Apprentissage Statistique
M2 Probabilités et Statistiques, Université Paris-Sud

Cours 1 : Théorie de l'apprentissage statistique: de Vapnik à la localisation (1/2)

SYLVAIN ARLOT ET FRANCIS BACH
NOTES DE COURS INITIALEMENT PRISES PAR THOMAS PUMIR ET ELODIE VERNET (2012)

Table des matières

1. Théorie de l'apprentissage : généralités 1
1.1. Cadre général de la prédiction 1
1.2. Régression 3
1.3. Classification (binaire supervisée) 3
1.4. Estimateur, notion de consistance 5
1.5. Quelques règles de classification classiques 6
1.6. No free lunch theorem : on n'a rien sans rien 8
2. Minimisation du risque structurel 9
2.1. Minimisation du risque empirique 9
2.2. Analyse du risque empirique sur un modèle 10
2.3. Choix de modèles 11
Références 13

1. Théorie de l'apprentissage : généralités

1.1. **Cadre général de la prédiction.** On dispose de n observations $D_n = (X_i, Y_i)_{1 \leqslant i \leqslant n} \in (\mathcal{X} \times \mathcal{Y})^n$. Ces observations sont n réalisations indépendantes d'une variable aléatoire (X, Y) de loi commune P.

Si on se donne une nouvelle réalisation (X_{n+1}, Y_{n+1}) de la variable aléatoire (X, Y), l'objectif est de prédire Y_{n+1} sachant X_{n+1} en se trompant aussi rarement que possible.

Classiquement, $\mathcal{X} = \mathbb{R}^p$. Chacune des coordonnées de la variable explicative $X \in \mathcal{X}$ est alors une variable décrivant l'objet d'étude.

La variable d'intérêt $Y \in \mathcal{Y}$ décrit une caractéristique de l'objet d'étude ; on appelle Y l'étiquette associée à X. L'ensemble $\mathcal{Y}$ peut être discret ou continu. Dans le cas discret, si $\mathcal{Y} = \{0, 1\}$ on parle de classification binaire, si $\mathcal{Y} = \{0, 1, ..., m\}$ avec $m \in \mathbb{N}$ on parle de classification multivariée.

Date: 2 Février 2015.

Exemple 1. *Dans le cas de l'aide au diagnostic médical*
- *X représente l'ensemble des paramètres observables (âge, taille, résultats d'examens médicaux...)*
- *Y représente l'étiquette associée au patient.*

  Par exemple : $Y = \begin{cases} 0 & \text{si le patient est sain,} \\ 1 & \text{si le patient est malade.} \end{cases}$

Il existe plusieurs types d'apprentissage :
- Apprentissage supervisé : on a observé tous les Y_i. C'est le cadre que l'on considérera dans la suite du cours.
- Apprentissage semi-supervisé : on observe quelques Y_i.
- Apprentissage non-supervisé : on ne connaît aucun Y_i.

Définition 1. *On appelle prédicteur/classifieur toute application mesurable $t : \mathcal{X} \to \mathcal{Y}$. L'ensemble des prédicteurs/classifieurs est noté $\mathbb{S}$.*

Le but d'un classifieur est de fournir une étiquette $t(X_{n+1})$ à X_{n+1}. On espère bien évidemment faire coïncider $t(X_{n+1})$ et Y_{n+1}, c'est-à-dire qu'on cherche le meilleur classifieur. Or, pour parler de meilleur classifieur, il est nécessaire de définir une mesure de la qualité du classifieur. On mesure cette qualité à l'aide d'une « fonction de contraste ». La fonction de perte correspondante, qui est l'espérance de la fonction de contraste, mesure alors la qualité du classifieur.

Définition 2. *On appelle fonction de contraste toute fonction γ de la forme*

$$\begin{aligned} \gamma : \quad & \mathbb{S} \times (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R} \\ & (t, (x, y)) \mapsto \gamma(t, (x, y)) \ . \end{aligned}$$

L'objectif est désormais de trouver $t \in \mathbb{S}$ tel que $\gamma(t; (X_{n+1}, Y_{n+1}))$ est minimal « en moyenne », ce qui signifie minimiser la fonction de perte définie comme suit.

Définition 3. *La fonction de perte $P\gamma$ associée à la fonction de contraste γ est définie par*

$$P\gamma(t) = P\gamma(t; \cdot) = \mathbb{E}_{(X,Y)\sim P}\left[\gamma(t, (X, Y))\right]$$

pour tout $t \in \mathbb{S}$.

On remarquera que si t est aléatoire (ce qui est le cas quand t dépend de l'échantillon D_n), $P\gamma(t)$ l'est aussi.

Définition 4. *On appelle prédicteur de Bayes tout prédicteur s^* qui minimise la fonction de perte :*

$$s^* \in \operatorname{argmin}_{t\in\mathbb{S}} P\gamma(t) \ .$$

La valeur minimale de la perte

$$\inf_{t\in\mathbb{S}} P\gamma(t) = P\gamma(s^*)$$

est appelée risque de Bayes.

Sachant qu'on ne peut pas trouver un meilleur prédicteur que le prédicteur de Bayes, le but est alors de s'approcher au plus du prédicteur de Bayes au sens de minimiser la « perte relative » :

Définition 5. *On appelle perte relative d'un prédicteur t pour une fonction de contraste γ la quantité $\ell(s^*, t) = P\gamma(t) - P\gamma(s^*)$.*

Par définition du prédicteur de Bayes, la perte relative est toujours positive et est nulle pour $t = s^*$.

