ÉCOLE
SUPÉRIEURE
D'ÉLECTRICITÉ

Support de cours

Théorie de la généralisation

Didacticiel sommaire

Hervé Frezza-Buet, Supélec

28 juin 2012

Herve.Frezza-Buet@supelec.fr

Table des matières

1 Introduction et notation

1.1 Définition informelle du problème

Il s'agit dans ce document d'esquisser les grandes lignes de l'apprentissage statistique, sachant qu'une étude plus approfondie peut commencer par la lecture des ouvrages cités en bibliographie.

La théorie de la généralisation est une approche statistique des problèmes que l'on rencontre lorsqu'on prétend modéliser des données dont on ne connaît pas la loi qui préside à leur distribution. Modéliser signifie proposer une procédure, appelée hypothèse, qui se comporte comme la loi qui génère les données. Plus précisément on suppose qu'un processus qu'on ne connaît pas, disons un oracle, associe à un vecteur une étiquette, et on aimerait expliciter une procédure qui, voyant arriver le même vecteur, lui associe l'étiquette que l'oracle lui aurait associée. Notre procédure, l'hypothèse doit mimer le comportement de l'oracle. Comment faire ? On va partir d'une base d'exemples, constituée de vecteurs et des étiquettes que l'oracle a données à ces vecteurs. A partir de là, on va tâcher de généraliser [1] ces exemples par une hypothèse, pour que lorsqu'on reçoit un vecteur, on puisse utiliser l'hypothèse pour lui associer une étiquette, qui a la meilleur chance d'être celle que l'oracle aurait fourni.

Concrètement, imaginons qu'on veuille prédire le volume V d'un pain à partir du temps de cuisson t, de la température T du four, et des quantités F, l et s de farine, de levure

1. D'où le nom de théorie de la généralisation.

et de sel. Dans ce problème, à un vecteur (t, T, F, l, s), le boulanger qui suit cette recette fournit une étiquette v, lorsque son pain sort du four. C'est l'oracle de l'histoire. Après maintes expérimentations, faites à chaque fois avec un (t, T, F, l, s) et le v qui en découle, on peut constituer une base en consignant méticuleusement les $((t, T, F, l, s), v)$ observés. On souhaite proposer une formule mathématique, l'hypothèse, qui puisse deviner v en fonction des conditions (t, T, F, l, s). La construction de cette formule est ce qu'on appelle la généralisation à partir des exemples de la base, c'est le travail de nombreuses techniques d'apprentissage.

Une remarque : cela permet de prédire au boulanger le volume de son pain si il applique telle ou telle recette. En revanche, le problème de trouver la recette qui donne un pain de volume souhaité est appelé problème inverse. C'est autre chose...

1.2 Notation et formalisation

On pose les notations suivantes :

les $x_i \in X$
: Ce sont les vecteurs auxquels on doit associer une étiquette. Dans le cas du boulanger, il s'agit des (t, T, F, l, s).

$\mathcal{D}_X$
: Il s'agit de la distribution qui préside au tirage i.i.d [2] des x_i. On utilisera la notation $\mathcal{D}_\bullet$ pour d'autres ensembles.

Oracle
: L'oracle donne une étiquette $u_i \in U$ à chaque exemple x_i, selon une distribution $F(x, u) = P(u \mid x)$ inconnue.

$S \subset X \times U$
: S définit une base d'exemples d'après l'oracle, on note $|S|$ le nombre d'exemples.

$h \in H$
: H est un espace de fonctions connu, dans lequel on cherche une hypothèse qui calcule $u = h(x)$ de façon « proche » de la distribution F de l'oracle.

$l(u_1, u_2) \in \mathbb{R}^+$
: l, le coût, est une mesure d'erreur d'étiquetage, on s'en sert pour calculer l'erreur commise par une hypothèse h par rapport au couple (x_i, u_i) fournit par l'oracle. Cette erreur est $l(h(x_i), u_i)$.

$\mathcal{D}_{X \times U}$
: Il s'agit de la distribution induite par $\mathcal{D}_X$ et F sur $X \times U$ (inconnue car F est inconnue), à savoir celle qui préside à la constitution des bases d'exemples S.

$\mathcal{R}_{\text{réel}}(h) = \int_{(x,u) \in X \times U} l(u, h(x))\, dF(x, u)$
: Cette expression définit le risque réel. Il s'agit de l'erreur, mesurée sur tout les (x, u) possibles, en tenant compte de la probabilité $dF(x, u)$ d'avoir chacun d'eux. Bref, c'est l'espérance du coût de h.

$\mathcal{R}^S_{\text{emp}}(h) = \frac{1}{|S|} \sum_{(x,u) \in S} l(u, h(x))$
: Cette expression définit le risque empirique, qui estime le risque réel $\mathcal{R}_{\text{réel}}(h)$ par la moyenne des coûts observés sur la base d'exemples S.

2. C'est de l'anglais : *independently and identically*, ça veut dire que les tirages suivent tous la même loi, et qu'ils sont indépendants les uns des autres.

