\left[\rho(\omega),\pi\right]\mathbb{P}(d\omega)$ est alors égale à *l'information*

mutuelle entre ω (qui représente ici l'échantillon observé) et θ, lorsque ω est tiré suivant $\mathbb{P}$ et θ est tiré suivant $\rho(\omega,d\theta)$ une fois ω choisi. Ainsi, l'écart entre l'erreur de généralisation d'une règle de classification randomisée et l'erreur constatée sur l'échantillon observé est contrôlé par l'information mutuelle entre l'échantillon et le paramètre. En pratique on est tenu de choisir π indépendamment de $\mathbb{P}$ et ce que l'on perd est quantifié par l'identité $\int \mathcal{K}(\rho(\omega),\pi)\mathbb{P}(d\omega) = \int \mathcal{K}\big[\rho(\omega), \int \rho(\omega')\mathbb{P}(d\omega')\big]\mathbb{P}(d\omega) + \mathcal{K}\big[\int \rho(\omega')\mathbb{P}(d\omega'), \pi\big]$. On peut néanmoins aller plus loin de la façon suivante : quand π et λ sont fixés, la loi *a posteriori* optimale (c'est-à-dire qui minimise la borne) a pour densité $\dfrac{d\rho}{d\pi} = \dfrac{\exp\big[-\lambda r(\theta)\big]}{\int \exp\big[-\lambda r(\theta')\big]\pi(d\theta')}$. On la notera $\pi_{\exp(-\lambda r)}$. On peut alors revenir sur le choix de la loi *a priori* et la prendre de la forme $\pi_{\exp(-\beta R)}$. En travaillant un peu sur le lien entre $\pi_{\exp(-\beta R)}$ et sa version empirique $\pi_{\exp(-\beta r)}$, on parvient alors à prouver la borne en espérance

$$\int \left\{ \int r(\theta,\omega)\rho(\omega,d\theta) - \frac{\mathcal{K}\big[\rho,\pi_{\exp(-\beta r)}\big]}{\beta} \right\} \mathbb{P}(d\omega) \leqslant \int\int R(\theta)\rho(\omega,d\theta)\mathbb{P}(d\omega)$$

$$\leqslant \int \left\{ \frac{\int r(\theta,\omega)\rho(\omega,d\theta) + \dfrac{\mathcal{K}\big[\rho,\pi_{\exp(-\beta r)}\big]}{\beta}}{1 - \dfrac{2\beta}{N}} \right\} \mathbb{P}(d\omega)\,,$$

Une inégalité de déviation du même type peut aussi être prouvée. Le cas $\rho = \pi_{\exp(-\beta r)}$ est particulièrement intéressant : les termes d'entropie disparaissent, montrant ainsi que cette « loi de Gibbs » (comme diraient les physiciens) *a posteriori* ne souffre pas de sur-apprentissage : à une constante universelle près, elle a la même performance en espérance et sur l'échantillon observé. De plus la borne inférieure montre que l'encadrement est optimal à un facteur $\left(1 - \frac{2\beta}{N}\right)^{-1}$ près.

Bornes relatives

Une autre amélioration consiste à contrôler $r(\theta) - r(\widetilde{\theta})$, où $\widetilde{\theta}$ est une valeur inconnue du paramètre, par exemple $\arg\min_{\theta\in\Theta_1} R(\theta)$, où Θ_1 est une partie de Θ. On ne contrôle alors pas $R(\theta)$, mais uniquement $R(\theta) - R(\widetilde{\theta})$: dans certaines circonstances, on saura ainsi que l'on se trouve très près du taux d'erreur optimum dans le modèle de classification choisi, sans savoir avec une aussi grande précision quel est ce taux ! Cela se produira par exemple dans le cas binaire bruité où $|\mathcal{Y}| = 2$ et où $P(Y_i = f_{\widetilde{\theta}}(X_i) \mid X_i) = 1 - \alpha$, quand $0 < \alpha < 1/2$.

