Classification supervisée : des algorithmes et leur calibration automatique

Sylvain Arlot
Cnrs – École Normale Supérieure – Inria, Équipe-projet Willow
sylvain.arlotRETIRERCECI@ens.fr

http://www.di.ens.fr/∼arlot/

École Centrale de Paris, Cours de troisième année

6 Mars 2009

Table des matières

CHAPITRE 1

Classification supervisée : un panorama

Ces notes de cours sont largement inspirées de [**DGL96**]. À propos de la théorie statistique de l'apprentissage (qui comprend la classification supervisée), voir aussi [**Vap82, Vap98**]. Une autre référence intéressante est [**HTF01**].

1.1. Cadre, exemples

1.1.1. Description formelle. On observe $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \mathcal{Y}$, que l'on suppose être n réalisations indépendantes d'une variable aléatoire (X, Y) de loi inconnue P. L'objectif est de prédire Y_{n+1} à partir de X_{n+1} et $(X_i, Y_i)_{1 \leq i \leq n}$ en se trompant aussi rarement que possible, où (X_{n+1}, Y_{n+1}) est une nouvelle réalisation de (X, Y), indépendante de $(X_i, Y_i)_{1 \leq i \leq n}$.

La variable X est un ensemble de caractéristiques facilement observables d'un objet d'étude ; typiquement, $\mathcal{X}$ est une partie de $\mathbb{R}^\ell$ avec ℓ grand, et chacune des ℓ coordonnées de X est une variable (discrète ou continue) décrivant l'objet d'étude. La variable Y est une quantité discrète (disons, $\mathcal{Y} = \{0, \ldots, M-1\}$) représentant une caractéristique intéressante de l'objet d'étude, mais difficile ou impossible à observer ; Y est appelée *étiquette* associée à X.

On appelle indifféremment $(X_i, Y_i)_{1 \leq i \leq n}$ *l'échantillon*, *les observations* ou *les données*, et on le(s) note souvent D_n.

Remarque 1.1 (Classification et régression). On parle de *classification* car la quantité d'intérêt Y est discrète ; il s'agit au final de *classer* un objet (représenté par un point X de $\mathcal{X}$) dans l'une des M classes en lui attribuant son étiquette Y. Lorsque $\mathcal{Y} = [0, 1]$ ou $\mathbb{R}$ par exemple (quantité d'intérêt continue), on parle de *régression*. Nous ne traiterons pas de ce type de problème d'apprentissage dans ce cours.

Remarque 1.2 (Classification supervisée). La classification est dite *supervisée* car toutes les données sont étiquetées, c'est-à-dire que Y_i est observé pour $i = 1, \ldots, n$. Lorsqu'aucun des Y_i n'est observé, on parle de classification *non-supervisée*, et l'objectif est de regrouper les objets en M classes «cohérentes». Lorsque seule une (petite) partie des étiquettes est observée, on parle de classification *semi-supervisée* ; ce cadre se rencontre souvent lorsqu'il est très peu coûteux d'accéder à beaucoup de réalisations de X (par exemple, des photos téléchargées sur le Web) mais beaucoup plus coûteux de les étiqueter. Nous ne traiterons pas non plus dans ce cours ces deux types de problèmes de classification.

1.1.2. Exemples. Il existe un multitude de problèmes qui entrent dans le cadre de la classification supervisée, parmi lesquels :

(1) **Reconnaissance et identification de caractères manuscrits** : X est une image en niveaux de gris, *i.e.* $\mathcal{X} = [0, 1]^K$ où K est le nombre de pixels. L'étiquette Y indique le caractère représenté par X ($M = 11$ pour reconnaître les 10 chiffres, $M = 37$ si l'on ajoute les lettres, $M = 63$ en comptant les majuscules, et plus encore si l'on tient compte des accents et autres caractères spéciaux).

(2) Plus généralement, la **reconnaissance de formes** : X est une image (éventuellement en couleurs, donc $\mathcal{X} = [0,1]^{3K}$), et Y indique si l'image possède ou non une caractéristique donnée (contenir une voiture, un humain, *etc.* ; représenter un humain qui sourit, qui pleure, *etc.*) ; ici, $M = 2$ si la question posée est fermée (oui ou non), mais on peut aussi considérer des questions plus ouvertes.

(3) **Reconnaissance de parole :** X est un enregistrement sonore numérisé, $\mathcal{X} = \mathbb{R}^\ell$ avec ℓ très grand et Y indique qui parle (parmi un petit nombre de personnes possibles), ou bien ce qui est dit (parmi une petite liste de phrases possibles).

(4) **Catégorisation de textes** : X est un texte ($\mathcal{X}$ est l'ensemble des suites finies de caractères), Y indique qui a écrit le texte (par exemple, est-ce Shakespeare ? est-ce Corneille ou Racine ?) ou bien quelle est le thématique principale du texte, *etc.*.

(5) **Détection de spams** : X est le contenu d'un e-mail ($\mathcal{X} \subset \{0,1\}^K$ avec K grand est l'ensemble des fichiers binaires de taille < 5Mo), $Y = 1$ si c'est un spam et $Y = 0$ sinon.

(6) **Aide au diagnostic médical** : X est un ensemble de caractéristiques du patient (fréquence cardiaque, température corporelle, résultats d'examens médicaux, âge, sexe, antécédents personnels ou familiaux, *etc.*), Y donne une information sur l'état de santé réel du patient (est-il en danger de mort ? faut-il l'opérer ? cela vaut-il la peine de lui faire passer un examen dangereux ou très coûteux ?).
Dans le cadre de la cancérologie, on peut par exemple inclure dans X des données d'expressions de gènes (donc $\mathcal{X} = \mathbb{R}^\ell$ avec ℓ très grand), la quantité d'intérêt Y indiquant si le patient est ou non atteint d'un cancer, ou bien de quel type de cancer il s'agit (va-t-il y avoir des métastases ?).

(7) **Bioinformatique** : détection de «gènes» dans une séquence ADN ($\mathcal{X} = \{A, T, C, G\}^{\mathbb{N}}$), détection de sites actifs dans une protéine ($\mathcal{X} = \{\text{acides aminés}\}^{\mathbb{N}}$), catégorisation de protéines, *etc.*

Citons également l'attribution d'un prêt à un client d'une banque (étant données les caractéristiques d'un client et la conjoncture économique, va-t-il pouvoir rembourser son prêt ?), la vidéo-surveillance (détecter si un événement exceptionnel tel qu'un accident se déroule dans une vidéo), le contrôle parental (une image, une vidéo ou le contenu d'une page web peuvent-ils être visionnés par un mineur ?), l'interface cerveau-machine (étant donnés un ensemble de signaux électriques observés par des électrodes à la surface du crâne, déterminer à quoi pense le sujet ; voir par exemple Brain-Pong[1]), la classification de produits boursiers (en fonction de leur notation par des experts et de caractéristiques, sont-ils ou non sous-évalués ?), *etc.*

REMARQUE 1.3 (Définition du problème). Dans le processus de modélisation du problème, il est important de bien définir X et Y. Si possible, ne pas perdre des informations que l'on a déjà : si X est une image, poser $\mathcal{X} = [0,1]^K$ risque de faire oublier la structure spatiale des données ; si X est un e-mail, poser $\mathcal{X} = \{0,1\}^K$ risque de faire oublier son formatage particulier : en-tête, corps du message, fichier attaché, *etc.*

Une manière (parmi d'autres) d'incorporer des informations sur la nature des données est de procéder à un pré-traitement des données ; dans le cas d'un mail, on peut par exemple représenter X à l'aide de caractéristiques spécifiques uniquement (présence de certains mots, présence d'un fichier attaché exécutable, nom de domaine de l'expéditeur, liste des destinataires, présence d'erreurs de formatage de l'en-tête, *etc.*).

[1] `http://ida.first.fraunhofer.de/bbci/index_en.html`

REMARQUE 1.4 (Acquisition des données). Il faut également faire attention à la manière d'acquérir les données : minimiser le nombre d'erreurs d'étiquetage, veiller au caractère i.i.d. des données (si les données sont hétérogènes, incorporer cette hétérogénéité dans X_i ; par exemple, qui a étiqueté la i-ème donnée ?), réfléchir à la structure de la base de données pour faciliter l'accès (en fonction des méthodes de classification que l'on envisage d'utiliser).

REMARQUE 1.5 (Modélisation probabiliste). Pourquoi une modélisation probabiliste pour les (X_i, Y_i) ? Le caractère aléatoire de X est naturel (on a choisi un objet/patient «au hasard» parmi l'ensemble des possibles ; noter que choisir une personne aléatoirement uniformément parmi les patients d'un hôpital est différent de choisir une personne aléatoirement uniformément dans l'ensemble de la population française). Dans le cas de Y, c'est une variable aléatoire car Y dépend de X qui est aléatoire. Mais même conditionnellement à X, Y reste en général (on l'espère, faiblement) aléatoire . Cet aléa peut provient de plusieurs sources.

D'une part, les erreurs d'acquisition de X et Y sont inévitables, mais peuvent être réduites au minimum.

D'autre part, X ne contient pas toujours toute l'information nécessaire pour déterminer Y de manière unique. Par exemple, dans le cas de l'aide au diagnostic médical, X ne contient pas toute l'information sur la situation présente du patient (on ne peut pas faire tous les examens possibles sur tous les patients), et encore moins son *futur*. De même, pour la reconnaissance de caractères, certaines images sont trop ambiguës pour espérer obtenir une classification parfaite. Un 6 ou un 0 mal écrits sont souvent indistinguables, et l'image X ne contient que ce qui est écrit, pas ce que l'on a voulu écrire. Dans les deux cas, on peut formaliser le problème ainsi : Y est une fonction déterministe f de (X, Z), où X est observée mais pas Z. Du coup, si l'on a choisi (X, Z) aléatoirement parmi une population donnée, conditionnellement à X, $Y = f(X, Z)$ est encore une variable aléatoire sauf si Z est elle-même une fonction de X :

$$\mathbb{P}(Y = 1 \mid X = x) = \mathbb{P}(f(x, Z) = 1)$$

n'appartient pas nécessairement à $\{0, 1\}$. L'objectif de la classification est alors de faire au mieux pour prédire Y étant donné X ; notons que le problème de la reconnaissance de caractéres écrits par un français est donc légèrement différente de la reconnaissance de caractères écrits par un anglais.

Enfin, si l'on suppose que les (X_i, Y_i) sont i.i.d., c'est avant tout car ce cadre simplifie grandement l'analyse théorique tout en étant réaliste pour bon nombre de problèmes. Il existe d'autres cadres de classification où l'on ne suppose plus les données i.i.d. Par exemple, l'apprentissage «actif» s'intéresse à des cas où les données sont acquises séquentiellement, les X_i étant choisis en fonction des observations passées ; ainsi, on peut concentrer l'exploration de l'espace $\mathcal{X}$ sur les zones où le problème de classification est le plus difficile.

Par souci de simplicité, dans la suite de ce cours, on se concentrera sur le cas de la *classification binaire*, où il n'y a que deux étiquettes possibles ($\mathcal{Y} = \{0, 1\}$). La plupart des algorithmes s'étendent naturellement du cas binaire au cas multi-classes (ne serait-ce que parce qu'on peut décomposer un problème multi-classes en plusieurs problème de classification binaire), avec des difficultés techniques supplémentaires du côté statistique comme du côté algorithmique.

1.2. Classifieur, Risque, Classifieur de Bayes

Résoudre un problème de classification revient à trouver un classifieur, défini comme suit.

DÉFINITION 1.1 (Classifieur). On appelle *classifieur* toute application mesurable $t : \mathcal{X} \mapsto \mathcal{Y}$. L'ensemble des classifieurs est noté $\mathbb{S}$.

Étant donnée une nouvelle observation X_{n+1}, le classifieur t lui attribue l'étiquette $t(X_{n+1})$, que l'on espère coïncider avec Y_{n+1}.

Afin de différencier les bons classifieurs des mauvais pour un problème donné, il est nécessaire de définir une mesure de qualité du classifieur t. Dans la suite, on utilisera la quantité suivante.

DÉFINITION 1.2 (Risque d'un classifieur). Soit $t : \mathcal{X} \mapsto \mathcal{Y}$ un classifieur et P la distribution de la variable (X, Y) générant les données. Alors, le *risque* de t est défini comme la probabilité de mauvaise classification sous la loi P :

$$\mathcal{R}_P(t) = \mathcal{R}(t) := \mathbb{P}_{(X,Y)\sim P}(t(X) \neq Y) = \mathbb{P}(t(X) \neq Y) \quad . \tag{1.1}$$

Un classifieur t est alors d'autant meilleur que son risque $\mathcal{R}(t)$ est petit.

REMARQUE 1.6 (Classes asymétriques). La définition (1.1) ci-dessus n'est pas appropriée pour certains problèmes de classification. En effet, dans le cas de la détection de spams (comme dans l'aide au diagnostic médical), les deux classes ne jouent pas de rôles symétriques : il est plus grave de laisser passer un spam à tort que de classer «spam» un mail qui ne l'est pas. Dans ces situations, il est nécessaire d'incorporer cette asymétrie dans le risque en donnant un poids différent aux différents types d'erreurs :

$$\mathcal{R}(t) := w_1 \mathbb{P}(t(X) \neq Y \text{ et } Y = 1) + w_0 \mathbb{P}(t(X) \neq Y \text{ et } Y = 0) = \mathbb{E}[\gamma_w(t; (X, Y))]$$

avec $\gamma_w(t; (x, y)) := w_y \mathbb{1}_{t(x)\neq y}$. On retrouve la définition (1.1) du risque en prenant $w_0 = w_1 = 1$.