$h^\star = \underset{h \in H}{\operatorname{argmin}} \mathcal{R}_{\text{réel}}(h)$

$h^\star$ est l'hypothèse la meilleure que l'on puisse trouver dans H pour simuler le comportement de l'oracle.

$\hat{h}_S$

Cette notation désigne une hypothèse produite par une recherche qui se fonde sur la base d'exemple S. En effet, la généralisation des observations prises dans S conduit à élaborer une hypothèse, dont il convient de marquer sa dépendance par rapport à la base qui a servi à la construire. Une autre base S' pourrait très bien nous amener à construire une autre hypothèse $\hat{h}_{S'}$ qui soit différente de $\hat{h}_S$, alors que les deux prétendent approcher le comportement du même oracle.

2 Les différentes sources d'erreur

2.1 Le biais inductif

Le biais inductif est donné par $\mathcal{R}_{\text{réel}}(h^\star)$. S'il est non nul, cela signifie que le meilleur des $h \in H$ que l'on trouve n'est pas capable de reproduire parfaitement le comportement de l'oracle. Le biais inductif est une erreur inhérente à l'inadéquation de notre espace d'hypothèse H pour simuler l'oracle. Intuitivement, c'est le cas quand H n'est pas assez complexe pour rendre compte de la complexité de l'oracle.

2.2 La variance

La variance est une source d'erreur plus subtile. De la variance de quoi parle-t-on ? Lorsqu'on construit $\hat{h}_S$ en fonction de S, on a dans l'idée que $\hat{h}_S$ est une bonne approximation de $h^\star$. Cela signifie que d'une base d'exemple S à l'autre, les $\hat{h}_S$ qu'on en déduit ne devraient pas trop *varier* les unes des autres. Si cette variance est grande, cela signifie que le processus de construction de $\hat{h}_S$ à partir de S est trop dépendant de S, on peut dire intuitivement que ce processus colle trop aux données, sans les généraliser. C'est ce qu'on appelle le sur-apprentissage [3], ou apprentissage par cœur. En général, si H est riche, c'est-à-dire s'il s'y trouve des hypothèses complexes, on risque le sur-apprentissage.

3 Compromis biais/variance

Sachant les définitions précédentes du biais inductif et de la variance, on peut formuler l'essence même de tout problème de généralisation. Pour réduire le biais inductif, on va complexifier l'espace d'hypothèses H, et ainsi augmenter la variance. C'est très pénible en pratique, car il est difficile de trouver le bon compromis.

Prenons le cas d'une régression aux moindres carrés, ce qui veut dire que $l(u_1, u_2) = (u_1 - u_2)^2$, avec $X = \mathbb{R}$, $U = \mathbb{R}$, et où $|S| = 5$, comme illustré figure 1.

Choisissons comme espace d'hypothèse H les polynômes de degré 1, les droites. Un bon algorithme d'apprentissage devrait conduire à une fonction $\hat{h}_S$ du style de celle de la figure 2. On voit bien que, du fait que l'on a des points qui, de toute façon, ne sont pas alignés, on aura toujours une erreur. L'espérance de cette erreur dans le meilleur des cas est justement le biais inductif.

Essayons maintenant avec comme espace d'hypothèses H les polynômes de degré 2. Un bon algorithme d'apprentissage devrait conduire à une fonction $\hat{h}_S$ du style de celle de la

3. *Overfitting* en anglais.

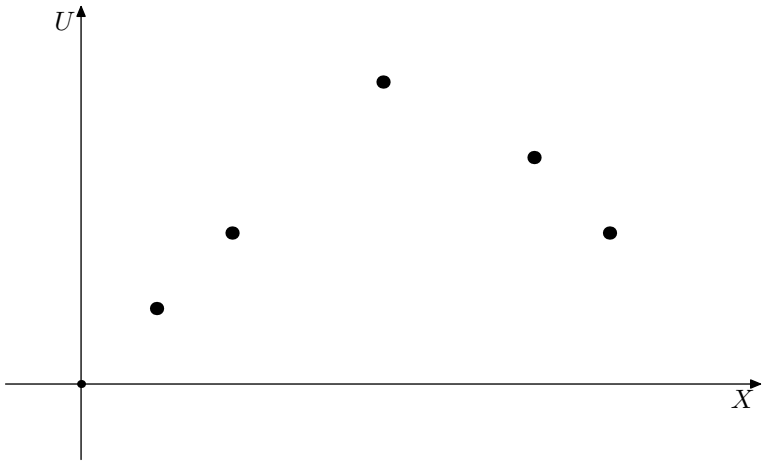

FIGURE 1 – Exemple avec $X = \mathbb{R}$, $U = \mathbb{R}$ et $|S| = 5$.