1.2. **Régression.** Dans le cas de la régression on considère que : $\mathcal{Y} = \mathbb{R}$. Ici, Y est donc continue.

On peut toujours écrire que Y et X sont reliés par la relation

$$Y = \eta(X) + \varepsilon \qquad \text{avec} \qquad \eta(X) = \mathbb{E}[Y|X] \ .$$

Ceci implique que $\mathbb{E}[\varepsilon|X] = 0$ p.s. ; η est appelée fonction de regression.

Définition 6. *On définit le contraste des moindres carrés par*

$$\gamma(t, (x, y)) = (t(x) - y)^2 \ .$$

Proposition 1. *Si $\mathcal{Y} = \mathbb{R}$ et γ est le contraste des moindres carrés, alors, pour tout $t \in \mathbb{S}$,*

$$P\gamma(t) = \mathbb{E}[(t(X) - \eta(X))^2] + P\gamma(\eta) \geqslant P\gamma(\eta)$$

si bien que $\eta = s^*$ est un prédicteur de Bayes et que la perte relative s'écrit

$$\ell(s^*, t) = \mathbb{E}((t(x) - \eta(X))^2) \ .$$

Démonstration. Pour tout $t \in \mathbb{S}$,

$$\begin{aligned} P\gamma(t) &= \mathbb{E}[(t(X) - Y)^2] \\ &= \mathbb{E}[(t(X) - \eta(X) - \varepsilon)^2] \\ &= \mathbb{E}[(t(X) - \eta(X))^2] + \mathbb{E}[\varepsilon^2] - 2\mathbb{E}[\mathbb{E}[\varepsilon(t(X) - \eta(X))|X]] \ . \end{aligned}$$

Or

$$\mathbb{E}\left[\mathbb{E}\left[\varepsilon(t(X) - \eta(X))|X\right]\right] = \mathbb{E}\Big[\big(t(X) - \eta(X)\big)\underbrace{\mathbb{E}[\varepsilon|X]}_{=0}\Big] = 0 \ ,$$

ce qui prouve le premier résultat. La suite de la proposition s'en déduit directement. □

On remarque que $\ell(s^*, t) = \|t - \eta\|^2_{\mathbb{L}^2(P_X)}$. Minimiser la perte relative revient donc à un problème d'estimation de η en norme $\mathbb{L}^2(P_X)$.

1.3. **Classification (binaire supervisée).** Dans le cas de la classification binaire (supervisée) on a : $\mathcal{Y} = \{0, 1\}$.

Exemples :

– Reconnaissance d'un objet

$\mathcal{X}$ = ensemble de caractéristique des pixels d'une image

$$\mathcal{Y} = \begin{cases} 1 & \text{si il y a présence d'un objet particulier sur l'image (voiture...)} \\ 0 & \text{si il y a absence d'un tel objet ;} \end{cases}$$

– catégorisation de textes
$\mathcal{X}$ = suite des caractères d'un texte
$\mathcal{Y} = \begin{cases} 1 & \text{si le texte appartient à une catégorie donnée} \\ 0 & \text{si le texte n'appartient pas à la catégorie ;} \end{cases}$
– détection de spams
$\mathcal{X}$ = suite des caractères d'un mail
$\mathcal{Y} = \begin{cases} 1 & \text{si le mail est un spam} \\ 0 & \text{si le mail n'est pas un spam.} \end{cases}$

Exemples de fonctions de contrastes en classification binaire :
– contraste 0–1

$$\gamma_{0-1}(t,(x,y)) = \gamma(t,(x,y)) = \mathbb{1}_{t(x)\neq y}$$

– contraste asymétrique : étant donné $w = (w_0, w_1) \in \mathbb{R}^2$

$$\gamma_w(t,(x,y)) = w_0 \mathbb{1}_{t(x)\neq 0, y=0} + w_1 \mathbb{1}_{t(x)\neq 1, y=1}$$

Le contraste asymétrique est une généralisation du contraste 0–1. Il peut éventuellement pénaliser une erreur plus que l'autre. Par exemple, il est utile dans le cas de la détection de spam : il est plus grave de classifier un e-mail « normal » en tant que spam qu'un spam en tant qu'e-mail « normal ».

Proposition 2. *On considère le contraste 0-1, $\gamma(t,(x,y)) = \mathbb{1}_{t(x)\neq y}$.*
(On rappelle que $\eta(X) = \mathbb{E}[Y|X] = \mathbb{P}(Y = 1|X)$.)
Le classifieur de Bayes est alors $s^*(X) = \mathbb{1}_{\eta(X)\geqslant\frac{1}{2}}$ p.s. sauf sur $\{\eta(X) = \frac{1}{2}\}$.
Alors $P\gamma(s^\star) = \mathbb{E}[\eta(X) \wedge (1-\eta(X))]$. Et pour tout $t \in \mathbb{S}$, $\ell(s^\star, t) = \mathbb{E}\left[|2\eta(X) - 1|\, \mathbb{1}_{t(X)\neq s(X)}\right]$

Démonstration.

