LA MISE EN OEUVRE DES IDEES DE VLADIMIR N. VAPNIK

Léon Bottou

SUMMARY

There is a fundamental difference between the Vapnik-Chervonenkis theorem and the usual statistical results. The Vapnik-Chervonenkis theorem is neither reducible to the Law of Large Numbers nor to the Bayes formula. This amazing difference leads to the development of a statistical theory able to encompass these new results and to stimulate new solutions to applied problems. This contribution presents original concepts of the Statistical Learning Theory (Vapnik, 1995) and focuses on their applications. We discuss the Structural Risk Minimisation principle, the links with the Regularisation theory, the necessary refinements to the VC-Dimension concept. We briefly mention the analysis of Non-Parametric Algorithms and the Optimal Margin Algorithms.

Le théorème de Vapnik et Chervonenkis diffère fondamentalement des résultats statistiques classiques. En effet, ce théorème ne résulte *ni de la loi des grand nombres, ni de la formule de Bayes* mais repose sur un argument combinatoire nouveau. Cette différence permet le développement d'une théorie statistique qui incorpore ces résultats nouveaux et qui suscite des solutions nouvelles aux problèmes appliqués.

Cette intervention présente quelques constructions originales de la Théorie Statistique de L'Apprentissage (Vapnik, 1995) en mettant l'accent sur ses

applications. Le sujets couverts sont le principe de la Minimisation Structurelle du Risque, ses liens avec la Régularisation, les raffinements nécessaires du concept de dimension de Vapnik-Chervonenkis. Sont également mentionnés le traitement des méthodes Non Paramétriques et les Algorithmes à Marges Optimales.

1. Rechercher un Principe d'Induction

Une grande famille de problèmes d'apprentissage peuvent être représentés par la minimisation d'une fonction de risque de la forme suivante :

$$R(w) = \int Q(x,w)\, dp(x)\,, \quad w \in \Lambda$$

Cette fonction de risque représente l'erreur de généralisation en fonction de la valeur des paramètres w. Elle est composée de deux termes:

- La distribution de probabilité inconnue $p(x)$ régit le comportement des observations x. Elle représente les "lois de la nature" qu'il faut découvrir.
- La fonction de perte (loss function) $Q(x,w)$ indique le coût associé à une observation x particulière. Ce coût dépend d'un paramètre $w \in \Lambda$ qui définit le système que l'on cherche à faire fonctionner.

Exemple 1: Régression.

$$R(w) = \int (f(x,w) - y)^2\, dp(x,y)$$

Le minimum w^* est atteint lorsque la variance de ($f(x,w)$ - E($y|x$)) est minimale. Si l'on suppose que la fonction de régression E($y|x$) appartient à la famille de fonctions paramétrique $\{f(.,w),\ w \in \Lambda\}$, cela signifie que $f(x,w^*)$ = E($y|x$).

Exemple 2: Estimation de Densité.

$$R(w) = \int - \log f(x,w)\, dp(x) \quad \text{avec} \quad \int f(x,w)\, dx = 1 \quad \text{pour tout} \ \ w \in \Lambda$$

Le minimum w^* est atteint lorsque la distance de Kullback-Leibler entre les densités $f(x,w)$ et $p(x)$ est minimale. Si on suppose que la densité à estimer $p(x)$ appartient à la famille de fonctions paramétrique $\{f(.,w),\ w \in \Lambda\}$, cela signifie que $f(x,w^*) = p(x)$.

Lorsque la distribution $p(x)$ est connue, il suffit de résoudre analytiquement ou numériquement ces problèmes d'optimisation. Nous nous intéressons cependant au cas où la distribution $p(x)$ est inconnue, mais où l'on dispose d'un échantillon de L réalisations indépendantes $x_1, x_2 \ldots x_L$.

Cet échantillon ne permet pas de résoudre rigoureusement notre problème d'optimisation. On appelle "Principe d'Induction" tout principe qui permet de trouver une solution approchée à partir de notre échantillon.

1.1. Minimisation Empirique du Risque (ERM)

Le principe d'induction le plus naturel consiste à minimiser l'approximation empirique de notre fonction de risque (i.e. l'erreur d'apprentissage.)

$$R_L(w) = \frac{1}{L} \sum_k Q(x_k, w)$$

La loi des grands nombres assure en effet que cette somme approche la fonction de risque lorsque la taille de l'échantillon augmente.

Les exemples ci-après montrent que les méthodes statistiques fondamentales sont des cas particuliers du principe de Minimisation du Risque Empirique.

Exemple 1: Moindre Carrés

$$R(w) = \int (f(x,w) - y)^2 \, dp(x,y)$$

$$R_L(w) = \frac{1}{L} \sum_k \left(f(x_k, w) - y_k \right)^2$$

Minimiser $R_L(w)$ consiste à appliquer la méthode des Moindres Carrés.

Exemple 2: Maximum de Vraisemblance.

$$R(w) = \int - \log f(x,w) \, dp(x) \quad \text{avec} \quad \int f(x,w) \, dx = 1 \quad \text{pour tout} \quad w \in \Lambda$$

$$R_L(w) = -\frac{1}{L} \sum_k \log f(x_k, w) = -\frac{1}{L} \log \prod_k f(x_k, w)$$

Minimiser la quantité $R_L(w)$ consiste à appliquer la méthode du Maximum de Vraisemblance.