REMARQUE 1.7 (Cadre transductif). La définition (1.1) du risque mesure la probabilité de mauvaise classification d'une nouvelle observation X *tirée suivant la même loi* P *que les données.* Cette hypothèse n'est pas toujours réaliste. Par exemple, lorsque l'une des deux classes apparaît rarement dans l'ensemble de la population (patients atteints de formes rares de cancer, événements exceptionnels dans des vidéos tels qu'un accident, *etc.*), on risque fort de ne jamais l'observer si n n'est pas très grand. Pour pallier ce problème, une solution naturelle est d'acquérir un échantillon où l'on impose la présence des deux classes, par exemple dans des proportions fixées (approximativement) à l'avance. Autrement dit, l'échantillon est tiré suivant une loi P_{ech} différente de la loi P_{pop} de la variable (X, Y) dans la population générale. Lorsque l'échantillon est i.i.d., ceci correspond au cadre d'apprentissage *transductif*; l'objectif est de bien classifier une réalisation (X_{n+1}, Y_{n+1}) tirée suivant P_{pop}, la définition du risque $\mathcal{R}(t)$ est modifiée en conséquence.

Dans la suite, on supposera toujours le risque défini par (1.1). La distribution P des données étant inconnue, la valeur de $\mathcal{R}_P(t)$ n'est pas accessible en pratique, il est donc impossible de la minimiser directement. Il est toutefois intéressant d'identifier plus précisément, s'il(s) existe(nt), le(s) classifieur(s) minimisant le risque $\mathcal{R}_P(t)$ parmi l'ensemble $\mathbb{S}$ des classifieurs. La proposition suivante répond à cette question dans notre cadre.

PROPOSITION 1.1 (Classifieur de Bayes). *Soit* $\eta : \mathcal{X} \mapsto [0, 1]$ *la fonction définie par* $\eta(x) = \mathbb{P}(Y = 1 \mid X = x)$ *pour tout* $x \in \mathcal{X}$. *Alors, le classifieur* $s^\star$ *défini par*

$$s^\star(x) = \mathbb{1}_{\eta(x)\geq \frac{1}{2}} \tag{1.2}$$

vérifie

$$\forall t \in \mathbb{S}, \qquad \mathcal{R}_P(t) \geq \mathcal{R}_P(s^\star) = \mathbb{E}[\min\{\eta(X), 1 - \eta(X)\}] = \frac{1}{2} - \frac{1}{2}\mathbb{E}[|2\eta(X) - 1|] \quad . \tag{1.3}$$

Le classifieur $s^\star$ *est appelé* classifieur de Bayes, *et son risque* $\mathcal{R}_P(s^\star) = \mathcal{R}_P^\star = \mathcal{R}^\star$ *est appelé* risque de Bayes.

De plus, pour tout classifieur $t \in \mathbb{S}$, on a

$$\mathcal{R}_P(t) - \mathcal{R}_P(s^\star) = \ell(s^\star, t) = \mathbb{E}\left[|2\eta(X) - 1| \mathbb{1}_{t(X) \neq s^\star(X)}\right] \geq 0 \ . \tag{1.4}$$

La quantité $\ell(s^\star, t)$ est appelée excès de risque *du classifieur t. L'expression* (1.4) *implique que tout classifieur t minimisant le risque $\mathcal{R}_P(t)$ coïncide avec le classifieur de Bayes $s^\star$ presque partout (relativement à la distribution P_X de X), sauf peut-être sur l'ensemble*

$$\left\{ x \in \mathcal{X} \ \textit{t.q.}\ \eta(x) = \frac{1}{2} \right\} \ .$$

Il est à noter que ce résultat est très intuitif. En effet, étant donné un point x, $s^\star$ choisit l'étiquette Y la plus probable selon la loi de Y sachant $X = x$. Un classifieur t différant de $s^\star$ sur un ensemble de mesure positive (pour P_X) où η ne prend jamais la valeur $1/2$ a donc nécessairement un risque $\mathcal{R}(t)$ strictement plus grand que $\mathcal{R}(s^\star)$.

Démonstration. Soit $t \in \mathbb{S}$ un classifieur et $x \in \mathcal{X}$ quelconque. Alors,

$$\begin{aligned}
\mathbb{P}(t(X) \neq Y \mid X = x) &= \mathbb{P}(t(X) \neq 0 \text{ et } Y = 0 \mid X = x) + \mathbb{P}(t(X) \neq 1 \text{ et } Y = 1 \mid X = x) \\
&= \mathbb{1}_{t(x)=1}\mathbb{P}(Y = 0 \mid X = x) + \mathbb{1}_{t(x)=0}\mathbb{P}(Y = 1 \mid X = x) \\
&= t(x)(1 - \eta(x)) + (1 - t(x))\eta(x) = t(x)(1 - 2\eta(x)) + \eta(x) \qquad (1.5) \\
&\geq \min\{1 - \eta(x), \eta(x)\} = \frac{1}{2} - \frac{|2\eta(x) - 1|}{2} \ . \qquad (1.6)
\end{aligned}$$

Or, lorsque $t = s^\star$, (1.6) est une égalité car soit $\eta(x) = 1/2$ et il y a toujours égalité, soit $\eta(x) \neq 1/2$ et alors $s^\star(x) = 1$ équivaut à $1 - \eta(x) = \min\{\eta(x), 1 - \eta(x)\}$. On en déduit (1.3) en intégrant par rapport à X.

D'après (1.5),

$$\mathbb{P}(t(X) \neq Y \mid X = x) - \mathbb{P}(s^\star(X) \neq Y \mid X = x) = (t(x) - s^\star(x))(1 - \eta(x)).$$

On en déduit (1.4) en intégrant par rapport à X. □

Remarque 1.8 (Classifieur randomisé). La Proposition 1.1 est pratiquement inchangée si l'on étend l'ensemble des classifieurs aux classifieurs randomisés. En effet, un classifieur (binaire) randomisé t étant défini comme une application mesurable $\mathcal{X} \mapsto [0, 1]$, son risque conditionnellement à $X = x$ vaut

$$\begin{aligned}
\mathbb{P}(t(X) \neq Y \mid X = x) &= \mathbb{P}(t(X) \neq 0 \text{ et } Y = 0 \mid X = x) + \mathbb{P}(t(X) \neq 1 \text{ et } Y = 1 \mid X = x) \\
&= t(x)(1 - \eta(x)) + (1 - t(x))\eta(x) = t(x)(1 - 2\eta(x)) + \eta(x)
\end{aligned}$$

qui est minimal lorsque $t(x) = s^\star(x)$, ce minimum étant strict sauf lorsque $\eta(x) = 1/2$. Par conséquent, (1.3) reste vrai et (1.4) doit être modifiée en

$$\mathcal{R}(t) - \mathcal{R}(s^\star) = \mathbb{E}\left[(1 - 2\eta(X))(t(X) - s^\star(X))\right] \geq 0 \ .$$

1.3. Notions de consistance

1.3.1. Classifieur, règle de classification. Si la Proposition 1.1 ci-dessus donne une formule explicite pour le meilleur classifieur, elle n'est d'aucune utilité en pratique car la fonction η est inconnue. L'objectif de la classification supervisée est d'inférer un classifieur t ayant un petit risque $\mathcal{R}(t)$ en n'utilisant que les données.

Autrement dit, on cherche une application mesurable $\widehat{s} : (\mathcal{X} \times \mathcal{Y})^n \mapsto \mathbb{S}$ qui à un échantillon $D_n = (X_i, Y_i)_{1 \leq i \leq n}$ associe un classifieur $\widehat{s}(D_n)$. La valeur en un point x du classifieur $\widehat{s}(D_n)$ est notée $\widehat{s}(x; D_n)$, ou $\widehat{s}(x)$ lorsque cela n'induit pas d'ambiguïté.

L'application $\widehat{s}$ est par extension également appelée *classifieur*. On utilisera systématiquement des notations avec un «chapeau» $\widehat{\cdot}$ pour de tels classifieurs (inférés à partir d'un échantillon), par opposition à des classifieurs notés t ou $s^\star$ qui ne dépendent pas de D_n.

La qualité d'un classifieur $\widehat{s}$ est naturellement mesurée par son risque $\mathcal{R}_P\left(\widehat{s}(D_n)\right)$. Il faut noter ici sur un point important. Dans la définition (1.1) du risque d'un classifieur t,

$$\mathcal{R}_P\left(t\right) = \mathbb{P}_{(X,Y)\sim P}\left(t(X) \neq Y\right) \ ,$$

la probabilité n'est prise que relativement à l'aléa de la variable de test (X, Y). Implicitement, cette probabilité est *conditionnelle au classifieur* t, y compris lorsque $t = \widehat{s}(D_n)$ est aléatoire car dépendant des données. Par conséquent,

$$\mathcal{R}_P\left(\widehat{s}(D_n)\right) = \mathbb{P}_{(X,Y)\sim P}\left(\widehat{s}(X; D_n) \neq Y \mid \ D_n\right)$$

dépend des observations D_n ; c'est une variable aléatoire.

Il y a deux façons principales de mesurer la qualité d'un classifieur $\widehat{s}$ indépendamment d'un échantillon particulier D_n : soit en moyenne, avec l'espérance du risque

$$\mathbb{E}\left[\mathcal{R}_P\left(\widehat{s}(D_n)\right)\right] \ ,$$

soit en probabilité, avec la queue de distribution du risque

$$\mathbb{P}\left(\mathcal{R}_P\left(\widehat{s}(D_n)\right) > \epsilon\right) \ , \qquad \text{pour tout } \epsilon \geq \mathcal{R}_P^\star \ .$$

Le risque de Bayes $\mathcal{R}_P^\star$ étant une borne inférieure déterministe pour $\mathcal{R}_P\left(\widehat{s}(D_n)\right)$, on souhaite un classifieur $\widehat{s}$ qui ait un risque aussi proche que possible de $\mathcal{R}_P^\star$ en moyenne ou avec grande probabilité. Ceci étant en général impossible à taille d'échantillon n fixée, on va chercher à obtenir cette propriété lorsque n tend vers l'infini. Avant d'aller plus loin, il nous faut définir l'équivalent d'un classifieur $\widehat{s}$ lorsque la taille d'échantillon n n'est pas fixée *a priori*.

Définition 1.3 (Règle de classification). Une règle de classification $\widehat{s}$ est une suite $\left(\widehat{s}_n\right)_{n\in\mathbb{N}}$ de classifieurs $\widehat{s}_n : \left(\mathcal{X} \times \mathcal{Y}\right)^n \mapsto \mathbb{S}$.

Par abus de notation, on écrit souvent $\widehat{s}$ à la place de $\widehat{s}_n$ pour une taille d'échantillon donnée, étant sous-entendu que $\widehat{s}(D_n) = \widehat{s}_n(D_n)$ pour un échantillon D_n de taille n.

1.3.2. Consistance. Une «bonne» règle de classification est une règle de classification dont le risque est proche du risque de Bayes lorsque la taille de l'échantillon est suffisamment grande. Ceci conduit aux deux définitions suivantes :

Définition 1.4 (Consistance faible). Une règle de classification $\widehat{s}$ est dite (*faiblement*) *consistante* pour la distribution P si

$$\mathbb{E}_{D_n\sim P^{\otimes n}}\left[\mathcal{R}_P\left(\widehat{s}_n(D_n)\right)\right] \xrightarrow[n\to\infty]{} \mathcal{R}_P^\star \ . \tag{1.7}$$

Définition 1.5 (Consistance forte). Une règle de classification $\widehat{s}$ est dite *fortement consistante* pour la distribution P si $\mathcal{R}_P\left(\widehat{s}_n(D_n)\right)$ converge presque sûrement vers $\mathcal{R}_P^\star$, c'est-à-dire

$$\mathbb{P}_{D_n\sim P^{\otimes n}}\left(\mathcal{R}_P\left(\widehat{s}_n(D_n)\right) \xrightarrow[n\to\infty]{} \mathcal{R}_P^\star\right) = 1 \ . \tag{1.8}$$

Remarque 1.9. Une règle $\widehat{s}$ est (faiblement) consistante lorsque $\mathcal{R}_P\left(\widehat{s}_n(D_n)\right)$ converge vers $\mathcal{R}_P^\star$ en espérance. Comme $\mathcal{R}_P\left(\widehat{s}_n(D_n)\right)$ est une suite de variables aléatoires uniformément bornées entre $\mathcal{R}_P^\star$ et 1, cette convergence est équivalente à sa convergence en probabilités :

$$\forall \epsilon > 0, \qquad \mathbb{P}_{D_n\sim P^{\otimes n}}\left(\mathcal{R}_P\left(\widehat{s}_n(D_n)\right) - \mathcal{R}_P^\star > \epsilon\right) \xrightarrow[n\to\infty]{} 0 \ .$$

Et comme la convergence presque sûre (1.8) implique la convergence en probabilités, ceci montre que toute règle de classification fortement consistante est nécessairement (faiblement) consistante, d'où cette terminologie.

1.3.3. Consistance universelle. La distribution P des données étant *a priori* inconnue, on attend d'une bonne règle de classification qu'elle soit consistante (resp. fortement consistante) pour toute distribution P sur $\mathcal{X} \times \mathcal{Y}$; on dit alors qu'elle est *universellement consistante* (resp. universellement fortement consistante).

Mieux encore, on aimerait disposer d'une règle de classification qui soit universellement consistante avec une vitesse de convergence uniforme sur l'ensemble des distributions P possibles ; une telle règle est dite *uniformément universellement consistante*.

Formellement, ces deux notions de consistance universelle peuvent s'écrire ainsi dans le cas de la consistance faible :

– $\widehat{s}$ est universellement consistante si et seulement si

$$\sup_P \lim_{n\to\infty} \left(\mathbb{E}_{D_n\sim P^{\otimes n}}\left[\mathcal{R}_P\left(\widehat{s}_n(D_n)\right)\right] - \mathcal{R}_P^\star\right) = 0 \ , \tag{1.9}$$

– $\widehat{s}$ est uniformément universellement consistante si et seulement si

$$\lim_{n\to\infty} \sup_P \left(\mathbb{E}_{D_n\sim P^{\otimes n}}\left[\mathcal{R}_P\left(\widehat{s}_n(D_n)\right)\right] - \mathcal{R}_P^\star\right) = 0 \ , \tag{1.10}$$

les deux $\sup_P$ étant pris sur l'ensemble des mesures de probabilité P sur $\mathcal{X} \times \mathcal{Y}$. Clairement, la consistance universelle uniforme implique la consistance universelle ; la réciproque n'a en revanche aucune raison d'être vraie (elle est d'ailleurs fausse en général).