Voici un exemple d'inégalité en moyenne (une inégalité de déviation de même type est aussi disponible). Considérons une partie Θ_1 de Θ (qui peut éventuellement être égale à Θ tout entier), $\widetilde{\theta} \in \arg\min_{\theta\in\Theta_1} R(\theta)$,

$$\widehat{\theta} \in \arg\min_{\theta\in\Theta} r(\theta),$$

$$d(\theta,\theta') = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\big[f_\theta(X_i) \neq f_{\theta'}(X_i)\big],$$

$$\psi(a) = \sup_{\theta\in\Theta_1} d(\theta,\widehat{\theta}) - a\big[r(\theta) - r(\widehat{\theta})\big],$$

et $g(a) = 2a^{-2}[\exp(a) - 1 - a]$, $a \in \mathbb{R}_+$. Pour tous paramètres réels β et λ tels que $0 \leqslant \beta < \lambda$, toute loi *a posteriori* $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$,

$$\int\int R(\theta)\rho(\omega,d\theta)\mathbb{P}(d\omega) \leqslant R(\widetilde{\theta}) + \int \left\{ \frac{\mathcal{K}\left[\rho(\omega),\pi_{\exp(-\lambda r)}\right]}{\lambda-\beta} + \inf_{a,0\leqslant a\leqslant \frac{2N(\lambda-\beta)}{g\left(\frac{2\lambda}{N}\right)\lambda^2}} \frac{g\left(\frac{2\lambda}{N}\right)\lambda^2 \quad (a)}{N(\lambda-\beta)} \right.$$

$$\left. + \left(1 + \frac{g\left(\frac{2\lambda}{N}\right)\lambda^2 a}{2N(\lambda-\beta)}\right)\left[\int r(\omega)\pi_{\exp\left[-(\beta-\frac{g\left(\frac{2\lambda}{N}\right)\lambda^2 a}{2N})r\right]}(d\theta) - r(\widehat{\theta})\right]\right\}\mathbb{P}(d\omega).$$

Cette inégalité fournit une « borne empirique sans biais » permettant de comparer le taux d'erreur moyen de ρ avec celui de la meilleure règle (inconnue) dans Θ_1. On dispose aussi d'une borne théorique correspondante, dans laquelle d est remplacée par D et ψ par $\varphi(a) = \sup_{\theta\in\Theta_1} D(\theta,\widetilde{\theta}) - a\left[R(\theta) - R(\widetilde{\theta})\right]$. Plus précisément

$$\int\int R(\theta)\rho(\omega,d\theta)\mathbb{P}(d\omega) \leqslant R(\widetilde{\theta}) + \inf_{a,0\leqslant a\leqslant \frac{2N(\lambda-\beta)}{g\left(\frac{2\lambda}{N}\right)\lambda^2}} \left(1 - \frac{g\left(\frac{2\lambda}{N}\right)\lambda^2 a}{2N(\lambda-\beta)}\right)^{-1}$$

$$\times \left\{ \frac{\int_\beta^\lambda \left[\int R(\theta)\pi_{\exp(-\gamma R)}(d\theta) - R(\widetilde{\theta})\right] d\gamma}{\lambda-\beta} + \frac{g\left(\frac{2\lambda}{N}\right)\lambda^2\varphi(a)}{2N(\lambda-\beta)} + \frac{\int \mathcal{K}\left[\rho(\omega),\pi_{\exp(-\lambda r)}\right]\mathbb{P}(d\omega)}{\lambda-\beta} \right\}.$$

Cette borne montre que la loi de Gibbs *a posteriori* $\pi_{\exp(-\lambda r)}$ a un taux d'erreur moyen qui peut atteindre dans certains cas des vitesses de convergence vers $\inf_\Theta R$ supérieures à $\sqrt{\frac{R(\widetilde{\theta})}{N}}$ (par exemple dans le cas binaire bruité évoqué plus haut, où $\varphi\left[(1-2\alpha)^{-1}\right] = 0$, la convergence est en $1/N$ dès que $\int R(\theta)\pi_{\exp(-\gamma R)}(d\theta) \leqslant \inf_{\theta\in\Theta} R(\theta) + \frac{c}{N}$, où c est une constante réelle positive).