Minimiser une approximation de la fonction de risque pose cependant des problèmes sérieux. Comme le montre la figure 1, l'existence d'un intervalle de confiance entre les valeurs de deux fonctions (i.e. savoir que deux fonctions sont proches dans 95% des cas) ne garantit aucunement que leur minimums vont être proches.

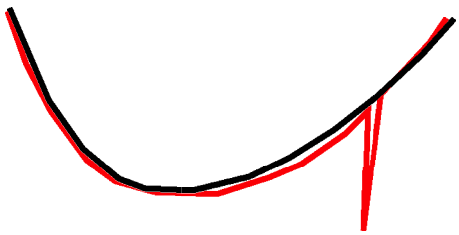

Figure 1: Grandes Déviations

Cette garantie est apportée par les théorèmes de Vapnik-Chervonenkis (Vapnik, Chervonenkis, 1968). Ces théorèmes fournissent un intervalle de confiance reliant $R(w)$ et $R_L(w)$ pour toute valeur de w et donc en particulier pour la valeur de w déterminée par la procédure d'apprentissage.

Le plus simple de ces résultats exprime qu'avec une probabilité 1-η,

$$\forall w \in \Lambda, \quad R_L(w) - D(L,h,\eta) \leq R(w) \leq R_L(w) + D(L,h,\eta)$$

$$\text{avec } D(L,h,\eta) = \sqrt{\frac{h}{L}\left(1+\log\frac{2L}{h}\right) - \frac{1}{L}\log\frac{\eta}{4}}$$

dans le cas ou $Q(x,w) \in [0,1]$

La largeur $D(L,h,\eta)$ de cet intervalle de confiance dépend de la taille L de l'échantillon, de la dimension de Vapnik (ou capacité) h de la famille de fonctions paramétrique $\{Q(.,w), \ w \in \Lambda\}$ et du niveau de confiance 1-η.

Ce résultat permet d'obtenir les traditionnels résultats sur la consistance des méthodes statistiques classiques. Il permet également de concevoir un nouveau principe d'induction plus performant.

1.2. Minimisation Structurelle du Risque (SRM)

Le principe de la Minimisation du Risque Empirique est efficace lorsque le ratio L/h est grand. La largeur $D(L,h,\eta)$ de l'intervalle de confiance est faible, le minimum du risque empirique $R_L(w)$ (i.e. l'erreur d'apprentissage) est proche du minimum du risque réel $R(w)$ (i.e. l'erreur de généralisation.)

En revanche, lorsque le ratio L/h est faible, la largeur de l'intervalle de confiance devient tout à fait significative.

Lorsque le nombre d'exemples L est fixé, il convient alors de minimiser le Risque Garanti, c'est à dire la somme du Risque Empirique $R_L(w)$ et de l'intervalle de confiance $D(L,h,\eta)$.

$$S_1 \subset S_2 \subset \ldots \subset S$$
$$h_1 < h_2 \subset \ldots \subset h$$

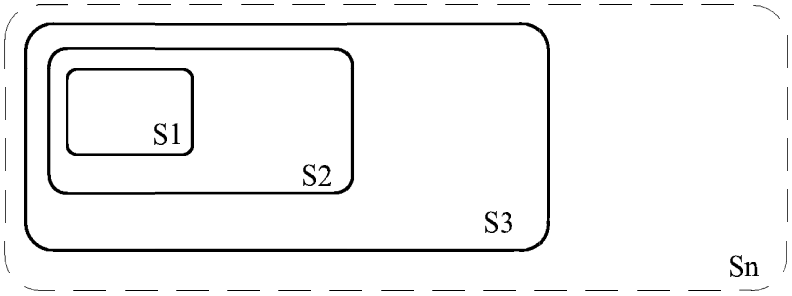

Figure 2: Structure sur les familles de fonctions

On définit donc une suite de sous-ensembles imbriqués $S_i=\{Q(.,w), w\in\Lambda_i\}$ de l'ensemble des fonctions $S=\{Q(.,w), w\in\Lambda\}$ implémentées par notre système d'apprentissage (Figure 2). Ces sous-ensembles possèdent des capacités croissantes h_1, h_2, ... h_N. On réalise alors une optimisation en deux étapes :

- Calcul du minimum du risque empirique dans chaque sous-ensemble.

$$R_L(w^*_i) = \mathrm{Min}_{w\in\Lambda_i} R_L(w)$$

- Sélection du sous-ensemble présentant le meilleur risque garanti.

$$\mathrm{Min}_i \quad R_L(w^*_i) + D(L, h_i, \eta)$$

Cette procédure est illustrée par la figure ci-après (Figure 3). Lorsque l'on choisit des sous-ensembles de capacité croissante, la valeur optimale du risque empirique diminue car on dispose d'une famille de fonctions plus riche. Dans le même temps, la largeur de l'intervalle de confiance augmente car la capacité h augmente.