Il n'est pas évident que la consistance universelle (uniforme ou pas, faible ou forte) soit possible. Ainsi, nous démontrerons en Section 2.2 qu'il est sans espoir d'obtenir une règle de classification universellement consistante(Théorème 2.3). De plus, «en moyenne» sur les distributions P, deux règles de classification sont toujours équivalentes. Il ne peut donc pas y avoir de «meilleure» règle de classification supervisée pour une distribution P générique.

Nous verrons en revanche plus loin dans ce cours qu'il existe des règles (non-uniformément) universellement consistantes, le premier résultat du genre avec $\mathcal{X}$ infini n'ayant été démontré qu'en 1977 **[Sto77]**.

L'objectif du domaine de la classification supervisée est de proposer différentes règles de classification et d'identifier pour chacune les familles de distributions P pour lesquelles elles sont (uniformément) optimales. On s'intéresse donc à des quantités telles que

$$\sup_{P\in\mathcal{P}} \left(\mathbb{E}_{D_n\sim P^{\otimes n}}\left[\mathcal{R}_P\left(\widehat{s}_n(D_n)\right)\right] - \mathcal{R}_P^\star\right) \ ,$$

asymptotiquement ou à n fixé, pour des familles $\mathcal{P}$ «raisonnables». Ensuite, en fonction du problème considéré, on utilise la règle qui fonctionne le mieux pour la famille $\mathcal{P}$ qui lui correspond le mieux.

Des exemples naturels de familles $\mathcal{P}$ sont l'ensemble des distributions «zéro-erreur» (c'est-à-dire telles que $\eta(x) \in \{0, 1\}$, soit $Y = s^\star(X)$), ou bien l'ensemble des distributions telles que η est «régulière».

1.4. La règle des k plus proches voisins

L'idée principale de cette règle de classification est que l'étiquette d'un point $x \in \mathcal{X}$ est correctement prédite par celle de ses voisins dans $\mathcal{X}$. On obtient une règle de classification en posant que l'étiquette d'un point $x \in \mathcal{X}$ est celle qui est majoritaire parmi les étiquettes de ses

k plus proches voisins dans l'échantillon. Le nombre k de voisins considérés et la distance sur $\mathcal{X}$ utilisée pour définir la notion de «voisin dans $\mathcal{X}$» sont à fixer librement par l'utilisateur. Plus formellement, on a la définition suivante.

DÉFINITION 1.6 (Règle des k plus proches voisins). Soit d une distance sur $\mathcal{X}$ et $k \geq 1$ un entier. *La règle de classification des k plus proches voisins* (k-ppv, k-NN en anglais) $\widehat{s}^{k-\mathrm{ppv}}$ est définie comme suit. Pour tout $(x_i, y_i)_{1\leq i\leq n} \in (\mathcal{X}\times\mathcal{Y})^n$ et $x \in \mathcal{X}$, on ordonne les x_i par distance croissante à x : il existe une unique permutation τ_x de $\{1,\dots,n\}$ telle que $d(x, x_{\tau_x(1)}) \leq \cdots \leq d(x, x_{\tau_x(n)})$ avec

$$\forall i \in \{1,\dots,n-1\}, \quad d(x, x_{\tau_x(i)}) < d(x, x_{\tau_x(i+1)}) \text{ ou } \tau_x(i) < \tau_x(i+1) \ .$$

On définit alors le classifieur des k plus proches voisins par

$$\widehat{s}^{k-\mathrm{ppv}}\left(x; (x_i, y_i)_{1\leq i\leq n}\right) := \mathbb{1}_{\widehat{\eta}_n^{k-\mathrm{ppv}}(x)\geq \frac{1}{2}} \qquad \text{où} \quad \widehat{\eta}_n^{k-\mathrm{ppv}}(x) := \frac{1}{k}\sum_{j=1}^{k} Y_{\tau_x(j)} \ . \tag{1.11}$$

REMARQUE 1.10 (Temps de calcul). D'un point de vue pratique, la règle des k-ppv est coûteuse en termes de stockage ($\mathcal{O}(n)$) car elle nécessite de conserver en mémoire tout l'échantillon (sauf cas très particulier).

Une implémentation naïve consiste à calculer la distance de x à tous les points de l'échantillon, d'où un algorithme en $\mathcal{O}(n)$. Des méthodes plus fines sont fondées sur la construction d'un arbre permettant d'évaluer $\widehat{s}^{k-\mathrm{ppv}}(x)$ avec un coût moyen de $\mathcal{O}(\ln(n))$; en revanche, le coup de construction de l'arbre est en $\mathcal{O}(n\ln(n))$. Voir par exemple 'kd-tree' sur la version anglaise de Wikipedia.

REMARQUE 1.11 (Choix de d). D'un point de vue pratique, le choix d'une distance sur $\mathcal{X}$ est crucial. En particulier, lorsque $X \in \mathbb{R}^\ell$ réunit ℓ variables qui ne sont pas nécessairement d'amplitudes comparables (par exemple, la fréquence cardiaque et la température corporelle), utiliser la distance euclidienne sur $\mathbb{R}^\ell$ est un choix très discutable. Dans ce cas précis, il faut au moins pondérer les différentes variables avant d'utiliser la distance euclidienne (ou la distance induite par $\|\cdot\|_p$ pour un certain $p \in [1,\infty]$).

Nous reviendrons en détail sur la règle des k-ppv au chapitre 3. Mentionnons simplement quelques propriétés importantes de cette règle :

- Pour tout k et P fixés, $\lim_{n\to\infty} \mathbb{E}\left[\mathcal{R}_P\left(\widehat{s}^{k-\mathrm{ppv}}\right)\right] = \mathcal{R}_P^{k-\mathrm{ppv}}$, avec

$$\mathcal{R}_P^\star \leq \cdots \leq \mathcal{R}_P^{(2k)-\mathrm{ppv}} = \mathcal{R}_P^{(2k-1)-\mathrm{ppv}} \leq \cdots \leq \mathcal{R}_P^{1-\mathrm{ppv}} = \mathbb{E}\left[2\eta(X)\left(1-\eta(X)\right)\right] \leq 2\mathcal{R}_P^\star \ .$$

- Pour tout k ne dépendant pas de n, la règle $\widehat{s}^{k-\mathrm{ppv}}$ n'est pas universellement consistante.
- Si l'on choisit $k = k_n$ de telle sorte que lorsque $\lim_{n\to\infty} k_n = \infty$ et $\lim_{n\to\infty} k_n/n = 0$, alors la règle $\left(\widehat{s}_n^{k_n-\mathrm{ppv}}\right)_{n\in\mathbb{N}}$ est universellement consistante, quelle que soit la distance d.

REMARQUE 1.12. Le fait que la règle des k-ppv soit consistante pour tout k «grand» et toute distance d ne signifie pas que tous ces choix sont équivalents. Par exemple, il existe des distributions P pour lesquelles la meilleure règle $\widehat{s}^{k-\mathrm{ppv}}$ est $\widehat{s}^{1-\mathrm{ppv}}$ (voir Section 3.3.2).

1.5. Choix de k par validation croisée

Intuitivement, pour une distribution P fixée, le choix de k (comme celui de d) est crucial pour que l'excès de risque $\ell\left(s^\star, \widehat{s}^{k-\mathrm{ppv}}\right)$ soit aussi petit que possible (en espérance ou en probabilité) :

- Lorsque η est très régulière vis-à-vis de d et souvent éloignée de 0 et de 1, mieux vaut prendre k grand pour minimiser le risque.

– À l'inverse, lorsque η est très irrégulière vis-à-vis de d et souvent proche de 0 ou de 1, mieux vaut prendre k petit.

Un principe général pour calibrer une règle de classification (ou, plus généralement, pour choisir parmi une famille $(\widehat{s}_\lambda)_{\lambda\in\Lambda}$ de règles concurrentes), est de construire pour chacune de ces règles un estimateur $\widehat{\mathcal{R}}(\widehat{s}_\lambda; D_n)$ de son risque, puis de choisir la règle $\widehat{s}_{\widehat{\lambda}}$ avec

$$\widehat{\lambda} \in \arg\min_{\lambda\in\Lambda} \left\{ \widehat{\mathcal{R}}(\widehat{s}_\lambda; D_n) \right\} \quad . \tag{1.12}$$

REMARQUE 1.13 (Validité de (1.12)). Lorsque Λ est infini ou de de cardinal «très grand», on le fait de manière approchée en se restreignant à un sous-ensemble $\Lambda_0 \subset \Lambda$ fini et de cardinal «raisonnablement grand». Ici, «très grand» signifie de taille $\exp(\alpha n)$ pour un $\alpha > 0$, tandis que «raisonnablement grand» signifie de taille au plus Cn^α pour de «petites» constantes C et α.

Sans faire d'hypothèse sur les règles $\widehat{s}_\lambda$, la manière la plus naturelle pour construire un bon estimateur $\widehat{\mathcal{R}}(\widehat{s}_\lambda; D_n)$ de $\mathcal{R}_P(\widehat{s}_\lambda(D_n))$ est de compter le nombre d'erreurs de classification commises par $\widehat{s}_\lambda(D_n)$ sur un *nouvel échantillon* $(X_{n+1}, Y_{n+1}), \ldots, (X_{n+m}, Y_{n+m})$ pour un certain m suffisamment grand. Malheureusement, par hypothèse, nous n'avons pas plus de données que $D_n = (X_1, Y_1), \ldots, (X_n, Y_n)$.

L'idée est de faire en sorte de disposer de «nouvelles» données en n'entraînant $\widehat{s}_\lambda$ qu'avec une partie de D_n. Ainsi, les observations restantes pourront servir à mesurer la performance de $\widehat{s}_\lambda$.

DÉFINITION 1.7 (Validation). Soit $n \in \mathbb{N}$, $D_n \in (\mathcal{X} \times \mathcal{Y})^n$ un échantillon et $\widehat{s}$ une règle de classification.

Soit $I^{(e)} \subset \{1, \ldots, n\}$ de cardinal $1 \leq n_e \leq n-1$, $I^{(v)} = \{1, \ldots, n\} \setminus I^{(e)}$ son complémentaire et $n_v = n - n_e$. On définit alors l'*échantillon d'entraînement*

$$D_n^{(e)} := (X_i, Y_i)_{i\in I^{(e)}}$$

et l'*échantillon de validation*

$$D_n^{(v)} := (X_i, Y_i)_{i\in I^{(v)}} \quad .$$

L'estimateur par *validation* du risque de $\widehat{s}(D_n)$ est alors défini par

$$\widehat{\mathcal{R}}^{\mathrm{val}}\left(\widehat{s}; D_n; I^{(e)}\right) = \widehat{\mathcal{R}}^{\mathrm{val}}(\widehat{s}; D_n) := \frac{1}{n_v} \sum_{i\in D_n^{(v)}} \mathbb{1}_{\widehat{s}\left(X_i; D_n^{(e)}\right) \neq Y_i} \quad . \tag{1.13}$$

Étant donnée une famille $(\widehat{s}_\lambda)_{\lambda\in\Lambda}$ de classifieurs, on «*choisit* λ *par validation*» en choisissant

$$\widehat{\lambda}^{\mathrm{val}} = \widehat{\lambda}^{\mathrm{val}}\left(D_n; I^{(e)}\right) \in \arg\min_{\lambda\in\Lambda} \left\{ \widehat{\mathcal{R}}^{\mathrm{val}}\left(\widehat{s}; D_n; I^{(e)}\right) \right\} \quad . \tag{1.14}$$

Lorsque cela n'induit pas d'ambiguïté, on n'écrira pas explicitement le fait que $\widehat{\mathcal{R}}^{\mathrm{val}}$ et $\widehat{\lambda}^{\mathrm{val}}$ dépendent de D_n ou de $I^{(e)}$.

REMARQUE 1.14. Pour que $D_n^{(v)}$ soit indépendant de $D_n^{(e)}$, il est nécessaire de choisir $I^{(e)}$ indépendamment des données. En particulier, si les $(X_i, Y_i)_{1\leq i\leq n}$ sont ordonnés en fonction des X_i, il est nécessaire de choisir $I^{(e)}$ après permutation aléatoire des indices.

Le défaut principal de la validation est de dépendre trop fortement du choix arbitraire de $I^{(e)}$. Il s'ensuit que l'estimateur $\widehat{\lambda}^{\mathrm{val}}$ est assez instable. Une manière naturelle de compenser ce défaut est d'estimer le risque de chaque classifieur par validation sur *plusieurs* découpages $I_1^{(e)}, \ldots, I_B^{(e)}$ et de moyenner ces estimations. Une telle méthode est appelée *validation croisée*.

DÉFINITION 1.8 (Validation croisée). Soit $n \in \mathbb{N}$, $D_n \in (\mathcal{X} \times \mathcal{Y})^n$ un échantillon et $\widehat{s}$ une règle de classification. Soit $B \geq 1$, $I_1^{(e)}, \dots, I_B^{(e)}$ une suite (aléatoire ou non) de sous-ensembles stricts de $\{1, \dots, n\}$ non-vides ; pour tout j, on note $I_j^{(v)}$ le complémentaire de $I_j^{(e)}$ et $D_{n,j}^{(e)}$ l'échantillon d'entraînement associé à $I_j^{(e)}$.