$$P\gamma(t) = \mathbb{P}(t(X) \neq Y) = \mathbb{E}[\mathbb{P}(t(X) \neq Y)|X]$$

Si $\eta(X) \geqslant \frac{1}{2}$ et si $t(X) = 0$, $P(0 \neq Y|X) = \eta(X) \geqslant \frac{1}{2}$.
De plus si $\eta(X) \geqslant \frac{1}{2}$ et si $t(X) = 1$, $\mathbb{P}(1 \neq Y|X) = 1 - \eta(X) \leqslant \frac{1}{2}$.
Ainsi si $\eta(X) \geqslant \frac{1}{2}$, $s^\star(X) = 1$ et $\mathbb{P}(s^\star(X) \neq Y|X) = \eta(X) \wedge (1-\eta(X))$.
De même, si $\eta(X) < \frac{1}{2}$, $s^\star(X) = 0$ et $\mathbb{P}(s^\star \neq Y|X) = \eta(X) \wedge (1-\eta(X))$.

On a ainsi défini l'estimateur de Bayes à un ensemble de mesure nulle près : $s^\star(X) = \mathbb{1}_{\eta(X)\geqslant\frac{1}{2}}$ et $P\gamma(s^\star) = \mathbb{E}[\eta(X) \wedge (1-\eta(X))]$.

De plus, $\ell(s^\star, t) = \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{(t(X)\neq Y} - \mathbb{1}_{s^\star(X)\neq Y}|X\right]\right]$.
Or, si $\eta(X) \geqslant \frac{1}{2}$, alors $s^\star(X) = 1$ et $\mathbb{E}\left[\mathbb{1}_{t(X)\neq Y} - \mathbb{1}_{s^\star(X)\neq Y}|X\right] = \mathbb{1}_{t(X)\neq s^\star(X)}(\eta(X) - (1-\eta(X))$.
Et si $\eta(X) < \frac{1}{2}$, alors $s^\star(X) = 0$ et $\mathbb{E}\left[\mathbb{1}_{t(X)\neq Y} - \mathbb{1}_{s^\star(X)\neq Y}|X\right] = \mathbb{1}_{t(X)\neq s^\star(X)}(1 - \eta(X) - \eta(X))$.
D'où $\ell(s^\star, t) = \mathbb{E}\left[|2\eta(X) - 1|\, \mathbb{1}_{t(X)\neq s(X)}\right]$. □

Exercice 1. *Déterminer le classifieur de Bayes, $P\gamma(s^\star)$, et $\ell\,(s^\star, t\,)$ dans le cas du contraste asymétrique $w_y \mathbb{1}_{t(x)\neq y}$.*

1.4. **Estimateur, notion de consistance. Définitions**

– On appelle *estimateur* $\widehat{s}$ toute application mesurable qui à un n-échantillon associe un classifieur :

$$\widehat{s} : (\mathcal{X} \times \mathcal{Y})^n \to \mathbb{S}.$$

– On appelle *algorithme d'apprentissage* (ou règle d'apprentissage) $\widehat{s}$ toute application mesurable qui pour tout n, à un n-échantillon associe un classifieur :

$$\widehat{s} : \bigcup_{n\geqslant 1} (\mathcal{X} \times \mathcal{Y})^n \to \mathbb{S}.$$

On notera D_n un n-échantillon $(X_i, Y_i)_{i=1..n}$. Et pour simplifier les notations, on confondra souvent l'application $\widehat{s}$ et sa valeur en un n-échantillon en notant $\widehat{s} = \widehat{s}(D_n) = \widehat{s}((X_i, Y_i)_{i=1..n})$. On remarquera que la perte $P\gamma(\widehat{s})$ est aléatoire :

$P\gamma(\widehat{s}) = \mathbb{E}\,[\,\gamma(\widehat{s}, (X, Y))|D_n\,] = \mathbb{E}\,[\,\gamma(\widehat{s}, (X, Y))|(X_i, Y_i)_{i=1..n}\,]$.

On appelle le risque l'espérance de la perte $\mathbb{E}\,[\,P\gamma(\widehat{s}(D_n))\,]$ et l'excès de risque l'espérance de la perte relative $\mathbb{E}\,[\,\ell\,(s^\star, \widehat{s}(D_n)\,)\,]$.

On définit ensuite différents types de consistance.

Définition 7. *On dit qu'il y a :*

– *consistance faible pour une loi P si le risque tend vers 0 quand la taille de l'échantillon tend vers l'infini :*

$$\lim_{n\to+\infty} \mathbb{E}\,[\,\ell\,(s^\star, \widehat{s}(D_n)\,)\,] = 0$$

– *consistance forte pour une loi P si la perte relative tend vers 0 $P^{\otimes n}$p.s. quand la taille de l'échantillon tend vers l'infini :*

$$\ell\,(s^\star, \widehat{s}(D_n)\,) \to 0 \ P^{\otimes n}\text{p.s.}$$

– *consistance universelle faible si il y a consistance faible pour toute loi :*

$$\forall P \text{ loi sur } \mathcal{X} \times \mathcal{Y}, \ \lim_{n\to+\infty} \mathbb{E}_{D_n\sim P^{\otimes n}}\,[\,\ell\,(s^\star, \widehat{s}(D_n)\,)\,] = 0$$

– *consistance universelle forte si il y a consistance forte pour toute loi :*

$$\forall P \text{ loi sur } \mathcal{X} \times \mathcal{Y}, \ \ell\,(s^\star, \widehat{s}(D_n)\,) \to 0 \ P^{\otimes n}\text{p.s.}$$

– *consistance universelle faible uniforme si le risque converge uniformément en les lois de probabilité vers 0 :*

$$\sup_{P \text{ loi sur } \mathcal{X}\times\mathcal{Y},} \mathbb{E}\,[\,\ell\,(s^\star, \widehat{s}(D_n)\,)\,] \to 0.$$

On remarquera que la consistance universelle faible uniforme implique la consistance universelle faible. De plus, la consistance universelle faible implique la consistance faible. De même, la consistance universelle forte implique la consistance forte.