La structure (i.e. la suite de sous-ensemble) représente donc un pari sur la solution de notre problème d'apprentissage. On espère que sélectionner un sous-ensemble de faible capacité ne pénalisera pas trop le risque empirique

(n'augmentera pas trop l'erreur d'apprentissage) mais réduira l'écart entre le risque empirique et le risque réel (i.e. l'écart entre erreur d'apprentissage et erreur de généralisation).

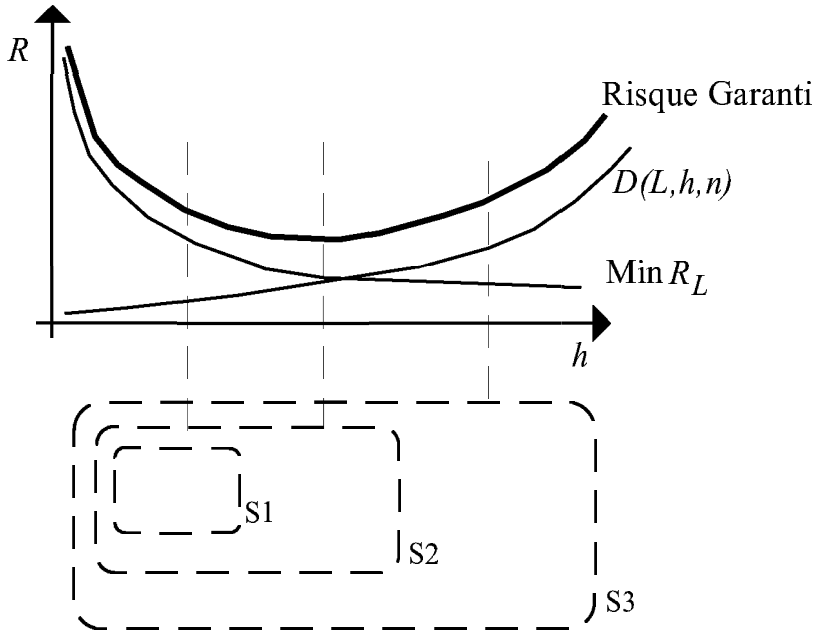

Figure 3: Minimisation Structurelle du Risque

1.3. Trois Exemples de SRM

Cette section reprend trois exemples linéaires concernant une application simplifiée de reconnaissance de caractères manuscrits (Guyon et al., 1991). Ces exemples illustrent bien la SRM mais soulèvent aussi des questions importantes.

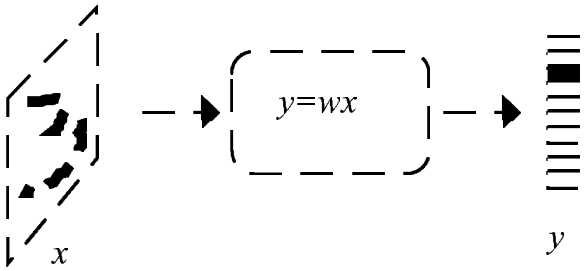

Figure 4 - Reconnaissance Optique de Caractères

L'application consiste à calculer un vecteur de dix scores à partir d'une image binaire représentant l'un des dix chiffres manuscrits (Figure 4). Un ensemble de 1000 exemples est disponible pour l'apprentissage. Chaque

exemple est constitué d'un vecteur x_k de 256 pixels représentant l'image et d'un vecteur y_k contenant les scores désirés (1 ou -1 selon le chiffre représenté). On recherche une matrice W permettant d'obtenir nos scores en calculant le produit Wx.

Système de Référence : Moindres Carrés

La méthode des Moindres Carrés consiste à calculer :

$$W^* = \text{Arg Min}_W \frac{1}{L} \sum_{k=1}^{L} (y_k - Wx_k)^2 = \text{Cov}(x_k, x_k)^{-1} . \text{Cov}(x_k, y_k)$$

où $\text{Cov}(x_k, y_k)$ représente la matrice de covariance empirique des vecteurs (x_k) et (y_k). Cette approche souffre de la forte corrélation des pixels de l'image. Le taux d'erreur mesuré sur un ensemble de 300 exemples distincts atteint13% environ.

Exemple 1 : Sélection de Variables Principales

L'Analyse en Composantes Principales des images x_k donne une suite de matrices de transfert Q_i telles que le produit $Q_i x$ retourne le vecteur des i composantes principales de l'image x. On considère alors la structure définie par les sous-ensembles

$$\Lambda_i = \{UQ_i, U \in \Re^{10 \times i}\}$$

Chaque élément de cette structure restreint en fait la régression linéaire au i premières composantes principales de l'image. La première étape consiste donc à calculer :

$$W_i^* = \left(\text{Arg Min}_U \quad \frac{1}{L} \sum_{k=1}^{L} (y_k - U Q_i x_k)^2 \right) Q_i$$
$$= \text{Cov}(Q_i x_k, Q_i x_k)^{-1} \text{Cov}(Q_i x_k, y_k) \; Q_i$$

La seconde étape consiste à sélectionner la solution qui présente le plus faible risque garanti. On montre que la capacité de l'élément i de notre structure est inférieure à $i+1$.