L'estimateur par *validation croisée* du risque de $\widehat{s}(D_n)$ est alors défini par

$$\widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{s}; D_n; \left(I_j^{(e)}\right)_{1\leq j\leq B}\right) = \widehat{\mathcal{R}}^{\mathrm{vc}}(\widehat{s}; D_n) := \frac{1}{B}\sum_{j=1}^{B} \widehat{\mathcal{R}}^{\mathrm{val}}\left(\widehat{s}; D_n; I_j^{(e)}\right) \tag{1.15}$$

Étant donnée une famille $(\widehat{s}_\lambda)_{\lambda\in\Lambda}$ de classifieurs, on «*choisit λ par validation croisée*» en choisissant

$$\widehat{\lambda}^{\mathrm{vc}} = \widehat{\lambda}^{\mathrm{vc}}\left(D_n; \left(I_j^{(e)}\right)_{1\leq j\leq B}\right) \in \arg\min_{\lambda\in\Lambda}\left\{\widehat{\mathcal{R}}^{\mathrm{val}}\left(\widehat{s}; D_n; \left(I_j^{(e)}\right)_{1\leq j\leq B}\right)\right\} \ . \tag{1.16}$$

Les exemples de validation croisée les plus utilisés sont les suivants :

(1) «Laisser-un-de-côté» (*leave-one-out*), aussi appelée souvent *validation croisée ordinaire* : $B = n$ et $I_j^{(e)} = \{1, \dots, n\} \setminus \{j\}$ pour tout j. Autrement dit, pour tout j, on entraîne chaque classifieur sur toutes les données sauf (X_j, Y_j), et on teste sa capacité à bien prédire Y_j. On note alors $\widehat{\mathcal{R}}^{\mathrm{loo}}$ l'estimateur du risque et $\widehat{\lambda}^{\mathrm{loo}}$ l'estimateur de λ.

(2) «Laisser-p-de-côté» (*leave-p-out*), avec $p \in \{1, \dots, n-1\}$: $B = C_n^p$ et $\left\{I_1^{(e)}, \dots, I_B^{(e)}\right\}$ est l'ensemble des parties de $\{1, \dots, n\}$ de taille p. On note alors $\widehat{\mathcal{R}}^{\ell\mathrm{po}}$ l'estimateur du risque et $\widehat{\lambda}^{\ell\mathrm{po}}$ l'estimateur de λ.
Noter que le leave-one-out en est un cas particulier ($p = 1$).

(3) *Validation croisée à V blocs* : $B = V$ et pour tout $j \in \{1, \dots, V\}$, $I_j^{(e)} = A_j^c$, où $A_1, \dots, A_V$ est une partition de $\{1, \dots, n\}$ (à choisir arbitrairement).

(4) *Apprentissage-test répétés* : $I_1^{(e)}, \dots, I_B^{(e)}$ sont des sous-ensembles aléatoires indépendants de $\{1, \dots, n\}$, chacun choisi uniformément parmi les sous-ensemble de taille n_e fixée.

REMARQUE 1.15 (Temps de calcul). Les deux premiers types de validation croisée reposent sur une exploration exhaustive des découpages de l'échantillon en deux parties de tailles fixées ($n-p$ et p). Il en résulte que ces algorithmes de calibration sont (très) coûteux en temps de calcul, sauf cas particulier.

À l'inverse, les deux derniers types de validation croisée (ainsi que la validation simple définie précédemment) peuvent être vus comme des approximations des estimateurs leave-p-out, puisqu'ils explorent uniquement un sous-ensemble des découpages en deux parties de tailles fixées. Si le temps de calcul s'en trouve fortement diminué (raison pour laquelle ces méthodes sont souvent préférées en pratique), on s'attend donc à ce que les performances statistiques soient dégradées.

Choisir donc parmi ces algorithmes de validation croisée requiert donc de trouver un bon compromis entre performance statistique et complexité algorithmique. Nous reviendrons sur cette problématique dans un chapitre suivant.

REMARQUE 1.16 (Taille de l'échantillon d'entraînement). Pour les différentes formes de validation croisée comme pour la validation simple, un paramètre dont le choix est crucial est la taille n_e de l'échantillon d'entraînement. En effet, dans le cas de la validation, puisque l'échantillon

d'entraînement $D_n^{(e)}$ est indépendant de l'échantillon de validation, on a

$$\begin{aligned}\mathbb{E}\left[\widehat{\mathcal{R}}^{\mathrm{val}}\left(\widehat{s};D_n;I^{(e)}\right)\right] &= \frac{1}{n_v}\sum_{i\in D_n^{(v)}}\mathbb{E}_{D_n^{(e)}}\mathbb{P}_{(X_i,Y_i)}\left(\widehat{s}\left(X_i;D_n^{(e)}\right)\neq Y_i\right)\\ &= \mathbb{E}_{D_n^{(e)}}\left[\mathcal{R}_P\left(\widehat{s}\left(D_n^{(e)}\right)\right)\right]\ ,\end{aligned} \tag{1.17}$$

et un résultat similaire est valable pour tout algorithme de validation croisée utilisant des échantillons d'entraînement de même taille n_e.

1.6. Minimisation du risque empirique

Lorsque $t\in\mathbb{S}$ est un classifieur donné, un autre estimateur naturel de son risque est le nombre d'erreurs de classification qu'il commet sur de nouvelles données. On l'appelle le *risque empirique*.

Définition 1.9 (Risque empirique). Soit $D_n=(X_i,Y_i)_{1\leq i\leq n}$ un échantillon et $t\in\mathbb{S}$ un classifieur. Le *risque empirique de t sur D_n* est défini par

$$\widehat{\mathcal{R}}_n\left(t;D_n\right)=\widehat{\mathcal{R}}_n\left(t\right):=\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{t(X_i)\neq Y_i}\ . \tag{1.18}$$

Autrement dit, $\widehat{\mathcal{R}}_n\left(t\right)$ est le nombre d'erreurs commises par t sur l'échantillon. Comme chacun des (X_i,Y_i) a pour distribution P, pour tout classifieur t déterministe, le risque empirique de t est un estimateur sans biais du risque de t :

$$\mathbb{E}_{D_n\sim P^{\otimes n}}\left[\widehat{\mathcal{R}}_n\left(t;D_n\right)\right]=\frac{1}{n}\sum_{i=1}^{n}\mathbb{P}_{(X_i,Y_i)\sim P}\left(t(X_i)\neq Y_i\right)=\mathbb{P}_{(X,Y)\sim P}\left(t(X)\neq Y\right)=\mathcal{R}_P\left(t\right)\ .$$

Remarque 1.17 (Mesure empirique). Si

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

désigne la mesure empirique (c'est une mesure de probabilités sur $\mathcal{X}\times\mathcal{Y}$, alors

$$\forall t\in\mathbb{S},\qquad \widehat{\mathcal{R}}_n\left(t\right)=\mathcal{R}_{P_n}\left(t\right)\ .$$

L'avantage de cette seconde expression est qu'elle se généralise sans difficulté à toute définition du risque $\mathcal{R}\left(t\right)$, y compris hors du cadre de la classification binaire.

Une règle de classification naturelle consiste à choisir un classifieur t minimisant son risque empirique. On aboutit aux deux définitions suivantes.

Définition 1.10 (Modèle). Un modèle S est un ensemble de classifieurs, c'est-à-dire une partie de $\mathbb{S}$.

Définition 1.11 (Minimiseur du risque empirique). Soit S un modèle. Alors, la règle de classification $\widehat{s}_S$ de *minimisation du risque empirique sur S* est définie comme suit :

$$\widehat{s}_S(D_n)\in\arg\min_{t\in S}\left\{\widehat{\mathcal{R}}_n\left(t;D_n\right)\right\}\ . \tag{1.19}$$

Remarque 1.18. La règle $\widehat{s}_{\mathbb{S}}$ est en général très mauvaise. En effet, si la loi de X est sans atome, tous les X_i sont différents p.s., si bien que minimiser le risque empirique sur $\mathbb{S}$ revient à attribuer l'étiquette Y_i lorsque $x=X_i$, et n'importe quelle étiquette aux autres $x\in\mathcal{X}$. En particulier, l'un des classifieurs réalisant le minimum dans (1.19) a un risque égal à $1-\mathcal{R}^\star>\mathcal{R}^\star$.

EXEMPLE 1.19 (Histogrammes). Pour toute partition $I_1, \dots, I_K$ de $\mathcal{X}$, on définit le modèle d'histogrammes suivant :

$$S_{\text{histo}}(I_1, \dots, I_K) := \left\{ \sum_{j=1}^{K} a_j \mathbb{1}_{x \in I_j} \text{ t.q. } a_1, \dots, a_K \in \{0,1\} \right\} .$$

En particulier, si $\mathcal{X} = [0,1]^\ell$, un modèle particulier d'histogrammes est le modèle associé à la partition de $\mathcal{X}$ en une grille régulière de pas $h \in (0,1)$ avec $h^{-1} \in \mathbb{N}$; son cardinal est $K = h^{-\ell}$.

EXEMPLE 1.20 (Séparation linéaire). Si $\mathcal{X} = \mathbb{R}^\ell$ avec $\ell \geq 1$ un entier, alors, le modèle des *séparateurs linéaires* dans $\mathcal{X}$ est défini par

$$S_{\text{lin}} := \left\{ x \mapsto \mathbb{1}_{a_0 + a^T x > 0} \text{ t.q. } a_0 \in \mathbb{R},\, a \in \mathbb{R}^\ell \right\} .$$

On peut également considérer des sous-ensembles de S_{lin} en imposant que a appartienne à un sous-espace vectoriel de $\mathbb{R}^\ell$, ou bien que le nombre de coordonnées non-nulles de a soit au plus égal à $s < \ell$. C'est en particulier nécessaire lorsque $\ell \geq n$, car n points en position générale dans $\mathbb{R}^n$ sont toujours séparables par un hyperplan, indépendamment des valeurs des étiquettes Y_i.

.

REMARQUE 1.21 (Temps de calcul). Dans le cas des histogrammes, calculer S_{histo} requiert $\mathcal{O}(n)$ opérations (dans chacun des I_j, on fait voter les observations pour déterminer la valeur du classifieur) ; il faut ensuite au maximum $\mathcal{O}(K)$ opérations pour les stocker.

En général, comme $t \mapsto \mathbb{1}_{t(x) \neq y}$ n'est pas une fonction convexe, c'est un problème difficile.

REMARQUE 1.22 (Calibration). En général, il ne faut pas utiliser le risque empirique pour la calibration. Par exemple, supposons que l'on souhaite choisir k parmi les règles des k-ppv, une distance d sur $\mathcal{X}$ étant fixée. Utiliser le risque empirique pour choisir k revient à choisir

$$\widehat{k} \in \arg\min_{1 \leq k \leq n} \left\{ \widehat{\mathcal{R}}_n \left(\widehat{s}^{k-\text{ppv}}(D_n); D_n \right) \right\} ,$$

et donc $\widehat{k} = 1$ (avec peut-être des valeurs de $k > 1$ ex-aequo).

Dans le cas zéro-erreur ($Y = s^\star(X)$), cela peut donner un bon résultat. Mais en général, cela conduit au *sur-apprentissage*.

Le même problème se produit pour choisir parmi des minimiseurs du risque empirique sur différents modèles. Nous verrons dans la section suivante comment corriger ce défaut du risque empirique.

1.7. Sélection de modèle

1.7.1. Décomposition approximation–estimation. Choisir un modèle S sur lequel réaliser la minimisation du risque empirique est problématique. En effet, lorsque S est très petit, la contrainte $\widehat{s}_S \in S$ implique en général que le risque de $\widehat{s}_S$ est grand. À l'inverse, lorsque S est très grand (par exemple, $S = \mathbb{S}$), minimiser le risque empirique revient à «coller» aux données et ne tient pas compte du fait que Y n'est pas nécessairement une fonction déterministe de X. Il est donc nécessaire de trouver un compromis entre ces deux situations extrêmes.

Cette problématique se comprend formellement via la décomposition suivante du risque :

$$\mathcal{R}_P(\widehat{s}_S(D_n)) - \mathcal{R}_P^\star = \left[\mathcal{R}_P(\widehat{s}_S(D_n)) - \inf_{t \in S} \{\mathcal{R}_P(t)\} \right] + \left[\inf_{t \in S} \{\mathcal{R}_P(t)\} - \mathcal{R}_P^\star \right] . \quad (1.20)$$

Supposons, pour simplifier, que l'inf du risque sur S est atteint en un certain $s_S^\star \in S$ (pas nécessairement unique). Alors, le premier terme de (1.20) s'écrit

$$\mathcal{R}_P\left(\widehat{s}_S(D_n)\right) - \mathcal{R}_P\left(s_S^\star\right) \tag{1.21}$$

et est appelé *erreur d'estimation*. Il est en général d'autant plus grand que S est grand. Par ailleurs, le second terme de (1.20) s'écrit

$$\mathcal{R}_P\left(s_S^\star\right) - \mathcal{R}_P^\star = \ell\left(s^\star, s_S^\star\right) \tag{1.22}$$

et est appelé *erreur d'approximation*. Il est en d'autant plus petit que S est grand. Choisir un bon modèle S revient donc à réaliser un *compromis entre estimation et approximation*.

Remarque 1.23 (Consistance). L'erreur d'approximation ne dépendant pas de n, $\widehat{s}_S$ ne peut être consistant que si $\ell\left(s^\star, s_S^\star\right) = 0$, ce qui ne peut pas être universellement vrai à moins que S ne soit quasi égal à $\mathbb{S}$ (auquel cas l'erreur d'estimation ne convergera en général pas vers 0). Par conséquent, $\widehat{s}_S$ avec un modèle S fixe *n'est pas universellement consistant*.

1.7.2. Borne générale sur l'erreur d'estimation. Soit S un modèle et $\widehat{s}_S$ un minimiseur du risque empirique sur S. Alors,

$$\begin{aligned}
\mathcal{R}_P\left(\widehat{s}_S(D_n)\right) - \mathcal{R}_P\left(s_S^\star\right) &= \left[\widehat{\mathcal{R}}_n\left(\widehat{s}_S(D_n)\right)P - \widehat{\mathcal{R}}_n\left(s_S^\star\right)P\right] + \left[\mathcal{R}_P\left(\widehat{s}_S(D_n)\right) - \widehat{\mathcal{R}}_n\left(\widehat{s}_S(D_n)\right)P\right] \\
&\quad + \left[\widehat{\mathcal{R}}_n\left(s_S^\star\right)P - \mathcal{R}_P\left(s_S^\star\right)\right] \\
&\leq \sup_{t\in S}\left\{\mathcal{R}_P(t) - \widehat{\mathcal{R}}_n(t)\right\} + \sup_{t\in S}\left\{\widehat{\mathcal{R}}_n(t) - \mathcal{R}_P(t)\right\} \\
&\leq 2\sup_{t\in S}\left|\mathcal{R}_P(t) - \widehat{\mathcal{R}}_n(t)\right| \quad .
\end{aligned}$$

Notons que cette borne supérieure sur l'erreur d'estimation est une fonction croissante de S. Bien que cette majoration ne soit pas toujours très précise, elle l'est la plupart du temps et fournit un bon indicateur de la «complexité» du modèle S.