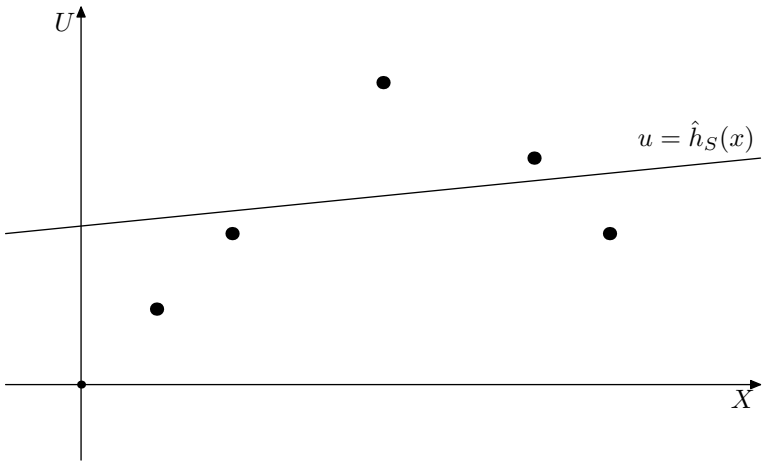

FIGURE 2 – Avec H les polynômes de degré 1.

figure 3. On voit que cette famille d'hypothèses permet de trouver en son sein une hypothèse meilleure que lorsqu'on utilise la famille des droites. Le biais inductif est moindre, à en croire ce S-là du moins.

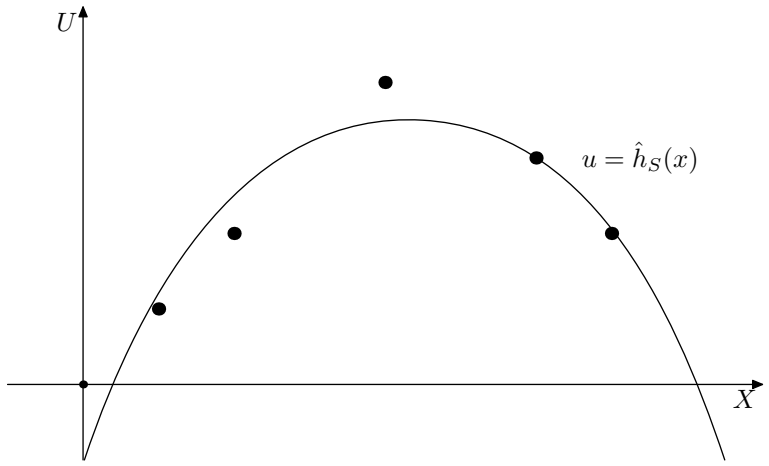

FIGURE 3 – Avec H les polynômes de degré 2.

Essayons d'enrichir encore l'espace d'hypothèses, en prenant les polynômes de degré 25. Un bon algorithme d'apprentissage devrait conduire à une fonction $\hat{h}_S$ du style de celle de la figure 4. Le risque empirique $\mathcal{R}^S_{\text{emp}}\left(\hat{h}_S\right)$ est pourtant nul puisque la courbe passe par tous les points... mais on voit bien que $\hat{h}_S$ généralise mal le comportement de l'oracle.

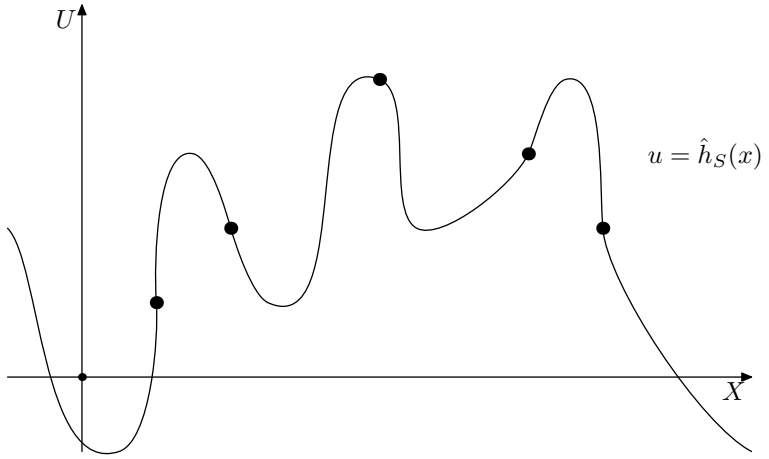

FIGURE 4 – Avec H les polynômes de degré 25. On constate un sur-apprentissage.

4 Quelques principes inductifs

Le problème de l'apprentissage, dans notre cas, consiste à trouver, à partir de S, une hypothèse $\hat{h}_S$, en essayant d'obtenir que $\mathcal{R}_{\text{réel}}\left(\hat{h}_S\right)$ soit petit, proche du biais inductif.

Induire, dans ce cas, c'est induire l'effet sur le risque réel du processus qui produit $\hat{h}_S$ en fonction de S.

4.1 Principe de minimisation du risque empirique

Ce principe consiste à trouver l'hypothèse pour laquelle, le risque empirique est minimal sur S. On augure alors que son risque réel l'est aussi. Pour résumer, le principe de minimisation du risque empirique [4] s'écrit :

$$\hat{h}_S = \underset{h \in H}{\operatorname{argmin}} \mathcal{R}^S_{\text{emp}}(h) \tag{1}$$

C'est ce que font pas mal de méthodes, telles que les régressions aux moindres carrés, les perceptrons multi-couches...

4.2 Principe de décision bayésienne

Non traité ici.

4.3 Principe de régulation d'hypothèse

La régulation d'hypothèse consiste à considérer, lors de l'induction, la complexité des hypothèses comme facteur pénalisant.