Cet exemple viole l'intuition selon laquelle la structure doit être déterminée a priori. On remarquera cependant que la structure ne dépend que de la distribution des images x alors que notre problème concerne la dépendance entre l'image et le vecteur de scores y.

Exemple 2 : Weight Decay (ou Ridge Regression)

La sensibilité de Wx au bruit dans l'image x augmente avec le module de la matrice W. Cette constatation suggère une nouvelle structure définie par les sous-ensembles

$$\Lambda_i = \{W, \ ||W||^2 < C_i\} \quad \text{avec} \ \ 0 < C_1 < C_2 < ... < C_N .$$

On calcule la matrice W optimale dans chaque sous-ensemble. D'après le théorème de Kuhn-Tücker, il existe des multiplicateurs de Lagrange $\lambda_1 > ... > \lambda_N > 0$ tels que :

$$W_i^* = \text{Arg Min}_W \left(\frac{1}{L} \sum_{k=1}^{L} (y_k - Wx_k)^2 \right) + \lambda_i ||W||^2$$
$$= (\text{Cov}(x_k , x_k) + \lambda_i I)^{-1} . \text{Cov}(x_k , y_k)$$

Adopter une structure définie par $\{W, K(W)<C\}$ correspond donc à introduire un terme de régularisation $\lambda K(W)$ dans la fonction de coût empirique. La SRM justifie l'emploi et quantifie l'effet des méthodes de régularisation (Ivanov, 1976) en statistiques.

On remarquera qu'imposer une contrainte sur $||W||$ ne change pas la capacité au sens strict. Considérons que les fonctions $\{(y-Wx)^2+b, \ ||W||<C\}$ permettent d'effectuer toutes les dichotomies d'un ensemble de h points (x_i , y_i) pour une valeur fixée de b. Il est alors évident que les fonctions $\{(y-Wx)^2+b, \ ||W||<C/2\}$ permettent d'effectuer toutes les dichotomies de l'ensemble de h points $(2x_i , 2y_i)$.

Il n'en va pas de même si on considère que les coefficients des vecteurs x ou y ne peuvent prendre que deux valeurs +1 ou -1. Il est donc nécessaire d'introduire ce type de considération dans la formulation initiale des théorèmes de Vapnik-Chervonenkis.

Cette technique permet cependant de réduire l'erreur de classification à 6.3%.

Exemple 3 : Lissage

Notre troisième exemple consiste effectuer un lissage préalable des images. L'opération de lissage par un noyau de largeur β peut être représenté la matrice S_β ci-dessous.

$$(S_\beta)_{i,j} = \frac{1}{Z(\beta)} \exp\left(-\frac{d(i,j)^2}{\beta^2}\right)$$

où $d(i,j)$ représente la distance entre les pixels i et j
et $Z(\beta)$ est un terme de normalisation.

Ce lissage ne suffit pas à définir une structure. Les ensembles de fonctions réalisables $\Lambda_\beta = \{ US_\beta, U \in \Re^{10\times 256} \}$ sont en effet tous égaux car S_β est inversible. On peut cependant considérer la double structure définie par les ensembles $\Lambda_{\beta,C} = \{ US_\beta, \|U\|^2 < C \}$.

Cette structure permet d'obtenir une erreur de classification de 4.3%.

1.4. Mise en oeuvre de la SRM.

Toutes ces approches ont été comparées dans (Guyon & al., 1991). En particulier, un même graphe rassemble le risque garanti et l'erreur de généralisation mesurés sur un ensemble de test distinct. Ce graphe permet de constater que le minimum de l'erreur garantie signale bien le minimum de l'erreur de généralisation. Ce travail révèle néanmoins un certain nombre de problèmes pratiques et théoriques:

- Comme nous l'avons signalé plus haut, il est nécessaire d'adapter les théorèmes généraux pour tenir compte de nos connaissances concernant les données.
- Le calcul analytique de la capacité n'est possible que dans les cas les plus simples. Nous avons dû utiliser des techniques *ad hoc* pour mesurer la capacité pour les exemples 2 et 3.
- L'application naïve des théorèmes donne une borne très large du risque garanti. En outre, le calcul d'intervalles de confiance plus étroits présente de sérieuses difficultés techniques.

Ces problèmes rendent difficile la mise en pratique de la SRM au sens strict. Il est souvent plus simple d'utiliser des techniques de validation croisée pour déterminer l'optimum de la capacité.

1. Une première approche consiste à réserver quelques données pour constituer un ensemble de validation. On peut alors estimer l'erreur de généralisation sur cet ensemble et sélectionner l'élément de notre structure qui se comporte le mieux. C'est l'approche *la plus simple et la plus utilisée*. Cependant, lorsque les données sont rares ou complexes, il

est coûteux de se priver d'une partie importante des données pendant l'apprentissage de notre système.

2. Une seconde approche consiste à effectuer une *mesure indirecte de la capacité*. Si on dispose de L exemples, on peut mesurer les erreurs d'apprentissage et de généralisation obtenues en optimisant notre système sur un ensemble d'apprentissage de taille $L/2$. Si on connaît la forme des courbes d'apprentissage, cette mesure nous permet de prédire ce qui se passe avec un ensemble d'apprentissage de taille L.