Nous verrons au chapitre 4 différentes mesures de complexité d'un modèle S et leur lien avec $\sup_{t\in S}\left|\mathcal{R}_P(t) - \widehat{\mathcal{R}}_n(t)\right|$.

1.7.3. Choix de modèle par pénalisation. Le modèle S étant un paramètre de la règle $\widehat{s}_S$, on peut utiliser la validation (croisée ou non) pour choisir un modèle parmi une famille $(S_m)_{m\in\mathcal{M}}$. Cependant, une autre approche est possible, appelée *pénalisation*. Elle est particulièrement naturelle pour choisir parmi des minimiseurs du risque empirique, et est en réalité une généralisation des méthodes de choix de modèle par validation ou validation croisée.

Définition 1.12 (Pénalisation). Soit $(S_m)_{m\in\mathcal{M}}$ une collection au plus dénombrable de modèles et pen : $\mathcal{M} \mapsto \mathbb{R}$ une fonction appelée *pénalité*. Alors, le minimiseur du risque empirique pénalisé est $\widehat{s}_{\widehat{m}}$ avec

$$\widehat{m} \in \arg\min_{m\in\mathcal{M}}\left\{\widehat{\mathcal{R}}_n\left(\widehat{s}_{S_m}\right) + \mathrm{pen}(m)\right\} \quad . \tag{1.23}$$

Exemple 1.24. Si S_m est fini, on peut utiliser la pénalité :

$$\mathrm{pen}(m) = L\sqrt{\frac{\ln\left(\operatorname{Card} S_m\right)}{n}} \quad .$$

Si S_m est un modèle d'histogrammes sur une partition de taille K_m, on peut utiliser :

$$\mathrm{pen}(m) = L\sqrt{\frac{K_m \ln(en)}{n}}$$

Si S_m est un modèle de séparateurs linéaires avec $s_m \geq 0$ coefficients non-nuls, on peut utiliser :

$$\mathrm{pen}(m) = L\sqrt{\frac{(s_m+1)\ln(en)}{n}}$$

Plus généralement, la pénalité doit être (à constante multiplicative près) une mesure de complexité appropriée pour le modèle S_m.

L'objectif est de choisir un classifieur $\widehat{s}_{S_{\widehat{m}}}$ ayant un risque (presque) aussi bon que celui de l'*oracle*, défini comme le minimiseur de $\mathcal{R}_P(\widehat{s}_{S_m})$ en $m \in \mathcal{M}$. Lorsque c'est le cas, on parle d'inégalité oracle.

DÉFINITION 1.13 (Inégalité oracle). On dit que $\widehat{m}$ satisfait une *inégalité oracle* lorsque, en espérance ou avec «grande» probabilité,

$$\ell(s^\star, \widehat{s}_{S_{\widehat{m}}}) \leq C \inf_{m\in\mathcal{M}} \{\mathcal{R}_P(\widehat{s}_{S_m}) + \epsilon_m\} \quad . \tag{1.24}$$

Une telle inégalité est d'autant meilleure que C est proche de 1 et ϵ_m est négligeable (ou du même ordre que) $\mathrm{pen}(m)$.

D'après (1.23), la meilleure pénalité est la différence entre le risque de $\widehat{s}_{S_m}$ et son risque empirique. On l'appelle pénalité idéale :

DÉFINITION 1.14 (Pénalité idéale). Soit S_m un modèle. La pénalité idéale de m est alors définie par

$$\mathrm{pen}_{\mathrm{id}}(m) := \mathcal{R}_P(\widehat{s}_{S_m}) - \widehat{\mathcal{R}}_n(\widehat{s}_{S_m}) \quad . \tag{1.25}$$

Évidemment, la pénalité idéale n'est pas accessible en pratique, mais nous verrons au Chapitre 5 qu'une pénalité estimant bien la pénalité idéale conduit à une inégalité oracle.

1.8. D'autres exemples de règles de classification

Parmi de nombreux exemples (voir notamment **[DGL96, HTF01]**), citons ici :

(1) classifieurs plug-in :
$$\widehat{s}(x) = \mathbb{1}_{\widehat{\eta}(x)\geq 1/2} \ ,$$
où $\widehat{\eta}$ est un estimateur de η.

(2) classifieurs par moyennage local (voir chapitre 3)

(3) Si $\mathcal{X} = \mathbb{R}^\ell$, un *réseau de neurones* à une couche cachée est un minimiseur du risque empirique sur
$$S(k_{\max}, \beta) = \left\{ x \in \mathbb{R}^\ell \mapsto \sum_{j=1}^{k} c_j \sigma\left(a_j (1, x_1, \cdots, x_\ell)^T\right) + c_0 \text{ t.q. } k \leq k_{\max}, \right.$$
$$\left. a_1, \ldots, a_k \in \mathbb{R}^{\ell+1},\ c_0, c_1, \ldots, c_k \in \mathbb{R} \text{ et } \sum_{j=0}^{\ell} |c_j| \leq \beta \right\} \ ,$$
où $\sigma : \mathcal{X} \mapsto \mathbb{R}$ est une fonction croissante telle que $\lim_{-\infty}\sigma = 0$ et $\lim_{+\infty}\sigma = 1$ (appelée *sigmoïde*). Par exemple, $\sigma(x) = \mathbb{1}_{x\geq 0}$ ou $\sigma(x) = 1/(1+e^{-x})$.

(4) machines à vecteurs de support (SVM) : voir **[STC00]**

(5) arbres de décision (CART) : voir **[BFOS84]**

CHAPITRE 2

Quelques résultats généraux sur la classification

L'objectif principal de ce chapitre est de montrer deux résultats qui permettent de mieux identifier ce qui est facile et ce qui est impossible en classification supervisée.

D'une part, le Théorème 2.1 montre un exemple de problème «simple» ($\mathcal{X}$ fini de cardinal négligeable par rapport au nombre d'observations n) pour lequel une règle très naïve fonctionne uniformément bien. Dans de telles situations, il est donc inutile d'utiliser des algorithmes complexes, qui risquent de ne pas avoir une meilleure performance statistique pour un temps de calcul bien plus élevé.

D'autre part, le Théorème 2.3 montre qu'il n'y a en général pas de règle de classification uniformément bonne. Par conséquent, pour chaque problème particulier, il faut identifier un (petit) ensemble de distributions $\mathcal{P}$ auquel P est susceptible d'appartenir, puis utiliser une règle de classification dont on aura pu montrer qu'elle fonctionne (uniformément) «mieux» que les autres lorsque $P \in \mathcal{P}$. Par exemple, la minimisation du risque empirique pénalisé est un moyen d'utiliser une telle information sur $\mathcal{P}$, *via* le choix d'une famille de modèles $(S_m)_{m\in\mathcal{M}}$ adéquate.

2.1. Consistance universelle uniforme lorsque $\mathcal{X}$ est fini

THÉORÈME 2.1. *Soit $\mathcal{X}$ un ensemble de cardinal fini $K \geq 1$. Soit la règle de classification $\widehat{s}^{\mathrm{maj}}$ (règle de majorité) définie comme suit : pour tout $n \in \mathbb{N}$, $n \geq 1$, pour tout $x \in \mathcal{X}$,*

$$\widehat{s}_n^{\mathrm{maj}}(x) := \begin{cases} 1 & \text{si } \mathrm{Card}\left\{ i \text{ t.q. } X_i = x \text{ et } Y_i = 1 \right\} \geq \mathrm{Card}\left\{ i \text{ t.q. } X_i = x \text{ et } Y_i = 0 \right\} \\ 0 & \text{sinon.} \end{cases}$$

Alors, pour tout $n \in \mathbb{N}$,

$$\sup_P \left(\mathbb{E}_{D_n \sim P^{\otimes n}} \left[\mathcal{R}_P\left(\widehat{s}_n^{\mathrm{maj}}(D_n) \right) \right] - \mathcal{R}_P^\star \right) \leq \frac{\sqrt{K \ln(2)} + \sqrt{\pi}}{\sqrt{2n}} \ . \tag{2.1}$$

En particulier, $\widehat{s}^{\mathrm{maj}}$ est uniformément universellement consistante.

REMARQUE 2.1. La règle de majorité $\widehat{s}^{\mathrm{maj}}$ est l'une des règles minimisant le risque empirique sur $\mathbb{S}$. On a juste précisé l'étiquette à attribuer à un point x où l'on a observé autant de fois $Y = 1$ et $Y = 0$.

REMARQUE 2.2. On peut montrer que (2.1) est vérifié avec une borne supérieure plus petite, proportionnelle à $K/\sqrt{n}$. Le début de la preuve est identique, le résultat s'obtenant grâce à une majoration fine de $\sup_{t\in S} \left| \widehat{\mathcal{R}}_n(t) - \mathcal{R}_P(t) \right|$ pour S fini que nous verrons dans le chapitre 4.

Énonçons tout d'abord deux résultats dont nous avons besoin pour prouver le Théorème 2.1. Le premier est un résultat général sur les règles qui minimisent le risque empirique.

LEMME 2.1 (Vapnik et Chervonenkis, 1974 **[VC74]**). *Soit $\widehat{s}$ une règle minimisant le risque empirique sur $S \subset \mathbb{S}$. Alors,*

$$\mathcal{R}_P(\widehat{s}) - \inf_{t\in S} \mathcal{R}_P(t) \leq 2 \sup_{t\in S} \left| \widehat{\mathcal{R}}_n(t) - \mathcal{R}_P(t) \right| \ . \tag{2.2}$$

DÉMONSTRATION. Soit $\epsilon > 0$ et $t_\epsilon \in S$ tel que

$$\mathcal{R}_P(t_\epsilon) \leq \inf_{t\in S} \mathcal{R}_P(t) + \epsilon \ .$$

Par définition de t_ϵ et de $\widehat{s}$,

$$\begin{aligned}\mathcal{R}_P(\widehat{s}) - \inf_{t\in S} \mathcal{R}_P(t) &= \mathcal{R}_P(\widehat{s}) - \widehat{\mathcal{R}}_n(\widehat{s}) + \widehat{\mathcal{R}}_n(\widehat{s}) - \inf_{t\in S} \mathcal{R}_P(t) \\ &\leq \mathcal{R}_P(\widehat{s}) - \widehat{\mathcal{R}}_n(\widehat{s}) + \widehat{\mathcal{R}}_n(t_\epsilon) - \mathcal{R}_P(t_\epsilon) + \epsilon \\ &\leq 2\sup_{t\in S} \left|\widehat{\mathcal{R}}_n(t) - \mathcal{R}_P(t)\right| + \epsilon \ .\end{aligned}$$

Comme ϵ peut-être pris arbitrairement proche de zéro, on en déduit le résultat. □

Le second résultat est l'inégalité de Hoeffding, déjà prouvée et utilisée dans le cadre du cours sur la prédiction de suites individuelles. Rappelons-la ici, dans le cas particulier de variables indépendantes.

THÉORÈME 2.2. *Soient $\xi_1, \ldots, \xi_n$ des variables aléatoires indépendantes telles que pour tout i, $a_i \leq \xi_i \leq b_i$ p.s. pour des réels $a_1, \ldots, a_n$ et $b_1, \ldots, b_n$. Alors,*

$$\mathbb{P}\left(\sum_{i=1}^n \xi_i - \sum_{i=1}^n \mathbb{E}[\xi_i] \geq \epsilon\right) \leq \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) \ . \tag{2.3}$$

REMARQUE 2.3. En appliquant ce même résultat aux variables $-\xi_i$, on obtient

$$\mathbb{P}\left(\left|\sum_{i=1}^n \xi_i - \sum_{i=1}^n \mathbb{E}[\xi_i]\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) \ . \tag{2.4}$$

DÉMONSTRATION DU THÉORÈME 2.1. La règle de majorité minimisant le risque empirique sur $S = \mathbb{S}$, on a $\inf_{t\in S} \mathcal{R}_P(t) = \mathcal{R}_P^\star$ et le Lemme 2.1 montre qu'il suffit de majorer l'espérance de

$$2\sup_{t\in\mathbb{S}} \left|\widehat{\mathcal{R}}_n(t) - \mathcal{R}_P(t)\right| \ .$$

Or, pour tout $t \in \mathbb{S}$, l'inégalité de Hoeffding (2.4) appliquée aux variables $\xi_i = n^{-1}\mathbb{1}_{t(X_i)\neq Y_i} \in [0; n^{-1}]$ qui sont indépendantes et d'espérance $n^{-1}\mathcal{R}_P(t)$, montre que

$$\mathbb{P}\left(\left|\widehat{\mathcal{R}}_n(t) - \mathcal{R}_P(t)\right| \geq \epsilon\right) \leq 2\exp\left(-2n\epsilon^2\right) \ .$$

Comme $\mathbb{S}$ est fini de cardinal 2^K, on en déduit que

$$\mathbb{P}\left(\sup_{t\in\mathbb{S}} \left|\widehat{\mathcal{R}}_n(t) - \mathcal{R}_P(t)\right| \geq \epsilon\right) \leq \sum_{t\in\mathbb{S}} \mathbb{P}\left(\left|\widehat{\mathcal{R}}_n(t) - \mathcal{R}_P(t)\right| \geq \epsilon\right) \leq 2^{K+1}\exp\left(-2n\epsilon^2\right) \ .$$

Remarquons que pour tout $u \geq 0$

$$\begin{aligned}
\mathbb{E}\left[\sup_{t\in\mathbb{S}}\left|\widehat{\mathcal{R}}_n(t) - \mathcal{R}_P(t)\right|\right] &= \int_0^\infty \mathbb{P}\left(\sup_{t\in\mathbb{S}}\left|\widehat{\mathcal{R}}_n(t) - \mathcal{R}_P(t)\right| \geq \epsilon\right) \mathrm{d}\epsilon \\
&\leq u + \int_u^\infty \mathbb{P}\left(\sup_{t\in\mathbb{S}}\left|\widehat{\mathcal{R}}_n(t) - \mathcal{R}_P(t)\right| \geq \epsilon\right) \mathrm{d}\epsilon \\
&\leq u + 2^{K+1}\int_u^\infty \exp\left(-2n\epsilon^2\right) \mathrm{d}\epsilon \\
&\leq u + 2^{K+1}e^{-2nu^2}\int_0^\infty \exp\left(-2n\epsilon^2\right) \mathrm{d}\epsilon \\
&= u + e^{-2nu^2}\frac{2^K\sqrt{\pi}}{\sqrt{2n}} \ .
\end{aligned}$$

Le résultat s'en déduit en prenant $u = \sqrt{K\ln(2)/(2n)}$. □

Il n'est en revanche pas évident que la consistance universelle (uniforme ou pas, faible ou forte) soit possible lorsque $\mathcal{X}$ est infini, comme le souligne la section suivante. Nous verrons plus loin dans ce cours qu'il existe des règles (non-uniformément) universellement consistantes pour $\mathcal{X}$ infini, le premier résultat du genre n'ayant été démontré qu'en 1977 par Stone [**Sto77**].