4.3.1 Régularisation

On procède de même que pour la minimisation du risque empirique, si ce n'est qu'on ajoute un terme pénalisant les hypothèses complexes. Il faut se doter d'une fonctionnelle Ω, qui à une des hypothèse associe un nombre, élevé si l'hypothèse est complexe. Le principe de régularisation s'écrit :

$$\hat{h}_S = \underset{h \in H}{\operatorname{argmin}} \left[\mathcal{R}^S_{\text{emp}}(h) + \Omega(h) \right] \tag{2}$$

4.3.2 Minimisation du risque structurel

Il s'agit ici de pénaliser non pas chaque hypothèse, mais une classe entière d'hypothèses. Ainsi, au lieu d'avoir des termes de pénalisation de la forme $\Omega(h)$, comme dans le cas de la régularisation, on a une pénalisation de la forme $\Omega(H)$. Il faut alors construire une suite d'espaces d'hypothèses $\{H_d\}_d$, telle que $H_1 \subset H_2 \subset \cdots \subset H_d \subset \cdots$. Par exemple, on a bien ce type d'inclusion si H_d est l'ensemble des polynômes de degré inférieur ou égal à d. Se fixant une base d'exemple S, on note $\hat{h}^d_S$ l'hypothèse induite d'après S sur H_d. La suite $\left\{ \mathcal{R}^S_{\text{emp}} \left(\hat{h}^d_S \right) \right\}_d$ est décroissante, les $\{H_d\}_d$ étant imbriqués, et tend vers 0 puisqu'à complexifier l'espace d'hypothèse, on finira bien par sur-apprendre sur la base d'exemple S.

Le principe de minimisation du risque structurel consiste à trouver le d qui convient, c'est-à-dire celui qui vérifie :

$$d^\star = \underset{d}{\operatorname{argmin}} \left[\mathcal{R}^S_{\text{emp}} \left(\hat{h}^d_S \right) + \Omega(H_d) \right] \tag{3}$$

La difficulté réside dans le fait de trouver une fonction Ω pertinente...

4. ERM : *Empirical Risk Minimization* en anglais.

5 Analyse PAC

L'analyse PAC[5] désigne le fait de considérer comme correct un processus quand la probabilité qu'il soit incorrect est bornée par une petite valeur.

Cette analyse est valable quand on a une relation sur les risques exprimée sur la figure 5. Cette relation est assez naturelle, mais pas toujours vérifiée, on ne rentrera pas dans le détail sur ce point. Plus la base d'exemple S est grande, plus $\mathcal{R}_{\text{réel}}\left(\hat{h}_S\right)$ et $\mathcal{R}^S_{\text{emp}}\left(\hat{h}_S\right)$ se rapprochent, puisque le second est un estimateur du premier, et qu'en augmentant la taille de l'échantillon, la moyenne se rapproche de l'espérance. Pour de petites tailles de corpus, l'espace d'hypothèse H considéré est capable de sur-apprendre, et ne s'en prive pas, d'où un risque empirique qui stagne à 0. Ce qui ne va pas de soi, mais est raisonnable en pratique, c'est que $\mathcal{R}_{\text{réel}}\left(\hat{h}_S\right)$ tende vers le biais inductif $\mathcal{R}_{\text{réel}}\left(h^\star\right)$.

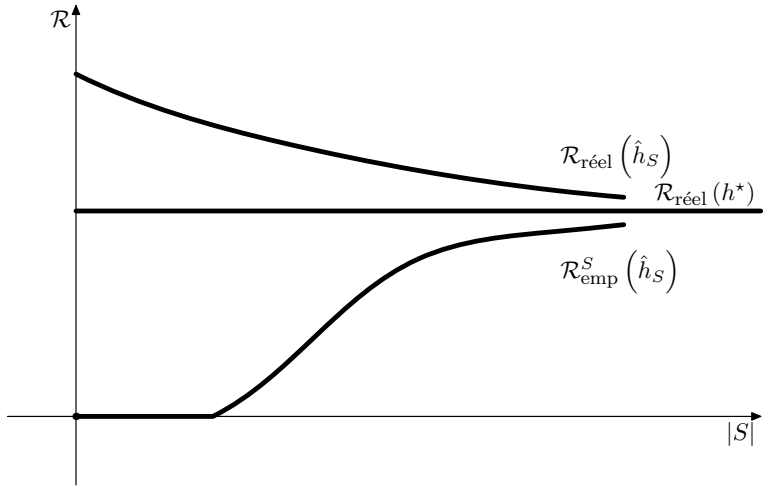

FIGURE 5 – Relation nécessaire entre les risques pour l'analyse PAC.

Cette remarque faite, définissons le problème posé par l'analyse PAC. Soient $\epsilon > 0$ et $\delta > 0$ fixés, trouver $|S|$ suffisamment grand pour assurer :

$$P\left(\left|\mathcal{R}_{\text{réel}}\left(h^\star\right) - \mathcal{R}_{\text{réel}}\left(\hat{h}_S\right)\right| \geq \epsilon\right) < \delta \tag{4}$$

Cela revient à dire que pour une base assez grande, l'expressivité de H étant limitée, on ne pourra plus avoir de sur-apprentissage, et on pourra borner par δ la probabilité que le risque réel de l'hypothèse bâtie sur S s'éloigne de plus de ϵ du biais inductif. Intuitivement, en multipliant la taille de la base d'exemples, sachant qu'ils sont tirés aléatoirement, on diminue le risque de tomber sur des exemples particulièrement favorables[6].