Cette seconde approche requiert une grande augmentation de la précision des formules de Vapnik-Chervonenkis. De nombreuses avancées ont été faites dans cette direction.

1.5. Pourquoi les Réseaux Multicouches fonctionnent.

Avant d'aborder ce sujet, il est intéressant de montrer comment la Minimisation Structurelle du Risque permet d'expliquer le fonctionnement des Réseaux de Neurones multicouches (MLP).

De multiples arguments en effet semble indiquer que les MLP ne peuvent pas fonctionner. Or la pratique indique clairement le contraire. Cette anomalie mérite une explication.

- Un MLP est un dispositif statistique grossièrement sur-paramétré. Certains réseaux de neurones pour la reconnaissance de caractères contiennent quelques dizaines de milliers de paramètres, sont entraînés avec quelques milliers d'exemples seulement, et fonctionnent néanmoins fort bien.
- Un MLP est un dispositif numérique ridicule. Aucun numéricien ne peut admettre qu'il est possible d'optimiser une fonction non linéaire de 10000 variables avec un simple algorithme de descente de gradient.

Cette contradiction peut être résolue par l'argument suivant : l'inefficacité de la descente de gradient crée une structure implicite sur l'ensemble des fonctions représentables par le réseau.

1. Les poids d'un MLP sont en général initialisés avec des valeurs assez faibles. Cela assure que seule la partie centrale des sigmoïdes est utilisée. Le réseau possède donc un comportement linéaire.

2. L'algorithme de descente de gradient est lent. Cela assure que l'apprentissage cherchera d'abord une solution linéaire. Les non-linéarités entrent progressivement en jeu lorsque la norme des poids augmentent.

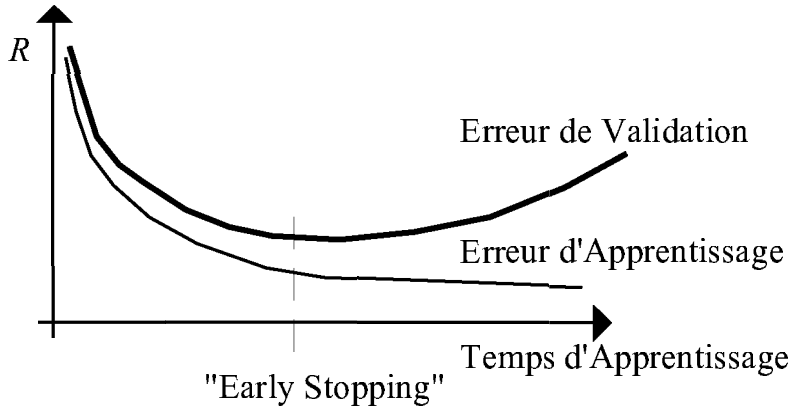

Figure 5 - "Early Stopping" dans un MLP

3. On arrête la procédure d'apprentissage lorsque la performance mesurée sur un ensemble de validation distinct culmine (Figure 5). Cette procédure (*early stopping*) permet de sélectionner une solution efficace.

Cette interprétation du fonctionnement des MLP est supportée par un grand nombre d'observations concernant le rôle de l'initialisation des poids, le régime initialement linéaire de l'apprentissage, ou la croissance du module des poids tout au long de l'apprentissage. Cette interprétation donne en outre un éclairage utile sur les astuces utilisées pour améliorer le fonctionnement des MLP : équilibrage du module des poids dans chaque couche du réseau (poids initiaux, forme des sigmoïdes), ralentissement de la croissance des poids (weight decay), méthodes d'arrêt de l'algorithme (validation croisée).

Cependant, il n'est pas toujours satisfaisant d'appliquer la méthode SRM en se reposant sur le manque d'efficacité de notre algorithme d'optimisation. Il est parfois possible de contrôler rigoureusement cet aspect des réseaux de neurones pour obtenir des performances légèrement améliorées.

2. Raffiner les Théorèmes de Vapnik Chervonenkis

Les théorèmes de Vapnik et Chervonenkis fournissent des intervalles de confiance reliant erreur d'apprentissage et erreur de généralisation. Dans le cadre d'une utilisation pratique, il faut bien sûr de considérer les intervalles de confiance les plus étroits possible.

2.1. Bornes Relatives

Les expressions de ces intervalles de confiances sont fortement liées au résultat de Hoeffding concernant l'estimation de la moyenne de L variables aléatoires indépendantes et identiquement distribuées dans $\{0,1\}$.

Considérons L variables aléatoires $x_1 ... x_L$ prenant la valeur 1 avec une probabilité μ et prenant la valeur 0 avec une probabilité $1-\mu$. Le théorème de Hoeffding précise que

$$P\{ \mid \mu - \frac{1}{L} \sum x_i \mid > \varepsilon \} \leq 2 \exp (-2L\varepsilon^2)$$

Ce résultat est notoirement imprécis lorsque la moyenne des variables est faible. Cela signifie que la borne de Vapnik est très imprécise lorsque l'erreur est faible. On peut donc dériver de nouvelles bornes fondées sur des résultats plus précis que le résultat de Hoeffding.