Echantillon fantôme et bornes de Vapnik

En introduisant un échantillon fantôme $(X_{N+1}, Y_{N+1},\ldots,X_{(k+1)N},Y_{(k+1)N})$, et en utilisant l'échangeabilité de la loi jointe de l'échantillon total $(X_1,Y_1,\ldots,X_{(k+1)N},Y_{(k+1)N})$, on peut montrer des inégalités similaires aux précédentes, dans lesquelles la loi *a priori* π, au lieu d'être fixe, a le droit de dépendre du « design » $(X_1,\ldots,X_{(k+1)N})$, pourvu que cette dépendance soit invariante par permutation des indices. On voit alors, dans cette approche, que seules comptent les restrictions $f_\theta : \{X_i, 1\leqslant i\leqslant (k+1)N\} \to \mathcal{Y}$ des règles de classification à $(k+1)N$ données. Ces restrictions sont en nombre fini (puisque $\mathcal{Y}$ est supposé être un ensemble fini de classes), au plus égal à $|\mathcal{Y}|^{(k+1)N}$. En fait, elles sont souvent bien moins nombreuses, par exemple dans le cas de la classification binaire, quand la famille de règles a une dimension de Vapnik Cervonenkis (dont nous n'avons pas la place de donner ici la définition) inférieure à h, le nombre de règles est inférieur à $\left(\frac{e(k+1)N}{h}\right)^h$ (c'est le cas par exemple de l'ensemble des règles obtenues en séparant $\mathbb{R}^d$ par des hyperplans affines, qui a pour dimension de Vapnik Cervonenkis $h = d+1$). En choisissant π uniforme sur ces règles réduites à l'échantillon total, on ramène le terme d'entropie $\mathcal{K}(\rho,\pi)$ à une valeur maximale de $h\log\left(\frac{(k+1)eN}{h}\right)$ dans les inégalités exposées ci-dessus, y compris lorsque l'on

prend comme loi *a posteriori* $\rho(\omega, d\theta)$ une masse de Dirac. De plus les versions localisées des bornes permettent de réduire le terme d'entropie, voire de l'annuler dans le cas particulier où on considère $\pi_{\exp(-\beta r)}$ comme loi *a posteriori*. Dans ce cadre, les supports vector machines de Vapnik offrent un modèle de classification très intéressant : il consiste, dans le cas binaire, à séparer les données par un hyperplan dans un espace de Hilbert « virtuel » que l'on manipule uniquement à l'aide du « noyau » $K(X_i, X_j)$ qui donne le produit scalaire entre X_i et X_j dans l'espace transformé. Il suffit en fait que la matrice $m_{i,j} = K(X_i, X_j)$ soit symétrique positive pour qu'une telle représentation dans un Hilbert existe, ce qui permet un grand choix de noyaux. En particulier, quand les formes sont représentées initialement dans $\mathbb{R}^d$, on choisit souvent un noyau exponentiel $K(X_i, X_j) = \exp(-\gamma||X_i - X_j||^2)$ qui possède la propriété intéressante d'envoyer les X_i sur des points de la sphère linéairement indépendants les uns des autres dans l'espace transformé. On peut alors séparer dans l'espace transformé les X_i, $1 \leqslant i \leqslant (k+1)N$ de *toutes les manières possibles* : on a fabriqué une représentation linéaire de toutes les règles de classification possibles de l'échantillon total. On peut les ranger en fonction de leur marge, la distance entre l'hyperplan séparateur et le nuage des points transformés des X_i, en pondérant plus fortement sous π les règles de plus forte marge (dans l'approche PAC-Bayesienne). Parmi tous les hyperplans qui séparent les X_i de la même façon on choisira dans cette approche « un hyperplan canonique », c'est-à-dire de marge maximum. Il se trouve que le calcul de cet hyperplan ne fait intervenir que les points placés à distance minimum de l'hyperplan séparateur, appelés *vecteurs de support*. Une autre façon de structurer les modèles consiste à s'appuyer sur le nombre de vecteurs de support. Les support vector machines apparaissent alors comme un cas particulier des « schémas de compression » de Littlestone et Warmuth. En effet, les règles dont la définition ne dépend que de la valeur de h données sur $(k+1)N$ sont au plus au nombre de $\binom{(k+1)N}{h} \leqslant \left(\frac{(k+1)eN}{h}\right)^h$, ce qui permet un contrôle des termes de complexité dans les inégalités dans lequel h joue un rôle similaire à celui de la dimension de Vapnik Cervonenkis. Une règle de classification qui ne dépend que de h données s'appelle un schéma de compression. De tels modèles de règles peuvent être construits de façons extrêmement variées et intuitives. Il suffit pour cela de se poser la question : comment ferais-je pour classer h données ? C'est ensuite cet ensemble restreint de h exemples (appelé ensemble de compression) qui vient paramétrer la famille de règles ainsi construite. En fait, on voit que l'on peut de cette façon construire un schéma de compression à partir de n'importe quelle règle d'apprentissage, en formant la famille de règles obtenue en entraînant la méthode de classification initiale successivement sur tous les sous-ensembles d'apprentissage restreints aux parties à h éléments de l'échantillon de départ. Cette méthode peut en particulier fournir un cadre théorique pour aborder le problème de la sélection et de l'agrégation de caractéristiques (features en anglais). D'autres techniques moins faciles à qualifier sur le plan théorique ont aussi remporté des succès pratiques, comme le boosting, dans lequel on sélectionne pas à pas une suite de combinaisons linéaires seuillées de règles de classification de base, en utilisant un critère pondéré dans lequel le poids des exemples mal classés à une étape augmente à l'étape suivante. On obtient un comportement qui reproduit qualitativement celui des schémas de compression, dans le cas « faiblement bruité » où il y a relativement peu d'exemples mal classés : en effet la règle construite au final dépendra dans ce cas essentiellement d'un petit nombre d'exemples.