Ce qui est fait dans la littérature est une analyse en pire cas, afin de s'affranchir de la distribution des exemples. Ne reste alors que la relation entre la complexité de H et l'espace X où sont tirés les exemples. Nous présentons une analyse basée sur cette démarche dans la section suivante.

5. *Probably Approximately Correct* en anglais.

6. Exemples pour lesquels on bâtit une hypothèse mauvaise ($\mathcal{R}_{\text{réel}}\left(\hat{h}_S\right)$ trop élevé) sans s'en rendre compte ($\mathcal{R}^S_{\text{emp}}\left(\hat{h}_S\right)$ faible), par sur-apprentissage.

6 Étude de Vapnick

Nous présentons dans cette section l'étude menée par Vapnick sur le cas de la classification [7] ; il s'agit d'une analyse PAC. Cela nous permettra d'introduire les notions de dimension VC et d'ensembles pulvérisés par un espace de fonctions.

6.1 S'affranchir de $\mathcal{D}_{X\times U}$ dans le contrôle du risque

Le but est de borner (par δ) la probabilité d'observer un risque empirique nul pour une hypothèse h sur la base d'exemple S, alors qu'en fait, le risque réel de cette hypothèse h est supérieur à ϵ. Dans ce cas, dont on souhaite donc borner la probabilité, on peut dire que la base d'exemple S est un cas particulièrement favorable, et donc trompeur, car on croit « à tort » que l'hypothèse h convient. Dit de façon plus formelle, on souhaite trouver des conditions pour avoir l'équation [8] 5 :

$$P_{\mathcal{D}_{X\times U}}\left(S \ : \ \exists\, h\in H \ : \ \mathcal{R}^S_{\text{emp}}(h)=0 \text{ et } \mathcal{R}_{\text{réel}}(h)>\epsilon\right)<\delta \tag{5}$$

Ce qui serait pratique, vu que calculer $\mathcal{R}_{\text{réel}}(h)$ est infaisable, ça serait de pouvoir utiliser une base de test $T\in X\times U$, de même taille que S, pour estimer $\mathcal{R}_{\text{réel}}(h)$. Dans ce cas, on élabore toujours h à partir de S, ce qui risque de conduire à un sur-apprentissage, mais on calcule mantenant un risque empirique sur *une autre* base, ce qui intuitivement n'entache pas ce risque-là de problèmes liés à la variance du processus d'apprentissage. Le risque empirique $\mathcal{R}^T_{\text{emp}}(h)$ ainsi mesuré, sur T, devrait être un bon estimateur de $\mathcal{R}_{\text{réel}}(h)$, contrairement à $\mathcal{R}^S_{\text{emp}}(h)$ qui peut être excessivement optimiste en cas de sur-apprentissage.

La question est donc : peut-on se ramener à l'étude de l'équation 6 plutôt que d'utiliser directement 5 qui fait intervenir un risque réel incalculable en pratique ?

$$P_{\mathcal{D}_{X\times U}}\left(S \ : \ \exists\, h\in H \ : \ \mathcal{R}^S_{\text{emp}}(h)=0 \text{ et } \mathcal{R}^T_{\text{emp}}(h)>\epsilon\right)<\delta \tag{6}$$

La réponse est « oui », grâce aux inégalités de Chernoff qui permettent de majorer le membre gauche de l'inégalité 5 par une expression de même nature que le membre gauche de l'inégalité 6. On obtient l'inégalité 7, où ST désigne la base d'exemples obtenir par la concaténation des deux bases d'exemple S et T. On rappelle que $|S|=|T|$.

$$\begin{aligned} & P_{\mathcal{D}_{X\times U}}\left(S \ : \ \exists\, h\in H \ : \ \mathcal{R}^S_{\text{emp}}(h)=0 \text{ et } \mathcal{R}_{\text{réel}}(h)>\epsilon\right) \\ < \ & 2P_{\mathcal{D}_{X\times U}}\left(ST \ : \ \exists\, h\in H \ : \ \mathcal{R}^S_{\text{emp}}(h)=0 \text{ et } \mathcal{R}^T_{\text{emp}}(h)>\frac{\epsilon}{2}\right) \\ < \ & \delta \end{aligned} \tag{7}$$

Vu qu'on est dans un cas de classification en deux classes, le risque réel est l'erreur moyenne de classification, à savoir le nombre d'exemples mal classés divisé par le nombre total d'exemples. Pour avoir $\mathcal{R}^T_{\text{emp}}(h)>\frac{\epsilon}{2}$, il faut que le nombre k d'exemples mal classés soit supérieur à $\epsilon\,|S|\,/2$. Or vu que $\mathcal{R}^S_{\text{emp}}(h)=0$, on peut dire que l'événement mesuré par la deuxième probabilité de l'équation 7 est le cas où, sur un échantillon ST de taille $2\,|S|$, on se retrouve « par hasard [9] » avec tous les exemples bien classés sur la première

7. Classification signifie que U est un ensemble à deux éléments, classe A et classe B, ou usuellement classe -1 ou classe $+1$.