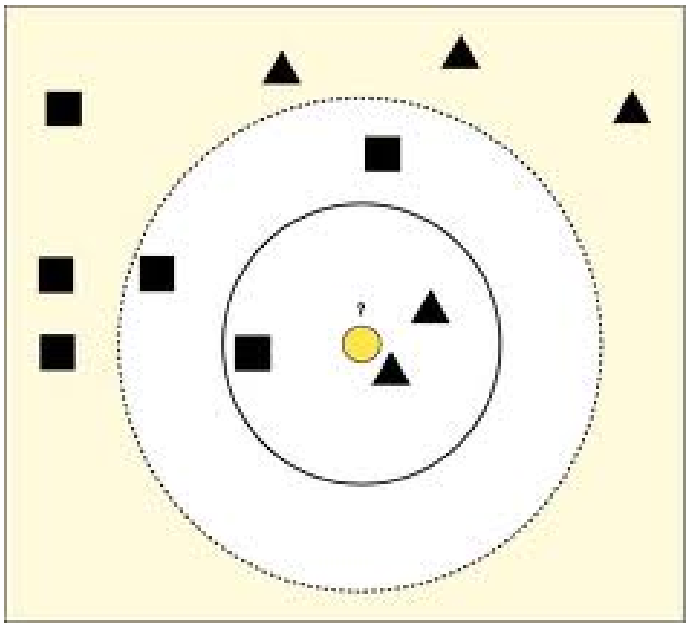

FIGURE 1. En jaune, point à classer grâce au k plus proches voisins. Pour $k = 3$, ce point serait classé dans la classe des triangles alors que pour $k = 5$, il serait classé dans la classe des carrés

1.5. **Quelques règles de classification classiques.**

– algorithme des k plus proches voisins

On travaille alors avec $\mathcal{X} = \mathbb{R}^p$ (normée). Supposons que nous avons un n-échantillon $(X_i, Y_i)_{i=1...n}$. Cette règle consiste à considérer les k-plus proches voisins (k-ppv) d'un point x, puis de comparer le nombre d'étiquettes 1 et 0 pour les Y_i parmi les k-plus proches voisins. On note $\hat{\eta}(x)$ le nombre moyen de Y_i ayant l'étiquette 1 parmi les k plus proches voisins de x

$$\hat{\eta}(x) = \frac{1}{k} \sum_{X_i \in \text{ k-ppv de x}} Y_i.$$

S'il y a plus d'étiquettes 1, on classera x en classe 1 et sinon on le classera dans la classe 0. Ainsi

$$\widehat{s} = \mathbb{1}_{\hat{\eta} \geqslant \frac{1}{2}}.$$

On remarquera la ressemblance de cette règle de classification avec celle de Bayes pour le contraste 0-1 ($s^\star = \mathbb{1}_{\eta \geqslant \frac{1}{2}}$). On remplace η par un estimateur de celui-ci : $\hat{\eta}$: ceci s'appelle un « plug-in ».

Théorème 3 (Stone)**.** *Avec $\mathbb{X} = \mathbb{R}^p$. On suppose que $(k_n)_n$ est une suite d'entiers naturels divergeant vers $+\infty$ et que $k_n/n \to 0$.*

Il y a alors consistance universelle des k_n plus proches voisins.

La règle des plus proches voisins est un exemple de règle par moyennage local.

– Classification par moyennage local

On appelle règle par moyennage local, une règle de la forme

$$\widehat{s} = \mathbb{1}_{\hat{\eta} \geqslant \frac{1}{2}},$$

où

$$\hat{\eta}(x) = \sum_i w_i(x) Y_i$$

avec $w_i(x)$ petit si x est loin de X_i ($w_i(x)$ est le poids de l'étiquette Y_i dans la moyenne), $w_i \geqslant 0$ et $\sum_i w_i = 1$.

– Par exemple avec $w_i(x) = \begin{cases} \frac{1}{k} & \text{si } X_i \in \text{ k-ppv de } x \\ 0 & \text{sinon} \end{cases}$, on retrouve la règle des k-plus proches voisins.

– un autre exemple est la règle par partition. Soit $\left((A_{k,n})_{k \in \mathbb{N}} \right)_{n \in \mathbb{N}}$ une suite de partitions de $\mathcal{X}$ (i.e. $\cup_{k \in \mathbb{N}} A_{k,n} = \mathcal{X}$ et $A_{k,n} \cap A_{j,n} = \oslash$ pour $k \neq j$). On note $A_n(x)$ l'élément de la n-ième partition contenant x et $N_n(x) = \text{Card}\{j \in \{1, \ldots, n\} \, : \, X_j \in A_n(x)\}$. Pour n donné, on considère les poids

$$w_i(x) = \frac{\mathbb{1}_{X_i \in A_n(x)}}{N_n(x)}$$

Lorsque $\mathbb{X} = \mathbb{R}^p$, cette règle est universellement consistante si $\text{diam}(A_n(X)) \to 0$ et $N_n(X) \to +\infty$ en probabilité lorsque $n \to +\infty$.