2.2. Pas de règle uniformément universellement consistante si $\mathcal{X}$ est infini

THÉORÈME 2.3 (No Free Lunch Theorem). *Si $\mathcal{X}$ est infini, alors pour tout $n \in \mathbb{N}$ et tout classifieur $\widehat{s} : (\mathcal{X}\times\mathcal{Y})^n \mapsto \mathbb{S}$,*

$$\sup_P \left\{ \mathbb{E}_{D_n\sim P^{\otimes n}}\left[\mathcal{R}_P\left(\widehat{s}\left(D_n\right)\right)\right] - \mathcal{R}_P^\star \right\} \geq \frac{1}{2} > 0 \ , \tag{2.5}$$

le sup *étant pris sur l'ensembles des mesures de probabilité sur $\mathcal{X}\times\mathcal{Y}$. En particulier, aucune règle de classification ne peut être uniformément universellement consistante lorsque $\mathcal{X}$ est infini.*

DÉMONSTRATION. Soit $n, K \in \mathbb{N}$, $\widehat{s} : (\mathcal{X}\times\mathcal{Y})^n \mapsto \mathbb{S}$ un classifieur. L'espace $\mathcal{X}$ étant infini, il contient au moins K points distincts $A_1, \ldots, A_K$.

Pour tout $q \in \{0,1\}^K$, notons P_q la distribution de probabilité sur $\mathcal{X}\times\mathcal{Y}$ définie par $\mathbb{P}_{(X,Y)\sim P}(X = A_j \text{ et } Y = q_j) = K^{-1}$ pour tout $j \in \{1,\ldots,K\}$. Autrement dit, X est choisi uniformément parmi les A_j, et Y est une fonction déterministe g_q de X. En particulier, pour tout q, $\mathcal{R}^\star_{P_q} = 0$.

L'idée de la preuve est que si $P = P_q$, le classifieur $\widehat{s}$ ne disposant que de n observations, il va devoir «deviner» au moins $K - n$ valeurs de Y correspondant aux X non-observés.

La preuve d'existence d'une «mauvaise» distribution P_q pour tout classifieur $\widehat{s}$ repose sur un argument probabiliste : si le risque $\mathbb{E}_{q\sim Q}\left[\mathcal{R}_{P_q}\left(\widehat{s}\right)\right]$ est grand, alors il existe un q tel que $\mathcal{R}_{P_q}\left(\widehat{s}\right)$ est grand.

Plus précisément, pour tout $q \in \{0,1\}^K$ (déterministe), on pose

$$F(q) = \mathbb{E}_{(X_i,Y_i)_{1\leq i\leq n}\sim P_q^{\otimes n}}\left[\mathcal{R}_{P_q}\left(\widehat{s}\left(D_n\right)\right)\right] \ .$$

La remarque-clé est que pour toute distribution de probabilité Q sur $\{0,1\}^K$,

$$\sup_{q\in\{0,1\}^K}\left\{F(q)\right\} \geq \mathbb{E}_{q\sim Q}\left[F(q)\right] \ .$$

Notons Q la distribution uniforme sur $\{0,1\}^K$, de telle sorte que $q \sim Q$ signifie que $q_1, \dots, q_K$ sont indépendantes et de même distribution Bernoulli $\mathcal{B}(1/2)$. Alors,

$$\begin{aligned}
\mathbb{E}_{q\sim Q}\left[F(q)\right] &= \mathbb{P}\left(\widehat{s}\left(X; D_n\right) \neq Y\right) \\
&= \mathbb{P}\left(\widehat{s}\left(X; D_n\right) \neq g_q(X)\right) \\
&= \mathbb{E}_{X,X_1,\dots,X_n}\left[\mathbb{P}_{q\sim Q}\left(\widehat{s}\left(X; D_n\right) \neq g_q(X) \,|\, X, X_1, \dots, X_n\right)\right] \\
&\geq \mathbb{E}_{X,X_1,\dots,X_n}\left[\frac{1}{2}\mathbb{1}_{\{X\neq X_1,\dots,X\neq X_n\}}\right] \\
&= \frac{1}{2}\mathbb{E}_X\left[\mathbb{P}\left(X_1 \neq X, \dots, X \neq X_n \,|\, X\right)\right] \\
&= \frac{1}{2}\mathbb{E}_X\left[\left(1-\frac{1}{K}\right)^n\right] = \frac{1}{2}\left(1-\frac{1}{K}\right)^n \quad .
\end{aligned}$$

Pour tout $n \in \mathbb{N}$ fixé, cette borne inférieure tend vers $1/2$ lorsque K tend vers 1, d'où le résultat. □

REMARQUE 2.4. Si $\mathcal{X}$ est infini, une règle de classification optimale «en pire cas», au sens où elle minimise

$$\sup_P \left\{\mathbb{E}_{D_n\sim P^{\otimes n}}\left[\mathcal{R}_P\left(\widehat{s}\left(D_n\right)\right)\right] - \mathcal{R}_P^\star\right\} \quad ,$$

est la règle de classification aléatoire «pile ou face» définie par

$$\widehat{s}^{\text{pile ou face}}(X; D_n) = B_X \quad ,$$

où $(B_X)_{x\in\mathcal{X}}$ sont des variables indépendantes entre elles (et indépendantes de D_n) suivant une distribution de Bernoulli de paramètre $1/2$. Par conséquent, l'optimalité «en pire cas» est un très mauvais critère pour choisir une règle de classification.

REMARQUE 2.5. Le même type de preuve montre que si $\widehat{s}_1$ et $\widehat{s}_2$ sont deux classifieurs, alors,

$$\mathbb{E}_q\mathbb{E}_{D_n\sim P_q^{\otimes n}}\left[\mathcal{R}_{P_q}\left(\widehat{s}_2\right)\right] + \epsilon \geq \mathbb{E}_q\mathbb{E}_{D_n\sim P_q^{\otimes n}}\left[\mathcal{R}_{P_q}\left(\widehat{s}_1\right)\right] \geq \mathbb{E}_q\mathbb{E}_{D_n\sim P_q^{\otimes n}}\left[\mathcal{R}_{P_q}\left(\widehat{s}_2\right)\right] - \epsilon \quad ,$$

où q suit une distribution uniforme sur $\{0,1\}^K$ pour un entier $K(\epsilon)$ assez grand. Autrement dit, tous les classifieurs sont équivalents en moyenne, la conséquence étant qu'aucun classifieur n'est uniformément meilleur que les autres. Tout gain de performances pour certaines distributions P (par exemple, celles pour lesquelles q_i est une fonction «régulière» de A_i) induit une perte de performance équivalente pour d'autres distributions. Le jeu est donc de caractériser ce qu'est un problème de classification «raisonnable» et de chercher un classifieur se comportant bien parmi cette classe de problèmes.

2.3. Estimateur plug-in

Supposons désormais que $\widehat{s}$ est un estimateur de type «plug-in», c'est-à-dire tel que

$$\widehat{s}(x) = \mathbb{1}_{\widehat{\eta}(x)\geq 1/2} \quad ,$$

où $\widehat{\eta}$ est un estimateur de η.

Alors, d'après (1.4), on a

$$\begin{aligned}
\mathcal{R}_P\left(\widehat{s}(D_n)\right) - \mathcal{R}_P\left(s^\star\right) &= \mathbb{E}\left[\left|2\eta(X)-1\right| \mathbb{1}_{\widehat{s}(X)\neq s^\star(X)} \,\middle|\, D_n\right] \\
&\leq 2\mathbb{E}\left[\left|\eta(X) - \widehat{\eta}(X)\right| \,|\, D_n\right]
\end{aligned}$$

puisque $\widehat{s}(x) \neq s^\star(x)$ entraîne que $|\eta(x) - \widehat{\eta}(x)| \geq |\eta(x) - 1/2|$.

En intégrant et en utilisant l'inégalité de Jensen (la fonction $x \mapsto x^2$ étant convexe), on obtient

$$\mathbb{E}\left[\mathcal{R}_P\left(\widehat{s}(D_n)\right) - \mathcal{R}_P\left(s^\star\right)\right] \leq 2\sqrt{\mathbb{E}\left[\left(\eta(X) - \widehat{\eta}(X)\right)^2\right]} \ . \tag{2.6}$$

Autrement dit, si $\widehat{\eta}$ est un estimateur consistant de η au sens du risque des moindre carrés, alors l'estimateur plug-in associé est consistant.

CHAPITRE 3

Règles par moyennage local

L'objet de ce chapitre est d'étudier la famille des règles de classification dites "par moyennage local", dont l'un des exemples les plus connus est la règle des k plus proches voisins.

3.1. Définition, exemples

DÉFINITION 3.1 (Règles par moyennage local). Pour tout $x \in \mathcal{X}$, soient

$$W_1(x) = W_1(x; n; X_1, \ldots, X_n), \ldots, W_n(x) = W_n(x; n; X_1, \ldots, X_n)$$

des réels positifs de somme 1 bien choisis. Une *règle de classification par moyennage local* est définie par

$$\widehat{s}_n\left(x; (X_i, Y_i)_{1 \le i \le n}\right) := \mathbb{1}_{\widehat{\eta}_n(x) \ge \frac{1}{2}} \qquad \text{où} \quad \widehat{\eta}_n(x) := \sum_{i=1}^{n} W_i(x; n; X_1, \ldots, X_n) Y_i \ . \tag{3.1}$$

EXEMPLE 3.1 (k plus proches voisins). La règle des k-ppv introduite en Section 1.4au premier cours est un cas particulier de règle par moyennage local, en prenant les poids

$$W_i(x) = \frac{1}{k} \mathbb{1}_{\{\, X_i \text{ fait partie des } k \text{ plus proches voisins de } x \text{ dans } X_1, \ldots, X_n \,\}} \ .$$

EXEMPLE 3.2 (partition). Soit $A_1, \ldots, A_k, \ldots$ une partition finie ou dénombrable de $\mathcal{X}$. Pour tout $x \in \mathcal{X}$, on note $A(x)$ l'élément de la partition qui contient x. La règle de classification par partition est définie par (3.1) avec les poids

$$W_i(x) = \frac{\mathbb{1}_{X_i \in A(x)}}{\sum_{j=1}^{n} \mathbb{1}_{X_j \in A(x)}} \ .$$

EXEMPLE 3.3 (noyau). Soient $K : \mathbb{R}^\ell \mapsto \mathbb{R}^+$ (noyau) et $h > 0$ (largeur du noyau, ou fenêtre). La règle de classification par noyau est définie par (3.1) avec les poids

$$W_i(x) = \frac{K\left(\frac{X_i - x}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{X_j - x}{h}\right)} \ ,$$

avec la convention $0/0 = 0$. Les exemples les plus courants de noyau K sont :

– le noyau fenêtre $K(x) = \mathbb{1}_{\|x\| \le 1}$

– le noyau gaussien $K(x) = \exp\left(-\|x\|^2\right)$,

où $\|\cdot\|$ est la norme euclidienne sur $\mathcal{X} = \mathbb{R}^\ell$.

3.2. Consistance universelle

3.2.1. Résultat général.

THÉORÈME 3.1 (Consistance faible pour les règles par moyennage local, Stone **[Sto77]**). *On suppose que $\mathcal{X} = \mathbb{R}^\ell$ est muni d'une norme $\|\cdot\|$, et que pour toute distribution de X sur $\mathcal{X}$, les poids W_j satisfont les trois conditions suivantes :*

(i) *Il existe une constante $c > 0$ telle que pour toute fonction mesurable $f : \mathcal{X} \mapsto [0, +\infty[$ telle que $\mathbb{E}[f(X)] < \infty$,*

$$\mathbb{E}\left[\sum_{j=1}^{n} W_j(X; n; X_1, \ldots, X_n) f(X_j)\right] \leq c\mathbb{E}[f(X)] \ .$$

(ii) *Pour tout $a > 0$,*

$$\lim_{n\to\infty} \mathbb{E}\left[\sum_{j=1}^{n} W_j(X; n; X_1, \ldots, X_n) \mathbb{1}_{\|X_j - X\| \geq a}\right] = 0 \ .$$

(iii)

$$\lim_{n\to\infty} \mathbb{E}\left[\max_{1\leq j\leq n} W_j(X; n; X_1, \ldots, X_n)\right] = 0 \ .$$

Alors, la règle $\widehat{s}_n$ définie par (3.1) *est (faiblement) universellement consistante.*

Démonstration. Pour simplifier, on notera $W_j(x; D_n)$ au lieu de $W_j(x; n; X_1, \ldots, X_n)$ tout au long de cette preuve. Soit P une mesure de probabilité sur $\mathcal{X} \times \mathcal{Y}$, et μ la loi de X avec $(X, Y) \sim P$.