8. : se lit « tel que ».

9. Hasard suivant la distribution $\mathcal{D}_{X\times U}$.

moitié S de la base d'une part, et avec k exemples mal classés sur la deuxième moitié T de la base d'autre part.

C'est là qu'intervient l'analyse en pire cas. Bien que le hasard dont on parle soit « piloté » par $\mathcal{D}_{X\times U}$, on ne va pas en tenir compte et simplement réaliser un dénombrement, sans profiter de la distribution. On peut montrer que la probabilité que, sachant que k exemples sont mal classés dans ST, ces k exemples tombent tous dans T est inférieure à $1/2^k$. Donc pour une hypothèse h, regardons tous les cas possibles. L'hypothèse h réalise une dichotomie sur ST, à savoir qu'elle différencie ceux qu'elle classe correctement et ceux qu'elle classe mal. Parmi les dichotomies qu'un h de H est capable de réaliser, certaines d'entre elles correspondent à notre problème, à savoir qu'elles mettent tous les exemples de S du côté « bien classé », ainsi que tous les exemples de T, sauf les k qui sont mal classés. Et encore, on ne s'intéresse qu'aux k supérieurs $\epsilon\left|S\right|/2$. Pour chacun de ces derniers, on peut majorer la probabilité que l'événement se produise par $1/2^{\epsilon|S|/2}$, car plus k est grand, plus cette probabilité est faible. Donc chacune des dichotomies qui correspond à notre problème à une probabilité de $1/2^{\epsilon|S|/2}$ de se réaliser. La question est alors de savoir quel est le nombre de ces dichotomies-là. Ne sachant le trouver, on va le majorer brutalement par le nombre totale de dichotomies que h peut réaliser sur ST. Ce nombre lui-même est majoré par une grandeur notée $G_H\left(2\left|S\right|\right)$ qui est le nombre maximal de dichotomie qu'une fonction de H puisse réaliser sur un ensemble d'exemples de taille $2\left|S\right|$.

Le résultat de cette majoration brutale est que la probabilité d'observer un risque empirique nul sur S tout en ayant plus de k erreurs sur T est majorée par $1/2^k \times G_H\left(2\left|S\right|\right)$. Comme nous considérons $k = \epsilon\left|S\right|/2$, on peut rajouter une ligne aux inéquations 7, ce qui donne les inéquations 8.

$$\begin{aligned} & P_{\mathcal{D}_{X\times U}}\left(S \ : \ \exists\, h\in H \ : \ \mathcal{R}^S_{\text{emp}}\left(h\right)=0 \text{ et } \mathcal{R}_{\text{réel}}\left(h\right)>\epsilon\right) \\ < \ & 2P_{\mathcal{D}_{X\times U}}\left(ST \ : \ \exists\, h\in H \ : \ \mathcal{R}^S_{\text{emp}}\left(h\right)=0 \text{ et } \mathcal{R}^T_{\text{emp}}\left(h\right)>\frac{\epsilon}{2}\right) \\ < \ & 2G_H\left(2\left|S\right|\right)2^{-\frac{\epsilon|S|}{2}} \\ < \ & \delta \end{aligned} \tag{8}$$

On voit apparaître une majoration qui dépend d'un terme $G_H\left(2\left|S\right|\right)$, qui est lié au pouvoir de séparation des éléments de l'espace d'hypothèse H au regard de X. C'est toute la subtilité de cette analyse que d'utiliser cette notion pour s'affranchir de la distribution effective des exemples et de leurs étiquettes.

6.2 Fonction de croissance et dimension VC

L'expression $G_H\left(l\right)$ désigne le nombre maximal de dichotomies qu'une hypothèse de H peut réaliser sur une base d'exemple de taille l. On l'appelle *fonction de croissance* de H. On a bien évidemment $G_H\left(l\right) \leq 2^l$, puisque 2^l est le nombre de dichotomies qu'on peut réaliser sur un ensemble de taille l.

Par exemple, soit $X = \mathbb{R}^2$ et H les fonctions linéaires, qui à $(x, y) \in \mathbb{R}^2$ associent signe(ax + by + c). Prenons un ensemble de 3 points x, y, z de X, et choisissons de leur affecter un étiquetage dichotomique arbitraire $+1$ où -1. On peut trouver un $h \in H$, c'est-à-dire un jeu de coefficients a, b, c, tel que les valeurs $h(x), h(y), h(z)$ correspondent à notre étiquetage. Autrement dit, pour cet ensemble de trois points, tout étiquetage est réalisable par au moins une des fonctions de H (cf. figure 6). On dit alors que cet ensemble de 3 points est *pulvérisé*[10] par H. C'est en effet vrai pour tout ensemble de trois points.

10. *shattered* en anglais.