– pour la règle par noyaux (ou règle de Nadaraya-Watson), associée à un noyau $K : \mathcal{X} \mapsto [0; +\infty[$ et une fenêtre $h > 0$, les poids sont

$$w_i(x) = \frac{K(\frac{x - X_i}{h})}{\sum_{j=1}^n K(\frac{x - X_j}{h})}$$

– Une autre règle de classification consiste à minimiser le risque empirique.

Définitions

– On note P_n la mesure empirique associée à l'échantillon D_n :

$$P_n = \frac{1}{n} \sum_{i=1}^n \delta_{(X_i, Y_i)}$$

– Le risque empirique est l'application :

$$P_n\gamma(t) : \begin{cases} \mathbb{S} & \to \mathbb{R} \\ t & \mapsto P_n\gamma(t) = P_n(\gamma(t, .)) = \frac{1}{n} \sum_{i=1}^n \gamma(t, (X_i, Y_i)) \end{cases}$$

Soit $S \subset \mathbb{S}$ un sous ensemble de prédicteurs, on appelle S un modèle. On appelle le minimiseur du risque empirique $s_S^\star$ (ERM : Empirical Risk Minimizer) sur le modèle S, l'estimateur qui minimise le risque empirique sur S :

$$s_S^\star \in \text{argmin}_{t \in S} \, P_n\gamma(t)$$

1.6. **No free lunch theorem : on n'a rien sans rien.** Le théorème suivant montre que dans le cas de la classification il n'est pas possible d'avoir une consistance universelle faible uniforme pour le contraste 0-1 sur l'ensemble des lois sur $\mathcal{X} \times \mathcal{Y}$ lorsque $\mathcal{X}$ est infini.

Théorème 4. *Si $\mathcal{X}$ est infini, $\mathcal{Y} = \{0,1\}$, $\gamma = \gamma_{0-1}$, alors pour tout entier $n \in \mathbb{N}$ et pour tout estimateur $\widehat{s} : (\mathcal{X}, \mathcal{Y})^n \to \mathbb{S}$*

$$\sup_{P \text{ loi sur } \mathcal{X}\times\mathcal{Y}} \{ \mathbb{E}_{D_n \sim P^{\otimes n}} [\ell (s^\star, \widehat{s}(D_n))] \} \geqslant \frac{1}{2} \ .$$

Démonstration. Soit un entier $K \geqslant 1$ et $A_1, \dots, A_K \in \mathcal{X}$. Pour simplifier, on suppose $A_i = i$ pour tout i.

Soit $r \in \{0,1\}^K$ fixé. On définit une loi P_r sur $\mathcal{X} \times \mathcal{Y}$ comme suit : $(X,Y) \sim P_r$ si et seulement si X suit une loi uniforme sur l'ensemble $\{1, \dots, K\}$ et $Y = r_X$ est une fonction de X uniquement. Ainsi, sous la loi P_r, $s^\star(X) = s_r^\star(X) = r_X$ et $P_r \gamma(s_r^\star) = 0$.

On écrit alors que

$$\begin{aligned}
&\sup_{P \text{ loi sur } \mathcal{X}\times\mathcal{Y}} \{ \mathbb{E}_{D_n \sim P^{\otimes n}} [\ell (s^\star, \widehat{s}(D_n))] \} \\
&\geqslant \sup_{r \in \{0,1\}^K} \left\{ \mathbb{E}_{D_n \sim P_r^{\otimes n}} \left[\ell_r \big(s_r^\star, \widehat{s}(D_n) \big) \right] \right\} \\
&= \sup_{r \in \{0,1\}^K} \left\{ \mathbb{P}_{D_n \sim P_r^{\otimes n}, (X,Y) \sim P_r} \big(\widehat{s}(D_n; X) \neq Y \big) \right\} \\
&\geqslant \mathbb{P}_{r \sim R, D_n \sim P_r^{\otimes n}, (X,Y) \sim P_r} \big(\widehat{s}(D_n; X) \neq Y \big)
\end{aligned}$$

où R est une loi quelconque sur $\{0,1\}^K$. Réécrivons la dernière probabilité écrite afin de pouvoir échanger l'ordre d'intégration (c'est-à-dire, prendre d'abord une moyenne vis-à-vis de r, et ensuite moyenner par rapport aux X_i et à X :

$$\begin{aligned}
&\mathbb{P}_{r \sim R, D_n \sim P_r^{\otimes n}, (X,Y) \sim P_r} \big(\widehat{s}(D_n; X) \neq Y \big) \\
&= \mathbb{P}_{r \sim R, X_1, \dots, X_n, X \sim \mathcal{U}(\{1,\dots,K\})} \big(\widehat{s}((X_i, r_{X_i})_{1 \leqslant i \leqslant n}; X) \neq r_X \big) \\
&= \mathbb{E}_{X_1, \dots, X_n, X \sim \mathcal{U}(\{1,\dots,K\})} \big[\mathbb{P}_{r \sim R} \big(\widehat{s}((X_i, r_{X_i})_{1 \leqslant i \leqslant n}; X) \neq r_X | X_1, \dots, X_n, X \big) \big]
\end{aligned}$$

On s'intéresse désormais à la probabilité sachant $X_1, \dots, X_n, X$ écrite ci-dessus. Il s'agit de la probabilité que l'on ait une certaine fonction de $(X, X_1, \dots, X_n, r_{X_1}, \dots, r_{X_n})$ égale à r_X. On souhaite choisir R telle que cette probabilité est plutôt grande.