La règle $\widehat{s}_n$ est de type "plug-in". D'après (2.6), il suffit donc de montrer que,

$$\mathbb{E}\left[(\eta(X) - \widehat{\eta}_n(X; D_n))^2\right] = \|\eta - \widehat{\eta}_n\|^2_{L^2(\mu\otimes P^{\otimes n})} \xrightarrow[n\to\infty]{} 0 \ .$$

Posons, pour tout $x \in X$,

$$\eta^\star_n(x; D_n) := \sum_{j=1}^{n} \eta(X_j) W_j(x; D_n) \ .$$

D'après l'inégalité triangulaire,

$$\|\eta - \widehat{\eta}_n\|_{L^2} \leq \|\eta - \eta^\star_n\|_{L^2} + \|\eta^\star_n - \widehat{\eta}_n\|_{L^2} \ , \tag{3.2}$$

où l'on a noté $\|\cdot\|_{L^2}$ pour $\|\cdot\|^2_{L^2(\mu\otimes P^{\otimes n})}$. Il suffit de montrer que chacun des deux termes du membre de droite de (3.2) tend vers zéro quand n tend vers l'infini.

Pour le premier terme, notons que pour tout $x \in \mathcal{X}$,

$$(\eta(x) - \eta^\star_n(x))^2 = \left(\sum_{j=1}^{n} W_j(x; D_n)(\eta(x) - \eta(X_j))\right)^2 \leq \sum_{j=1}^{n} W_j(x; D_n)(\eta(x) - \eta(X_j))^2 \tag{3.3}$$

en utilisant l'inégalité de Jensen et le fait que les W_j sont positifs et de somme 1.

Supposons dans un premier temps que η est uniformément continue. Pour tout $\epsilon > 0$, il existe $a > 0$ tel que

$$\sup_{x_1, x_2 \in \mathcal{X},\, \|x_1 - x_2\| \leq a} |\eta(x_1) - \eta(x_2)| \leq \epsilon \ .$$

En intégrant (3.3) par rapport à x, on obtient

$$\begin{aligned}\|\eta - \eta_n^\star\|_{L^2}^2 &\leq \mathbb{E}\left[\sum_{j=1}^n W_j(X;D_n)\left(\eta(X)-\eta(X_j)\right)^2\right] \\ &= \mathbb{E}\left[\sum_{j=1}^n W_j(X;D_n)\mathbb{1}_{\|X-X_j\|<a}\left(\eta(X)-\eta(X_j)\right)^2\right] \\ &\quad + \mathbb{E}\left[\sum_{j=1}^n W_j(X;D_n)\mathbb{1}_{\|X-X_j\|\geq a}\left(\eta(X)-\eta(X_j)\right)^2\right] \\ &\leq \epsilon^2 + \mathbb{E}\left[\sum_{j=1}^n W_j(X;D_n)\mathbb{1}_{\|X-X_j\|\geq a}\right] \ .\end{aligned}$$

En utilisant (ii), on en déduit que pour tout $\epsilon > 0$,

$$\limsup_{n\to\infty} \|\eta - \eta_n^\star\|_{L^2} \leq \epsilon \ ,$$

et donc que $\|\eta - \eta_n^\star\|_{L^2}$ converge vers zéro.

Lorsque η n'est pas uniformément continue, on utilise le fait que les fonctions continues à support compact sont denses dans $L^2(\mu)$. Par conséquent, pour tout $\epsilon > 0$, il existe $\widetilde{\eta} \in L^2(\mu)$ telle que $\widetilde{\eta}$ est continue à support compact et $\|\eta - \widetilde{\eta}\|_{L^2(\mu)} \leq \epsilon$. Comme $\widetilde{\eta}$ est uniformément continue, le raisonnement précédent s'applique en remplaçant η par $\widetilde{\eta}$ et $\eta_n^\star$ par

$$\widetilde{\eta}_n^\star(x;D_n) := \sum_{i=1}^n \widetilde{\eta}(X_j) W_j(x;D_n) \ .$$

Puisque

$$\|\eta - \eta_n^\star\|_{L^2} \leq \|\eta - \widetilde{\eta}\|_{L^2} + \|\widetilde{\eta} - \widetilde{\eta}_n^\star\|_{L^2} + \|\widetilde{\eta}_n^\star - \eta_n^\star\|_{L^2} \ ,$$

il reste uniquement à majorer le dernier terme. Or,

$$\begin{aligned}\|\widetilde{\eta}_n^\star - \eta_n^\star\|_{L^2}^2 &= \mathbb{E}\left[\left(\sum_{j=1}^n W_j(X;D_n)\left(\widetilde{\eta}(X_j)-\eta(X_j)\right)\right)^2\right] \\ &\leq \mathbb{E}\left[\sum_{j=1}^n W_j(X;D_n)\left(\widetilde{\eta}(X_j)-\eta(X_j)\right)^2\right] \\ &\leq c\mathbb{E}\left[\left(\widetilde{\eta}(X)-\eta(X)\right)^2\right] \leq c\epsilon^2 \ ,\end{aligned}$$

en utilisant successivement les définitions de $\widetilde{\eta}_n^\star$ et $\eta_n^\star$, l'inégalité de Jensen (comme pour prouver (3.3)), (i) avec $f(x) = \left(\widetilde{\eta}(x) - \eta(x)\right)^2$, et pour finir la définition de $\widetilde{\eta}$.

Pour conclure sur le premier terme du membre de droite de (3.2), nous avons prouvé que pour tout $\epsilon > 0$,

$$\|\eta - \eta_n^\star\|_{L^2} \leq \epsilon + \|\widetilde{\eta} - \widetilde{\eta}_n^\star\|_{L^2} + \sqrt{c}\epsilon \ ,$$

ce majorant tendant vers $\epsilon(1+\sqrt{c})$ quand n tend vers l'infini. Comme ϵ peut être pris arbitrairement proche de zéro, nous avons prouvé que $\|\eta - \eta_n^\star\|_{L^2}$ tend vers 0 quand n tend vers l'infini.

Pour le second terme du membre de droite de (3.2), remarquons que pour tout $i \neq j$,

$$\mathbb{E}\left[\left(Y_i - \eta(X_i)\right)\left(Y_j - \eta(X_j)\right) \mid X, X_1, \ldots, X_n\right] = 0$$

car (X_i, Y_i) est indépendant de (X_j, Y_j). Par conséquent,

$$\begin{aligned}
\|\eta_n^\star - \widehat{\eta}_n\|_{L^2}^2 &= \mathbb{E}\left[\left(\sum_{j=1}^n (\eta(X_j) - Y_j)\, W_j(X; D_n)\right)^2\right] \\
&= \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}\left[(\eta(X_i) - Y_i)\, W_i(X; D_n)\, (\eta(X_j) - Y_j)\, W_j(X; D_n)\right] \\
&= \sum_{j=1}^n \mathbb{E}\left[\left((\eta(X_j) - Y_j)\, W_j(X; D_n)\right)^2\right] \\
&\leq \mathbb{E}\left[\sum_{j=1}^n W_j(X; D_n)^2\right] \leq \mathbb{E}\left[\max_{1\leq j\leq n} W_j(X; D_n)\right]
\end{aligned}$$

et ce majorant tend vers zéro quand n tend vers l'infini d'après (iii). □

3.2.2. Règle par partition.

COROLLAIRE 3.2. *Soit $(A_{1,n}, A_{2,n}, \ldots)_{n\in\mathbb{N}}$ une suite de partitions dénombrables de $\mathcal{X}$. On suppose que*

$$\forall r > 0, \quad \lim_{n\to\infty} \frac{\mathrm{Card}\,\{k \text{ t.q. } A_{k,n} \cap \mathcal{B}(0,r) \neq \emptyset\}}{n} = 0 \quad \textit{et} \quad \lim_{n\to\infty} \max_k \{\mathrm{diam}(A_{k,n})\} = 0 \ ,$$

où $\mathcal{B}(0,r)$ est la boule de centre 0 et de rayon r pour une norme $\|\cdot\|$ sur $\mathcal{X}$ et $\mathrm{diam}(A) = \sup_{x_1,x_2\in A} \|x_1 - x_2\|$. Alors, la règle de classification par partition $\widehat{s}_n$ définie comme dans l'exemple 3.2 avec la partition $A_{1,n}, A_{2,n}, \ldots$ est universellement consistante.

EXEMPLE 3.4. Si $\mathcal{X} = \mathbb{R}^\ell$, une partition classique est la partition régulière de pas $h_n > 0$, où $A_1, A_2, \ldots$ sont des cubes de côté h_n.

Le Corollaire 3.2 montre que la règle par partition associée est universellement consistante lorsque $h_n \to 0$ et $nh_n^\ell \to \infty$ quand n tend vers l'infini.

DÉMONSTRATION. La preuve repose sur le Théorème de Stone (Théorème 3.1). Pour tout $x \in \mathcal{X}$ et $n \in \mathbb{N}$, on note $A_n(x)$ l'élément de la partition $A_{1,n}, A_{2,n}, \ldots$ qui contient x.

(i) Pour tout $k \in \mathbb{N}$, on note $N_{k,n} = \mathrm{Card}\,\{j \text{ t.q. } X_j \in A_{k,n}\}$.

$$\begin{aligned}
\mathbb{E}\left[\sum_{j=1}^n W_j(X; n) f(X_j)\right] &= \mathbb{E}\left[\sum_{j=1}^n \frac{f(X_j)\mathbb{1}_{X_j\in A_n(X)}}{\sum_{k\in\mathbb{N}} \mathbb{1}_{X_k\in A_n(X)}}\right] = \mathbb{E}\left[\sum_{k\in\mathbb{N}} \mathbb{1}_{X\in A_k} \frac{\sum_{j=1}^n f(X_j)\mathbb{1}_{X_j\in A_{k,n}}}{N_{k,n}}\right] \\
&= \sum_{k\in\mathbb{N}} \mathbb{P}\{X \in A_k\}\, \mathbb{E}\left[N_{k,n}^{-1} \mathbb{E}\left[\sum_{j=1}^n f(X_j)\mathbb{1}_{X_j\in A_{k,n}} \,\middle|\, N_{k,n}\right]\right] \\
&= \sum_{k\in\mathbb{N}} \mathbb{P}\{X \in A_k\}\, \mathbb{E}[f(X) \mid X \in A_{k,n}] = \mathbb{E}[f(X)] \ ,
\end{aligned}$$

où l'on a utilisé que, conditionnellement à $N_{k,n}$, $\sum_{j=1}^n f(X_j)\mathbb{1}_{X_j\in A_{k,n}}$ a la loi de la somme de $N_{k,n}$ variables indépendantes et de même loi $\mathcal{L}(X \mid X \in A_{k,n})$.

(ii) $\sum_{j=1}^n W_j(X; n; D_n)\mathbb{1}_{\|X_j - X\|\geq a}$ est nul si $\max_k \mathrm{diam}(A_{k,n}) < a$, ce qui a lieu pour tout k dès que n est assez grand.

(iii) On note μ la distribution de X. Soit $\epsilon > 0$. Il existe alors $r > 0$ tel que $\mu(\mathcal{B}(0,r)^c) < \epsilon$. Notons

$$\mathcal{K}_n(r) = \{\, k \text{ t.q. } A_{k,n} \cap \mathcal{B}(0,r) \neq \emptyset \,\} \quad \text{et} \quad K_n(r) = \operatorname{Card} \mathcal{K}_n(r) \ .$$

Alors,

$$\begin{aligned}\mathbb{E}\left[\max_j W_j(X;n;D_n)\right] &= \mathbb{E}\left[\sum_{k\in\mathcal{K}_n(r)} \mathbb{1}_{X\in A_{k,n}} \frac{\mathbb{1}_{N_{k,n}>0}}{N_{k,n}} + \mathbb{1}_{X\notin\mathcal{B}(0,r)}\right] \\ &= \sum_{k\in\mathcal{K}_n(r)} \mu(A_{k,n}) \mathbb{E}\left[\frac{\mathbb{1}_{N_{k,n}>0}}{N_{k,n}}\right] + \epsilon \ .\end{aligned}$$

Il s'agit donc de majorer l'espérance de l'inverse de $N_{k,n}$, qui suit une loi binômiale de paramètres $(n, \mu(A_{k,n}))$, ce que fait le Lemme 3.1 ci-dessous. On obtient alors

$$\mathbb{E}\left[\max_j W_j(X;n;D_n)\right] \leq \frac{2K_n(r)}{n+1} + \epsilon \xrightarrow[n\to\infty]{} \epsilon \ .$$

Cette majoration étant vraie pour ϵ arbitrairement proche de 0, on en déduit que la condition (iii) est vérifiée.