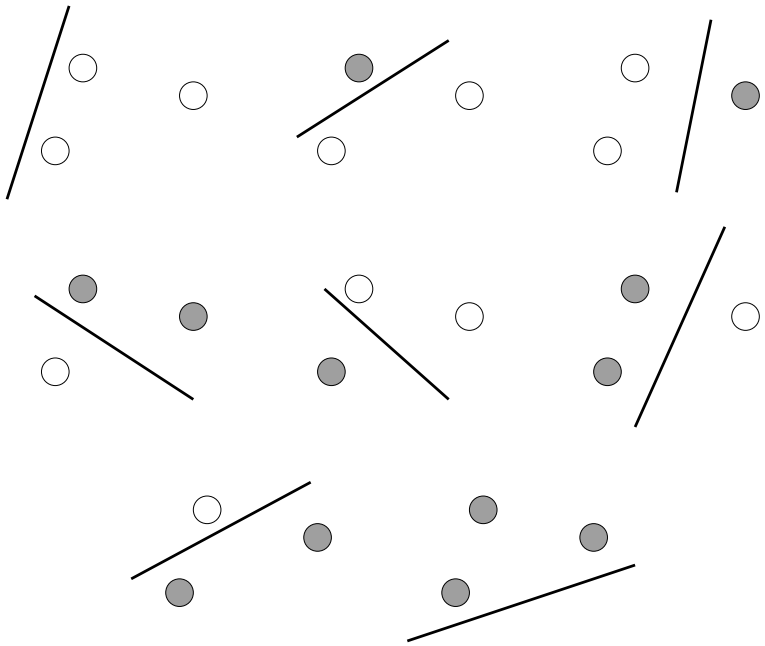

FIGURE 6 – Quel que soit l'étiquetage de ces trois points, on peut trouver une ligne qui les sépare. Cet ensemble est donc pulvérisé par l'ensemble des droites.

En revanche, comme l'illustre la figure 7, il existe des ensembles de 4 points pour lesquels une des dichotomies possible n'est pas réalisable par H.

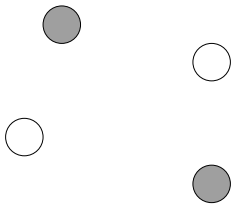

FIGURE 7 – On ne peut séparer ces point par une droite qui respecte l'étiquetage.

Pour en revenir à la notion de fonction de croissance, dire que $G_H(l) = 2^l$ signifie qu'il existe au moins un ensemble S tel que $|S| = l$ qui soit pulvérisé par H. On appelle *dimension de Vapnick-Chervonenkis* d'un espace d'hypothèses H, que l'on note $d_{VC}(H)$, le cardinal du plus grand ensemble pulvérisable par H. Cela revient à utiliser l'équation 9 pour définir cette notion :

$$d_{VC}(H) = \max_l \left\{ l \in \mathbb{N} \ : \ G_H(l) = 2^l \right\} \tag{9}$$

Si on reprend l'exemple où H est l'ensemble des séparateurs linéaires de $\mathbb{R}^2$, on a $d_{VC}(H) = 3$ car aucun des ensembles de 4 points n'est pulvérisable. De façon générale, si H est l'ensemble des séparateurs linéaires de $\mathbb{R}^d$, on a $d_{VC}(H) = d + 1$.

Pour espérer borner le risque de sur-apprentissage non détecté sur S via l'inéquation 8, il devient crucial que la croissance de $G_H(l)$ soit sub-exponentielle, ce qu'illustre la figure 8.

Le lemme de Sauer, donné par l'équation 10 assure que la croissance de $G_H(l)$ est polynomiale en l.

$$G_H(l) \leq \left(\frac{el}{d_{VC}(H)} \right)^{d_{VC}(H)} = \mathcal{O}\left(l^{d_{VC}(H)} \right) \tag{10}$$

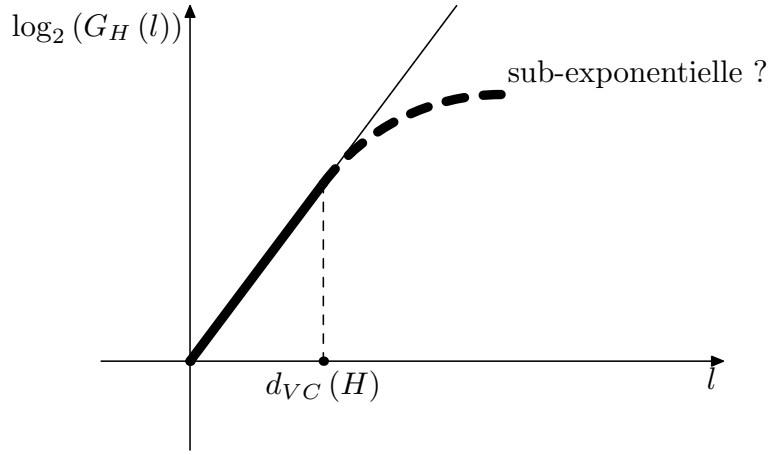

FIGURE 8 – La croissance de $G_H(l)$ est-elle sub-exponentielle ?