Une idée naturelle est de prendre $r_1, \dots, r_K$ indépendantes et de même loi $\mathcal{B}(\frac{1}{2})$; on suppose désormais que R est définie de la sorte. Ainsi, lorsque $X \notin \{X_1, \dots, X_n\}$ (ce qui se produit souvent lorsque $K \gg n$), $\widehat{s}((X_i, r_{X_i})_{1 \leqslant i \leqslant n}; X)$ est indépendante de r_X. On en déduit que

$$\mathbb{P}_{r \sim R} \big(\widehat{s}((X_i, r_{X_i})_{1 \leqslant i \leqslant n}; X) \neq r_X | X_1, \dots, X_n, X \big) \geqslant \frac{1}{2} \mathbb{1}_{X \notin \{X_1, \dots, X_n\}}$$

et donc

$$\begin{aligned}
&\mathbb{P}_{r\sim R, D_n\sim P_r^{\otimes n}, (X,Y)\sim P_r}\big(\widehat{s}(D_n;X)\neq Y\big)\\
&\geqslant \frac{1}{2}\mathbb{P}\big(X\notin\{X_1,\dots,X_n\}\big)\\
&= \frac{1}{2}\mathbb{E}\left[\mathbb{P}\left(X_1\neq X,\dots,X_n\neq X\right)|X\right]\\
&= \frac{1}{2}\mathbb{E}\left[\mathbb{P}\left(X_1\neq X|X\right)\times\cdots\times\mathbb{P}\left(X_n\neq X\right)|X\right]\\
&= \frac{1}{2}\left(1-\frac{1}{K}\right)^n\quad .
\end{aligned}$$

Récapitulons : on a démontré que pour tout $K\geqslant 1$,

$$\sup_{P \text{ loi sur } \mathcal{X}\times\mathcal{Y}}\left\{\mathbb{E}_{D_n\sim P^{\otimes n}}\left[\ell\left(s^\star,\widehat{s}(D_n)\right)\right]\right\}\geqslant\frac{1}{2}\left(1-\frac{1}{K}\right)^n\quad .$$

En faisant tendre K vers $+\infty$, on obtient la minoration cherchée. □

Le théorème se prouve grâce à la construction d'une loi de probabilité pathologique qui rend la classification difficile. Et pour cette loi, on ne peut pas faire mieux que prédire l'étiquette Y au hasard.

Ceci suggère qu'à la place de chercher à obtenir la consistance universelle uniforme, c'est-à-dire un résultat en pire cas sur l'ensemble des lois P possibles, il faut plutôt étudier des ensembles de lois P plus petits (mais « réalistes »), pour lesquels il est possible d'obtenir des *vitesses d'apprentissage* uniformes.

Exercice 2. *On considère le contraste* $0-1$ $\gamma(t;(x,y))=\mathbb{1}_{t(x)\neq y}$ *en classification binaire supervisée, et l'on suppose que* $\mathcal{X}$ *est fini. On définit la règle de majorité suivante : pour tout* $n\in\mathbb{N}$, $n\geqslant 1$, *pour tout* $x\in\mathcal{X}$,

$$\widehat{s}^{\,\mathrm{maj}}(x;D_n):=\begin{cases}1 & \text{si } \operatorname{Card}\{i \text{ t.q. } X_i=x \text{ et } Y_i=1\}>\operatorname{Card}\{i \text{ t.q. } X_i=x \text{ et } Y_i=0\}\\ 0 & \text{sinon.}\end{cases}$$

Alors, montrer que $\widehat{s}^{\,\mathrm{maj}}$ *est* uniformément universellement consistante.

Indication : utiliser l'inégalité de Hoeffding :

Théorème 5. *Soient* $\xi_1,\dots,\xi_n$ *des variables aléatoires indépendantes telles que pour tout* i, $a_i\leqslant\xi_i\leqslant b_i$ *p.s. pour des réels* $a_1,\dots,a_n$ *et* $b_1,\dots,b_n$. *Alors,*

$$\mathbb{P}\left(\sum_{i=1}^n\xi_i-\sum_{i=1}^n\mathbb{E}\left[\xi_i\right]\geqslant\varepsilon\right)\leqslant\exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n\left(b_i-a_i\right)^2}\right)\quad . \tag{1}$$

2. Minimisation du risque structurel

Références pour cette section : [3], [1, Chapitre 4], et éventuellement le premier cours de [2].

2.1. Minimisation du risque empirique.

On rappelle (voir la section 1.5) que le risque empirique est défini par :

$$P_n\gamma(t) = \frac{1}{n}\sum_{i=1}^{n}\gamma(t,(x_i,y_i)$$

On a alors $\mathbb{E}(P_n\gamma(t)) = P\gamma(t)$, c'est-à-dire que le risque empirique estime la perte sans biais pour tout $t \in \mathbb{S}$ fixé.