La borne initiale majore l'écart absolu entre erreur d'apprentissage et de généralisation. Cette borne possède la forme suivante (on supposera tout au long de cette section que $Q(x,w) \in [0,1]$) :

$$P\{ \sup_w \mid R(w) - R_L(w) \mid > \varepsilon \} \leq \left(\frac{2Le}{h} \right)^h 4 \exp (-L\varepsilon^2)$$

Il existe également une majoration de l'écart relatif entre ces erreurs. Cette majoration est beaucoup plus fine lorsque l'erreur d'apprentissage est faible.

$$P\{ \sup_w \frac{R(w) - R_L(w)}{\sqrt{R(w)}} > \varepsilon \} \leq \left(\frac{2Le}{h} \right)^h 4 \exp (-\frac{L\varepsilon^2}{4})$$

Par ailleurs, on sait que les coefficients qui apparaissent dans ces formules sont essentiellement dus aux problèmes techniques de la preuve. On travaille en pratique avec des coefficients déterminés de façon empirique :

$$P\{ \sup_w \frac{R(w) - R_L(w)}{\sqrt{R(w)\,(1-R(w))}} > \varepsilon \} \leq A \left(\frac{2Le}{h} \right)^h \exp (-B\, L\varepsilon^2) \qquad \text{[conjecturée]}$$

2.2. Conditions Nécessaires et Suffisantes

Les théorèmes de Vapnik et Chervonenkis fournissent un intervalle de confiance concernant l'écart entre l'erreur de généralisation et l'erreur d'apprentissage pour *toute* valeur de w (et pas seulement pour la valeur calculée par l'algorithme d'apprentissage). Lorsque la capacité est finie, la largeur de cet intervalle tend vers 0 lorsque le nombre d'exemples augmente.

Il a beaucoup été argumenté que cette convergence uniforme était une condition suffisante d'apprentissage et non une condition nécessaire. Un résultat récent (Vapnik, Chervonenkis, 1989) prouve au contraire que cette convergence uniforme est une condition nécessaire : en effet, si il est possible de créer un gros écart entre les erreurs d'apprentissage et de généralisation, alors un algorithme qui recherche la meilleure erreur d'apprentissage tombera sûrement dans ce piège.

Par ailleurs, les théorèmes de Vapnik et Chervonenkis donnent les conditions nécessaires et suffisantes pour que cette convergence uniforme ait lieu (Vapnik, 1995). Cela signifie que toute description complète des phénomènes d'apprentissage doit obligatoirement inclure un concept équivalent à la dimension de Vapnik-Chervonenkis.

2.3. Bornes Dépendant des Données

En général, les bornes de Vapnik et Chervonenkis ne dépendent pas de la distribution de probabilité $p(x)$. Aucune hypothèse n'est nécessaire sur la nature des données. En pratique cependant, on dispose d'informations à propos des données et l'on souhaite exploiter cette information pour affiner nos estimations.

Par exemple, la capacité d'une famille de fonctions discriminantes linéaires à n dimensions vaut n+1. Si on sait que toutes nos données résident dans un sous-espace de dimension n/2, on peut appliquer les bornes de Vapnik avec une capacité $h = n/2+1$.

Considérons un problème de reconnaissance des formes avec deux classes. On considère que notre système d'apprentissage peut implémenter un ensemble S de fonctions discriminantes. On note $\aleph_S(x_1...x_L)$ le nombre de dichotomies réalisées par les fonctions de S sur les points $x_1 \ldots x_L$.

On définit les fonctions suivantes :

$$H(L) = \underset{x_1 \ldots x_L}{\mathrm{E}} \log \aleph_S(x_1 \ldots x_L) \qquad \text{(VC-Entropy)}$$

$$H_{ann}(L) = \log \underset{x_1 \ldots x_L}{\mathrm{E}} \aleph_S(x_1 \ldots x_L) \qquad \text{(Annealed Entropy)}$$

$$H_{max}(L) = \log \sup_{x_1 \ldots x_L} \aleph_S(x_1 \ldots x_L) \qquad \text{(Log Growth Function)}$$

Ces fonctions vérifient la propriété suivante :

$$H(L) \leq H_{\text{ann}}(L) \leq H_{\max}(L) \leq h\,(1 + \log L/h)$$

où h désigne la VC-Dimension de S.

Ces quantités sont à l'origine des conditions nécessaires et suffisantes:

- Il y a convergence uniforme pour une distribution $p(x)$ fixée si et seulement si le ratio $H(L)/L$ tend vers 0 lorsque L augmente. Il est donc possible d'optimiser le risque pour cette distribution de probabilités particulière.
- Il y a convergence uniforme pour toute distribution $p(x)$ si et seulement si le ratio $H_{\max}(L)/L$ tend vers 0 lorsque L augmente. L'apprentissage est donc possible dans pour toute distribution de probabilités.