Ce résultat nous permet de poursuivre la majoration que nous avions établie par les inéquations 8, pour arriver à la majoration nous permettant de calculer des bornes, majoration exprimée par les inéquations 11.

$$\begin{aligned} & P_{\mathcal{D}_{X\times U}}\left(S \,:\, \exists\, h\in H \,:\, \mathcal{R}^S_{\text{emp}}(h)=0 \text{ et } \mathcal{R}_{\text{réel}}(h)>\epsilon\right) \\ < \;& 2P_{\mathcal{D}_{X\times U}}\left(ST \,:\, \exists\, h\in H \,:\, \mathcal{R}^S_{\text{emp}}(h)=0 \text{ et } \mathcal{R}^T_{\text{emp}}(h)>\frac{\epsilon}{2}\right) \\ < \;& 2G_H(2|S|)2^{-\frac{\epsilon|S|}{2}} \\ < \;& 2\left(\frac{2e|S|}{d_{VC}(H)}\right)^{d_{VC}(H)} 2^{-\frac{\epsilon|S|}{2}} \\ = \;& \delta \end{aligned} \tag{11}$$

Ce résultat permet de faire le calcul suivant. Si on reprend l'exemple où H est l'ensemble des séparateurs linéaires de $\mathbb{R}^2$. On a vu que $d_{VC}(H)=3$. Posons $\epsilon=1\%$ et $\delta=5\%$. la seule inconnue de l'égalité de l'équation 11 est $|S|$, et on trouve $|S|=2425$. On peut alors en conclure que, dans le cadre de la séparation linéaire, un classifieur qui ne fait pas d'erreurs sur 2425 exemples a un risque réel d'erreur inférieur à 1% (c'est le ϵ). Mais attention, en formulant cette affirmation, on a 5% (c'est δ) de chances de dire une ânerie.

6.3 Théorèmes de Vapnick et Chevonenkis

De façon générale, et c'est un théorème établi par Vapnick et Chervonenkis, on peut dire, au risque δ de se tromper, que toute hypothèse $h\in H$ consistante avec une base d'exemples S a un risque réel plus petit que

$$\frac{2}{|S|}\left(d_{VC}(H)\log\frac{2e|S|}{d_{VC}(H)}+\log\frac{2}{\delta}\right) \text{ , si } |S|\geq d_{VC}(H) \text{ et } |S|>\frac{2}{\epsilon}$$

Ce théorème s'assouplit un peu : on peut dire, au risque δ de se tromper, que toute hypothèse $h\in H$ qui fait k erreurs sur une base d'exemples S a un risque réel plus petit que

$$\frac{2k}{|S|}+\frac{4}{|S|}\left(d_{VC}(H)\log\frac{2e|S|}{d_{VC}(H)}+\log\frac{4}{\delta}\right) \text{ , si } |S|\geq d_{VC}(H)$$

Il existe dans la littérature des théorèmes du même genre, relatifs au cas particulier où

H est l'ensemble des machines à vecteur supports [11]. Dans ce cas, les marges, ainsi que le vecteur des *slack variables* interviennent dans le calcul de la borne.

C'est à partir de ce type d'analyse que l'on peut définir le terme $\Omega(H_d)$ qui apparaît dans l'induction par la minimisation du risque structurel, vu au paragraphe 4.3.2, page 7.

7 Étude empirique, validation croisée

Plutôt que de trouver des bornes au risque réel d'une hypothèse, on peut chercher à estimer ce risque par des mesures empiriques. C'est le principe de la validation croisée, très utilisée pour les réseaux de neurones et SVM. On dispose d'une base d'exemples S. On la divise en n parties égales, notés $\{S_i\}_{1\leq i\leq n}$. Notons également $\overline{S}_i = S \setminus S_i$ les exemples qui ne sont pas dans S_i. La validations croisée (dite à n plis) consiste à procéder comme suit pour estimer le risque réel. Pour chacune de ces parties S_i, on réalise un apprentissage sur $\overline{S}_i$, qui donne lieu à l'élaboration d'une hypothèse $\hat{h}_{\overline{S}_i}$. On calcule le risque empirique $R_i = \mathcal{R}_{\text{emp}}^{S_i}\left(\hat{h}_{\overline{S}_i}\right)$ de $\hat{h}_{\overline{S}_i}$ sur les exemples de S_i, qui n'ont pas été utilisés pour l'apprentissage. On obtient ainsi, pour chaque apprentissage sur $\overline{S}_i$, une estimation de $\mathcal{R}_{\text{réel}}\left(\hat{h}_{\overline{S}_i}\right)$. On fait alors la moyenne m de ces n estimations $\{R_i\}_{1\leq i\leq n}$, et on réalise un apprentissage sur toute la base. Le principe de validation croisée dit que l'hypothèse $\hat{h}_S$ ainsi calculée sur toute la base a un risque réel qui est correctement estimé par m.

Références

[1] Antoine Cornuéjols, Laurent Miclet, and Yves Kodratoff. *Apprentissage artificiel, Concepts et algorithmes*. Eyrolles, 2002. ISBN 2-212-11020-0.

[2] Nello Cristanini and John Shawe-Taylor. *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, 2000.

11. Concept non décrit dans ce document.