On peut alors estimer l'estimateur de Bayes en minimisant le risque empirique sur un modèle $S \subset \mathbb{S}$, cet estimateur est appelé l'ERM (Empirical Risk Minimizer) :

$$\widehat{s}_S(D_n) \in \operatorname{argmin}_{t\in S}\{P_n\gamma(t)\}$$

2.2. **Analyse du risque empirique sur un modèle.** On remarque la minoration de la perte relative suivante :

$$\ell\,(\,s^\star, \widehat{s}_S\,) \geqslant \inf_{t\in S} \ell\,(\,s^\star, t\,)\,.$$

On appelle cette borne inférieure l'erreur d'approximation de S et on la note :

$$\ell\,(\,s^\star, S\,) = \inf_{t\in S} \ell\,(\,s^\star, t\,)\,.$$

On suppose dans la suite que l'infimum est atteint, on note $s_S^\star$ un prédicteur réalisant cet infimum $s_S^\star \in \operatorname{argmin}_{t\in S}\{\ell\,(\,s^\star, t)\}$.

On peut alors décomposer la perte relative sous la forme suivante :

$$\ell\,(\,s^\star, s_S^\star\,) = \ell\,(\,s^\star, S\,) + (P\gamma(\widehat{s}_S) - P\gamma(s_S^\star)).$$

On appelle $P\gamma(\widehat{s}_S) - P\gamma(s_S^\star)$ (ou son espérance) l'erreur d'estimation. On retrouve alors un problème ressemblant au *compromis biais-variance*. Si le modèle S est petit, alors l'erreur d'estimation est faible mais l'erreur d'approximation est grande. C'est le problème de *sous-apprentissage*. Inversement, si le modèle S est grand, l'erreur d'approximation est petite mais l'erreur d'estimation est plus grande. Ce dernier problème est le problème de *sur-apprentissage* (overfitting) : l'estimateur suit les données plus le bruit, il apprend par cœur sans généraliser.

La propriété suivante donne une majoration de l'erreur d'estimation.

Propriété Si $\widehat{s}_S$ est un ρ minimiseur du risque empirique, c'est-à-dire si $(P_n\gamma(\widehat{s}_S) - P_n\gamma(s_S^\star) \leqslant \rho$, alors

$$P\gamma(\widehat{s}_S) - P\gamma(s_S^\star) \leqslant \rho + 2\sup_{t\in S}|(P-P_n)\gamma(t)|$$

Démonstration.

$$\begin{aligned}
& P\gamma(\widehat{s}_S) - P\gamma(s_S^\star) \\
= & \underbrace{(P-P_n)\gamma(\widehat{s}_S)}_{\leqslant \sup_{t\in S}|(P-P_n)\gamma(t)|} + \underbrace{P_n\gamma(\widehat{s}_S) - P_n\gamma(s_S^\star)}_{\leqslant \rho} + \underbrace{(P-P_n)\gamma(s_S^\star)}_{\leqslant \sup_{t\in S}|(P-P_n)\gamma(t)|} \\
\leqslant & \rho + 2\sup_{t\in S}|(P-P_n)\gamma(t)|
\end{aligned}$$

□

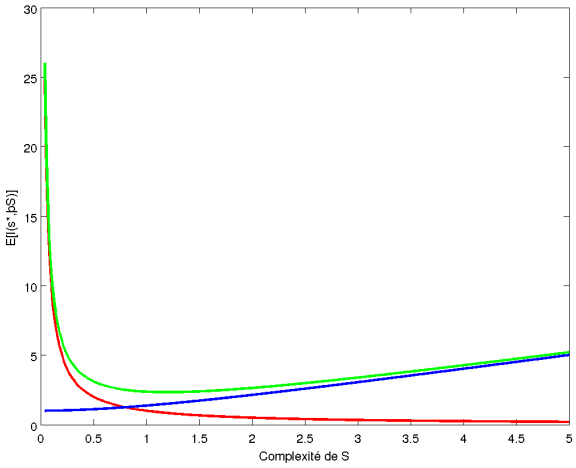

FIGURE 2. En fonction de la complexité, en rouge erreur d'approximation, en bleu erreur d'estimation, en vert somme des deux.

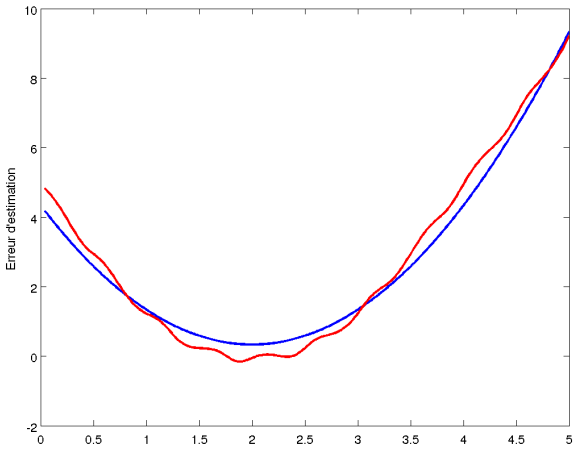

FIGURE 3. Risque en fonction de la complexité. En rouge : risque empirique, en bleu risque

D'après l'inégalité précédente et en remarquant que $\sup_{t\in S} |(P-P_n)\gamma(t)|$ est croissant en S, il est légitime de considérer ce dernier suprémum comme une mesure de complexité de S.