A ce jour, on ne connaît pas de bornes dépendant de la quantité $H(L)$. On dispose en revanche des bornes suivantes (Vapnik, 1995) qui dépendent de la quantité $H_{\text{ann}}(L)$.

$$\mathrm{P}\{\sup_w |R(w) - R_L(w)| > \varepsilon\} < 4\exp\left(H_{\text{ann}}(2L) - L\varepsilon^2\right)$$

$$\mathrm{P}\{\sup_w \frac{R(w) - R_L(w)}{\sqrt{R(w)}} > \varepsilon\} < 4\exp\left(H_{\text{ann}}(2L) - \frac{L\varepsilon^2}{4}\right)$$

$$\mathrm{P}\{\sup_w \frac{R(w) - R_L(w)}{\sqrt{R(w)(1-R(w))}} > \varepsilon\} < A\exp\left(H_{\text{ann}}(2L) - BL\varepsilon^2\right) \quad \text{[conjecturée]}$$

On peut souvent borner $H_{\text{ann}}(2L)$ par une expression plus favorable que $h(1+\log(2L/h))$. Cette technique fournit alors des bornes dépendant des donnée et donc beaucoup plus réalistes.

2.4. Mesurer la Dimension de Vapnik Chervonenkis

On peut également essayer de mesurer $H_{\text{ann}}(L)$ pour quelques valeurs de L. La valeur de $H_{\text{ann}}(L)$ augmente avec le nombre d'exemples L. Deux cas peuvent se présenter :

- Lorsque $H_{\text{ann}}(L)$ croit rapidement avec L (e.g. linéairement ou plus), l'apprentissage est très difficile voire impossible car $H(L)/L$ ne tend pas vers zéro avec rapidité.
- Lorsque $H_{\text{ann}}(L)$ croit modérément avec L (e.g. logarithmiquement), l'apprentissage est bien représenté par les bornes de la section ci-avant.

Dans ce dernier cas, on peut chercher la meilleure valeur h' permettant d'approcher les points $H_{ann}(L)$ par l'expression $h'(1+\log L/h')$. On appelle cette mesure "VC-Dimension Effective" (Vapnik, Levin, Le Cun, 1994).

La mesure de $H_{ann}(L)$ est assez délicate. Nous ne connaissons pas encore de procédure universelle. La méthode ci-après fonctionne pour des fonctions discriminantes *linéaires.*

1. Choisir au hasard deux ensembles d'exemples distincts E et F contenant L exemples.
2. Trouver les paramètres w^* qui maximisent l'écart $R_E(w)$-$R_F(w)$. Cela est aisément effectué dans le cas linéaire en utilisant notre algorithme d'apprentissage sur nos exemples après avoir inversé les classes des éléments de E.
3. Répéter ces opérations pour plusieurs choix des ensembles E et F.
4. Mesurer la médiane $M(L)$ de l'écart relatif maximal :

$$M(L) = \text{Median} \frac{R_E(w^*) - R_F(w^*)}{\sqrt{R(w^*)(1-R(w^*))}}$$

5. Répéter ces opérations pour plusieurs valeurs de L.
6. Déterminer les constantes A et B et les valeurs $H_{ann}(2L)$ telles que :

$$A \exp\left(H_{ann}(2L) - BL\, M(L)^2 \right) \approx 1$$

Cette procédure ne fonctionne pas très bien dans le cas non-linéaire. Les algorithmes d'optimisation ne permettent pas en effet de mener à bien la seconde étape avec une fiabilité suffisante.

3. Autres Principes d'Induction

La Minimisation Structurelle du Risque est une première étape dans la mise en oeuvre de la Théorie Statistique de l'Apprentissage. Nous mentionnons ici quelques directions prometteuses. Le lecteur intéressé trouvera plus d'information dans les références citées.

3.1. Minimisation Locale du Risque

Les techniques statistiques non paramétriques peuvent être traitées par un léger changement de la formulation du problème (Vapnik, Bottou, 1993). Il

ne s'agit plus de trouver une dépendance fonctionnelle, mais de trouver la valeur d'une fonction inconnue en un point x_0 fixé à l'avance.

Le risque à minimiser possède alors la forme suivante :

$$R(w) = \frac{\int Q(y,x,w)\, K(x-x_0, \beta)\, dp(x,y)}{\int K(x-x_0, \beta)\, dp(x,y)}$$

où $K(x-x_0, \beta)$ représente une fonction noyau. Cette fonction définit un voisinage de largeur β du point x_0 sur lequel on souhaite minimiser le coût $Q(y,x,w)$.

Si on connaissait la distribution $p(x,y)$, il suffirait de choisir $\beta=0$ et de minimiser analytiquement le risque mesuré sur le seul point x_0. Nous nous intéressons cependant au cas où l'on ne connaît pas la distribution $p(x,y)$ mais où l'on dispose d'un ensemble d'exemples (x_k, y_k). L'approximation empirique du risque s'écrit alors :

$$R(w) = \frac{\sum_{k=1}^{L} Q(y_k, x_k, w)\, K(x_k-x_0, \beta)}{\sum_{k=1}^{L} K(x_k-x_0, \beta)}$$

On ne peut pas alors considérer simplement un noyau de largeur $\beta=0$ car en général, aucun exemple ne tombe exactement sur le point x_0. La largeur du noyau β commande donc le nombre d'exemples utilisés pour effectuer notre approximation locale.