Conclusion

L'approche statistique s'est imposée ces dernières années comme l'une des voies les plus prometteuses de l'apprentissage automatique. On dispose en particulier actuellement à la fois de méthodes pratiques (support vector machines, boosting, …) qui donnent des résultats encourageants et d'une (et même plusieurs) théorie mathématique pour les étudier. Il reste néanmoins un certain écart entre la théorie et la pratique, les bornes théoriques ayant tendance à se montrer trop pessimistes par rapport aux performances réellement observées, ce qui limite leur pertinence quand on les utilise pour choisir des modèles de classification. D'autre part la théorie porte essentiellement sur la question de l'apprentissage supervisé, qui n'est, comme nous l'avons mentionné dans l'introduction qu'une partie – centrale mais insuffisante en elle-même – d'une méthode concrète de reconnaissance des formes, qui suppose des prétraitements et des post-traitements des données posant eux-mêmes des problèmes de complexité algorithmique et de choix de représentation difficiles et pouvant dans certains cas se prêter à une analyse mathématique fructueuse que nous n'avons pas la place d'aborder ici.

Pour en savoir plus

[B1] Birgé (L.), Model selection via testing : an alternative to (penalized) maximum likelihood estimators, *preprint PMA-862*, (http ://www.proba.jussieu.fr/prepublications.php) (2003).

[C1] Catoni (O.), Statistical learning theory and stochastic optimization, *Ecole d'été de Probabilités de Saint-Flour XXXI - 2001, J. Picard Ed., Lecture notes in mathematics*, **1851**, pp. 1-272, Springer (2004).

[C2] Catoni (O.), A PAC-Bayesian approach to adaptive classification, *preprint PMA-840* (2003) (http ://www.proba.jussieu.fr/prepublications.php).

[C3] Catoni (O.), Improved Vapnik Cervonenkis bounds, *preprint PMA-942* (2004) (http ://www.proba.jussieu.fr/prepublications.php).

[L1] Littlestone (N.), Warmuth (M.), Relating data compression and learnability, *Technical report*, University of California, Santa Cruz (1986).

[M1] McAllester (D. A.), Some PAC-Bayesian Theorems, *Proceedings of the Eleventh Annual Conference on Computational Learning Theory (Madison, WI, 1998)*, 230-234 (electronic), ACM, New York (1998).

[M2] McAllester (D. A.), PAC-Bayesian Model Averaging, *Proceedings of the Twelfth Annual Conference on Computational Learning Theory (Santa Cruz, CA, 1999)*, 164-170 (electronic), ACM, New York (1999).

[M3] Massart (P.), Concentration inequalities and model selection, Saint-Flour lecture notes, (2003) *Springer*, to appear.

[T1] Tsybakov (A.), Optimal aggregation of classifiers in statistical learning, Annals of Statistics, **32**(1), 2004.

[V1] Vapnik (V. N.), *Statistical learning theory*, Wiley, New York (1998).