2.3. **Choix de modèles.** On considère une famille de modèles $(S_m)_{m\in\mathcal{M}}$. On associe à chaque modèle un minimiseur du risque empirique (ERM) : $\widehat{s}_m$ pour $m \in \mathcal{M}$. On appelle oracle, le modèle $m^\star$ qui minimise la perte relative de l'ERM associé

$$m^\star \in \arg\min_{m\in\mathcal{M}} \{\ell(s^\star, \widehat{s}_m)\}.$$

Or ce modèle est inaccessible car on ne connait pas l'estimateur de Bayes $s^\star$. À la place, on cherche un modèle $\widehat{m}(D_n) \in \mathcal{M}$ qui vérifie l'inégalité suivante, dite inégalité oracle, avec une grande probabilité ou en espérance :

$$\ell\,(s^\star, \widehat{s}_{\widehat{m}}\,) \leqslant C \inf_{m\in\mathcal{M}} \{\,\ell\,(s^\star, \widehat{s}_m\,)\,\} + R_n.$$

Quelques exemples de familles de modèles en régression :
– On note m une partition finie de $\mathcal{X}$,

$$S_m = \{f : \mathcal{X} \to \mathbb{R} \text{ mesurable telle que } \forall \lambda \in m,\ f \text{ est constante sur } \lambda$$
$$= \mathrm{ev}\{\mathbb{1}_\lambda, \lambda \in m\}$$

– $S_m = \mathrm{ev}\{\Phi_1, \dots, \Phi_m\}$, où $(\Phi_k)_{k\geqslant 1}$ est une famille libre de $L^2(P_X)$
– $S_m = \{f : (x_1, \dots x_d) \to \sum_{j\in m} \alpha_j x_j, \alpha_j \in \mathbb{R}\}$, où $\mathcal{X} \subset \mathbb{R}^d$, $m \subset \{1, \dots, d\}$.

En classification, un exemple de famille de modèles est $S = \{f : \mathcal{X} \to \{0,1\}$ telle que $E_f = \{x : f(x) = 1\}$ est un demi espace de $\mathbb{R}^d\}$

Le principe de la sélection de modèle à l'aide des données est de trouver $\widehat{m}$ minimisant un critère $\mathrm{crit}(m)$ approximant la perte relative de l'ERM du modèle m pour tout $m \in \mathcal{M}_n$: c'est-à-dire, si $\mathrm{crit}(m) \approx \ell\,(s^\star, \widehat{s}_m(D_n)\,)$ pour tout $m \in \mathcal{M}_n$, on choisit

$$\widehat{m} \in \mathrm{argmin}_{m\in\mathcal{M}_n} \{\,\mathrm{crit}(m)\,\}.$$

On appelle le critère idéal, le critère suivant : $\mathrm{crit}_{\mathrm{id}}(m) = P\gamma(\widehat{s}_m)$ ou $\ell\,(s^\star, \widehat{s}_m\,)$.

Le lemme suivant donne une recette pour obtenir une égalité oracle.

Lemme 6. *Si* $\widehat{m} \in \arg\min_{m\in\mathcal{M}} \{\,\mathrm{crit}(m)\,\}$ *avec*

$$\forall m \in \mathcal{M}, \quad -B(m) \leqslant \mathrm{crit}(m) - \mathrm{crit}_{\mathrm{id}}(m) \leqslant A(m)$$

alors

$$\ell\,(s^\star, \widehat{s}_{\widehat{m}}\,) - B(\widehat{m}) \leqslant \inf_{m\in\mathcal{M}_n} \{\,\ell\,(s^\star, \widehat{s}_m\,) + A(m)\,\}$$

Démonstration. Par définition de $\widehat{m}$,

$$\mathrm{crit}(\widehat{m}) \leqslant \mathrm{crit}(m).$$

Ainsi,

$$\mathrm{crit}_{\mathrm{id}}(\widehat{s}) + (\mathrm{crit} - \mathrm{crit}_{\mathrm{id}})(\widehat{s}) - P\gamma(s^\star) \leqslant \mathrm{crit}_{\mathrm{id}}(m) + (\mathrm{crit} - \mathrm{crit}_{\mathrm{id}})(m) - P\gamma(s^\star)$$

D'où,

$$\ell\,(s^\star, \widehat{s}_{\widehat{m}}\,) - B(\widehat{m}) \leqslant \inf_{m\in\mathcal{M}_n} \{\,\ell\,(s^\star, \widehat{s}_m\,) + A(m)\,\}$$

□

Exemple : si $\mathrm{crit}(m) \geqslant \mathrm{crit}_{\mathrm{id}}(m)$ pour tout m, le lemme s'applique avec $B(m) = 0$ et $A(m) = \mathrm{crit}(m) - \mathrm{crit}_{\mathrm{id}}(m)$.

Références

[1] Sylvain Arlot. Classification supervisée : des algorithmes et leur calibration automatique, 2009. Notes d'un cours de troisième année à l'École Centrale Paris. `http://www.di.ens.fr/~arlot/enseign/2009Centrale/cours-classif.pdf.gz`.

[2] Sylvain Arlot. Sélection de modèles et sélection d'estimateurs pour l'apprentissage statistique, January 2011. Cours Peccot. Collège de France. `http://www.di.ens.fr/~arlot/peccot.htm`.

[3] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1996.