On montre alors que la taille de notre intervalle de confiance uniforme dépend de la largeur du noyau β. Cela permet de définir le principe de Minimisation Structurelle et Locale du Risque qui consiste à optimiser à la fois la capacité h et la largeur de noyau β

Diverses formes du coût Q et de la fonction noyau K permettent d'analyser les algorithmes non paramétriques usuels : Plus Proches Voisins, Fenêtres de Parzen, Algorithme de Watson-Nadaraya, LOESS, etc.

L'efficacité de cette méthode a été démontrée tant sur des problèmes de reconnaissance de caractères (Bottou, Vapnik, 1992) que sur des problèmes de prévision de consommation (Bottou, Driancourt, Ignace, 1995).

3.2. Algorithmes à Marges Optimales

Les algorithmes à Marges Optimales (ou algorithmes à Points de Support) fonctionnent en renversant l'ordre des optimisations. On cherche maintenant à minimiser notre intervalle de confiance tout en maintenant une erreur constante sur l'ensemble d'apprentissage.

Ces algorithmes semblent extrêmement intéressants :

- Ils permettent d'ajuster automatiquement capacité et nombre d'exemples,
- L'optimisation peut être réalisée rapidement et exactement en travaillant dans l'espace dual avec des techniques d'optimisation quadratique avec contraintes. Cette technique peut être appliquée à l'identique pour les problèmes linéaires ou non.
- La solution est représentée par une combinaison linéaire d'un petit nombre d'exemples appelés points de support. Il n'est pas nécessaire de représenter entièrement l'espace de paramètres. Cela permet de travailler dans des espaces de fonctions possédant un très grand nombre de paramètres (e.g. 10^{22}).
- Le nombre de points de support permet d'estimer directement une borne sur l'erreur retournée par la méthode de validation croisée "leave-one-out".
- Les points de support sont une caractéristique assez robuste. Ils représentent une grande partie de l'information portée par notre ensemble d'exemples.

Il n'est pas possible de décrire ici l'ensemble de ces méthodes. Le lecteur intéressé peut se reporter à (Boser, Guyon, Vapnik, 1992), au chapitre 5 de (Vapnik, 1995) et à (Cortes, Vapnik, 1995).

4. Conclusion

En élargissant les concepts statistiques usuels, la Théorie Statistique de l'Apprentissage réconcilie les divergences théoriques et empiriques entre méthodes statistiques et méthodes connexionnistes.

Ces explications permettent d'améliorer les méthodes d'apprentissage statistiques par l'introduction de techniques de contrôle de la VC-Dimension

(régularisation statistique, introduction de contraintes, pré-traitements, marges optimales, etc.)

L'enjeu consiste désormais à apporter les bonnes propriétés des réseaux de neurones (et celles là seulement) dans les centaines d'algorithmes statistiques utilisés et raffinés depuis quatre décennies.

Références

Boser B., Guyon I., Vapnik V. N. (1992) : "A Training Algorithm for Optimal Margin Classifiers", *Fifth Annual Workshop on Computational Learning Theory,* 144-152, ACM Pittsburgh.

Bottou L., Cortes C. Denker J. S., Drucker H., Guyon I., Jackel L. D., Le Cun Y., Müller U., Säckinger E., Simard P., Vapnik V. N. (1994) : "Comparison of Classifier Methods: A case study in Hand-written Digit Recognition", *Proceedings 12th IAPR International Conference on Pattern Recognition* **2**:77-83, IEEE Computer Society Press, Los Alamos.

Bottou L., Driancourt X., Ignace Ch. (1995) : "Quelques Applications de TL/Prévision", to appear in *Proceedings of ECANN 95.*

Bottou L., Vapnik V. N. (1992): "Local Learning Algorithms", *Neural Computation* **4**(6):888-901

Cortes C., Vapnik V. N. (1995): "Support Vector Networks", to appear in *Machine Learning.*

Guyon I., Vapnik V. N., Bottou L., Solla S. A. (1991) : "Structural Risk Minimization for Character Recognition", *Neural Information Processing Systems* **4**, Morgan-Kaufman, Denver.

Ivanov V. V. (1976) : *The Theory of Approximate Methods and their Application to the Numerical Solution of Singular Integral Equations*, Nordhoff International, Leyden.

Vapnik V. N. (1995) : *The Nature of Statistical Learning Theory*, Springer Verlag, à paraître.

Vapnik V. N., Bottou L. (1993): "Local Algorithms for Pattern Recognition and Dependencies Estimation", *Neural Computation* **5**(6):893-908

Vapnik V. N., Chervonenkis A. Ja., (1968) : "On the Uniform Convergence of Relative Frequencies of Events to their Probabilities", *Doklady Academii Nauk USSR,* **181**(4), (English transl. *Sov Math. Dokl*).

Vapnik V. N., Chervonenkis A. Ja., (1989) : "The necessary and sufficient conditions for consistency of the method of empirical risk minimization", *Pattern Recognition and Image Analysis* **1**(3):284-305.

Vapnik V. N., Levin E., Le Cun Y. (1994) : "Measuring the VC-Dimension of a Learning Machine", *Neural Computation* **6**(5).