□

LEMME 3.1. *Soit Z une variable aléatoire binômiale de paramètres (n,p). Alors*

$$\mathbb{E}\left[\frac{\mathbb{1}_{Z>0}}{Z}\right] \leq \frac{2}{(n+1)p} \ .$$

DÉMONSTRATION.

$$\begin{aligned}\mathbb{E}\left[\frac{\mathbb{1}_{Z>0}}{Z}\right] \leq \mathbb{E}\left[\frac{2}{1+Z}\right] &= 2\sum_{k=0}^{n} \frac{1}{k+1} C_n^k p^k (1-p)^{n-k} = \frac{2}{(n+1)p} \sum_{k=0}^{n} C_{n+1}^{k+1} p^{k+1} (1-p)^{n-k} \\ &\leq \frac{2}{(n+1)p} \sum_{k=0}^{n+1} C_{n+1}^{k} p^{k} (1-p)^{n-k+1} = \frac{2}{(n+1)p} \left(p + (1-p)\right)^{n+1} = \frac{2}{(n+1)p} \ .\end{aligned}$$

□

3.2.3. Règle par noyau.

COROLLAIRE 3.3. *Soit $\mathcal{X} \subset \mathbb{R}^\ell$ et $\mathcal{B}(0,r)$ la boule euclidienne de centre 0 et de rayon r. On suppose qu'il existe $0 < r < R$ et $b > 0$ tels que $b\mathbb{1}_{\mathcal{B}(0,r)} \leq K \leq \mathbb{1}_{\mathcal{B}(0,R)}$. Si $h = h_n$ vérifie $\lim_{n\to\infty} h_n = 0$ et $\lim_{n\to\infty} n h_n^\ell = +\infty$, la règle par partition définie par l'exemple 3.3 est universellement consistante.*

DÉMONSTRATION. La démonstration repose sur le Théorème de Stone (Théorème 3.1). Voir [**DGL96**], Chapitre 10 pour la preuve de la consistance forte de règles par noyaux, sous des hypothèses légèrement plus faibles. □

3.3. La règle des k plus proches voisins

3.3.1. Consistance universelle si $k = k_n$. La règle des k plus proches voisins étant une règle par moyennage local, le résultat suivant est une conséquence du Théorème de Stone (Théorème 3.1) :

COROLLAIRE 3.4. *Soit k_n une suite d'entiers telle que*

$$\lim_{n\to\infty} k_n = \infty \quad \textit{et} \quad \lim_{n\to\infty} k_n/n = 0 \ .$$

Alors, la règle des k_n plus proches voisins $\widehat{s}^{k_n-\text{ppv}}$ définie par (1.11) *pour une distance d induite par une norme quelconque sur $\mathcal{X} = \mathbb{R}^\ell$ est universellement consistante.*

DÉMONSTRATION.

(i) Lorsque d est la distance euclidienne sur $\mathbb{R}^\ell$, (i) est satisfaite avec $c = \left(1 + \frac{2}{\sqrt{2-\sqrt{3}}}\right)^d - 1$. Ce résultat est appelé Lemme de Stone [**Sto77**] ; voir [**DGL96**, Lemme 5.3] pour une preuve. Pour une preuve dans le cas d'une norme quelconque, voir [**DGL96**, Problème 5.1].

(ii) Posons $X_{(k)}(x)$ le k-ième plus proche voisin de x dans $X_1, \ldots, X_n$. On a alors,

$$\mathbb{E}\left[\sum_{j=1}^{n} W_j(X) \mathbb{1}_{\|X_j - X\| \geq a}\right] \leq \mathbb{P}\left(\left\|X_{(k)}(X) - X\right\| > a\right) \ ,$$

et cette probabilité tend vers 0 quand n tend vers l'infini dès lors que k_n/n tend également vers zéro [**DGL96**, Lemme 5.1].

(iii) est satisfaite dès lors que $k_n \to \infty$, car $\max_j W_j(x) \leq 1/k_n$ pour tout x.

□

REMARQUE 3.5. Si la distribution μ de X a une densité par rapport à la mesure de Lebesgue sur $\mathcal{X} = \mathbb{R}^\ell$, on peut montrer que la règle $\widehat{s}^{k_n-\text{ppv}}$ est fortement consistante sous les mêmes conditions [**DGL96**, Théorème 11.1]. Pour la consistance forte universelle, les cas d'égalité deviennent importants, et il faut modifier la définition de $\widehat{s}^{k_n-\text{ppv}}$; voir [**DGL96**, Section 11.2].

En pratique, il est fondamental de bien choisir k_n. La vraie distribution des données étant inconnue, on choisit le plus souvent k_n à l'aide des données, notamment par l'une des méthodes de validation croisée. Le résultat suivant donne une condition suffisante simple pour que la règle de classification qui en résulte soit universellement consistante.

COROLLAIRE 3.5. *Soit $\widehat{k}_n$ une suite d'entiers aléatoires telle que*

$$\widehat{k}_n \xrightarrow[n\to\infty]{(p)} \infty \quad \textit{et} \quad \widehat{k}_n/n \xrightarrow[n\to\infty]{(p)} 0 \ .$$

Alors, la règle des $\widehat{k}_n$ plus proches voisins $\widehat{s}^{\widehat{k}_n-\text{ppv}}$ définie par (1.11) *pour une distance d induite par une norme sur $\mathcal{X} = \mathbb{R}^\ell$ est universellement consistante.*

REMARQUE 3.6. Lorsque les deux hypothèses sur $\widehat{k}_n$ sont satisfaites presque sûrement, alors la règle $\widehat{s}^{\widehat{k}_n-\text{ppv}}$ est fortement consistante. Ce résultat, ainsi que le Corollaire 3.5, sont prouvés dans [**DGL96**, Section 26.1].

Nous reviendrons sur les règles automatiques de choix de k au Chapitre 5.

3.3.2. Valeur asymptotique du risque des k-ppv pour k fixé. Afin de mieux comprendre les enjeux du choix de k pour la règle des k plus proches voisins, il est intéressant d'étudier le comportement asymptotique de $\widehat{s}^{k-\text{ppv}}$ pour k fixé. Comme précédemment, on suppose que $\mathcal{X} = \mathbb{R}^\ell$ et que la distance d utilisée est induite par une norme sur $\mathcal{X}$.

PROPOSITION 3.2. *Pour tout entier k impair,*

$$\lim_{n\to\infty} \mathbb{E}\left[\mathcal{R}_P\left(\widehat{s}^{k-\text{ppv}}\right)\right] = \mathcal{R}_P^{k-\text{ppv}} = \mathbb{E}\left[F_k\left(\min\left\{\eta(X), 1-\eta(X)\right\}\right)\right] \ , \tag{3.4}$$

avec $F_k(t) = t + |1-2t|\,\mathbb{P}\left(\mathit{Binom}(k,t) > k/2\right)$, où $\mathit{Binom}(k,t)$ désigne une variable aléatoire binômiale de paramètres (k,t).

Ce résultat est prouvé dans [**DGL96**] pour $k = 1$ (Théorème 5.1) puis pour $k \geq 3$ quelconque (Théorème 5.3).

REMARQUE 3.7. La valeur limite du risque de $\widehat{s}^{k-\text{ppv}}$ ne dépend de P qu'à travers de la distribution de $\min\{\eta(X), 1-\eta(X)\}$. Elle est en particulier indépendante de la régularité de η (vis-à-vis de la distance d), ou à l'inverse de son manque de régularité. Seule la *vitesse de convergence* de $\mathbb{E}\left[\mathcal{R}_P\left(\widehat{s}^{k-\text{ppv}}\right)\right]$ vers $\mathcal{R}_P^{k-\text{ppv}}$ dépend de ces facteurs, qui jouent bien évidemment un rôle important lorsque n est fixé.

Rappelons que

$$\mathcal{R}_P^\star = \mathbb{E}\left[\min\{\eta(X), 1-\eta(X)\}\right] \ .$$

Par conséquent, pour tout k impair, $\mathcal{R}_P^{k-\text{ppv}} \geq \mathcal{R}_P^\star$ avec égalité si et seulement si $\eta(X) \in \{0, 1/2, 1\}$ presque sûrement. En particulier, $\widehat{s}^{k-\text{ppv}}$ avec k fixe (ou $k = k_n$ borné) est consistante que si et seulement si $\eta(X) \in \{0, 1/2, 1\}$ p.s.

Lorsque $k = 1$, (3.4) permet de donner une expression explicite de la valeur limite du risque :

$$\mathcal{R}_P^{k-\text{ppv}} = \mathbb{E}\left[2\eta(X)(1-\eta(X))\right] \leq 2\mathcal{R}_P^\star\left(1-\mathcal{R}_P^\star\right) \leq 2\mathcal{R}_P^\star \ .$$

Enfin, pour tout k impair, (3.4) implique la majoration suivante de $\mathcal{R}_P^{k-\text{ppv}}$:

$$\mathcal{R}_P^{k-\text{ppv}} \leq \mathcal{R}_P^\star + \sqrt{\frac{2\mathcal{R}_P^\star}{k}} \leq \mathcal{R}_P^\star + \sqrt{\frac{1}{k}} \ . \tag{3.5}$$

D'autres majorations de $\mathcal{R}_P^{k-\text{ppv}}$ sont prouvées dans [**DGL96**, Section 5.7].

DÉMONSTRATION DE (3.5). Soit Z une variable binômiale de paramètres (k, t) avec $0 \leq t \leq 1/2$. Alors,

$$\begin{aligned}
\mathbb{P}\left(Z > \frac{k}{2}\right) = \mathbb{P}\left(Z - kt > k\left(\frac{1}{2} - t\right)\right) &\leq \frac{\mathbb{E}\left[Z - kt\right]}{k\left(\frac{1}{2}-t\right)} \qquad \text{(Inégalité de Markov)} \\
&\leq \frac{\sqrt{\text{var}\left(Z\right)}}{k\left(\frac{1}{2}-t\right)} = \frac{2\sqrt{t(1-t)}}{\sqrt{k}(1-2t)} \qquad \text{(Inégalité de Jensen)} \ .
\end{aligned}$$

Par conséquent, (3.4) implique, en utilisant à nouveau l'inégalité de Jensen,

$$\begin{aligned}
\mathcal{R}_P^{k-\text{ppv}} - \mathcal{R}_P^\star &\leq 2k^{-1/2}\mathbb{E}\left[\sqrt{\eta(X)(1-2\eta(X))}\right] \\
&\leq 2k^{-1/2}\sqrt{\mathbb{E}\left[\eta(X)(1-2\eta(X))\right]} = \sqrt{\frac{2\mathcal{R}_P^\star}{k}} \leq \sqrt{\frac{1}{k}} \ .
\end{aligned}$$

□

3.3.3. Élements de comparaison des différentes valeurs de k. Une autre conséquence intéressante de (3.4) est que l'on peut comparer les valeurs de $\mathcal{R}_P^{k-\text{ppv}}$ pour différentes valeurs de k :

$$\mathcal{R}_P^\star \leq \cdots \leq \mathcal{R}_P^{(2k)-\text{ppv}} = \mathcal{R}_P^{(2k-1)-\text{ppv}} \leq \cdots \leq \mathcal{R}_P^{2-\text{ppv}} = \mathcal{R}_P^{1-\text{ppv}} \leq 2\mathcal{R}_P^\star \ . \tag{3.6}$$

La comparaison pour un nombre de voisins impair est prouvée par [**DGL96**, Théorème 5.5]. L'égalité $\mathcal{R}_P^{(2k)-\text{ppv}} = \mathcal{R}_P^{(2k-1)-\text{ppv}}$ est quant à elle vérifiée en attribuant à x l'étiquette de son plus proche voisin dans $X_1, \ldots, X_n$ lorsqu'il y a égalité [**DGL96**, Théorème 5.6].

L'inégalité (3.6) illustre le fait que les petites valeurs de k ne sont pas en général de bons choix pour la règle des k-ppv lorsque le nombre d'observations n tend vers l'infini.

Cependant, cet ordre de comparaison n'est pas vrai pour tout P. Considérons par exemple la distribution suivante sur $\mathcal{X} \times \{0, 1\}$ avec $\mathcal{X} = \mathbb{R}^\ell$. Soient $a, b \in \mathcal{X}$ tels que $\|a - b\| > 4$, S_0

la sphère de centre a et de rayon 1 et S_1 la sphère de centre b et de rayon 1. Soit Y tel que $\mathbb{P}(Y=0)=\mathbb{P}(Y=1)=1/2$ et X de loi uniforme sur S_Y.

Comme S_0 et S_1 sont disjointes, Y est une fonction déterministe de X. Pour tout k impair, la règle des k-ppv se trompe sur S_j si et seulement si le nombre de $X_i \in S_j$ est inférieur ou égal à $(k-1)/2$, soit

$$\begin{aligned}\mathbb{E}\left[\mathcal{R}_P\left(\widehat{s}^{1-\text{ppv}}\right)\right] &= \mathbb{P}\left(Y=0, \sum_{i=1}^{n} \mathbb{1}_{Y_i=1} \leq (k-1)/2\right) + \mathbb{P}\left(Y=1, \sum_{i=1}^{n} \mathbb{1}_{Y_i=0} \leq (k-1)/2\right) \\ &= \mathbb{P}\left(\text{Binom}(n,1/2) \leq (k-1)/2\right) \ ,\end{aligned}$$

qui est une fonction strictement croissante de k, valant 2^{-n} pour $k=1$ et au moins $(1+n)2^{-n}$ pour $k \geq 3$.

CHAPITRE 4

Minimisation du risque empirique

Ce chapitre a pour objet l'étude d'une approche alternative à celle du moyennage local : la minimisation du risque empirique.

4.1. Principe, définition, exemples

L'objectif de la classification supervisée est de trouver un classifieur $t \in \mathbb{S}$ dont le risque $\mathcal{R}_P(t)$ est aussi proche que possible du risque de Bayes $\mathcal{R}_P^\star = \min_{t\in\mathbb{S}} \mathcal{R}_P(t)$. Une approche naturelle est de construire pour chaque classifieur $t \in \mathbb{S}$ un estimateur $\widehat{\mathcal{R}}(t)$ de son risque, puis de choisir t qui minimise cette estimation.

Par exemple, si l'on remplace dans la définition du risque $\mathcal{R}_P(t)$ la distribution P des données par la *mesure empirique*

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

alors on obtient le *risque empirique*

$$\widehat{\mathcal{R}}_n(t; D_n) = \widehat{\mathcal{R}}_n(t) := \mathcal{R}_{P_n}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{t(X_i)\neq Y_i} \ . \tag{4.1}$$

Comme chacun des (X_i, Y_i) a pour distribution P, pour tout classifieur t déterministe, le risque empirique de t est un estimateur sans biais du risque de t :

$$\mathbb{E}_{D_n\sim P^{\otimes n}}\left[\widehat{\mathcal{R}}_n(t; D_n)\right] = \mathbb{P}_{(X,Y)\sim P}(t(X)\neq Y) = \mathcal{R}_P(t) \ .$$

Plus encore, comme nous l'avons vu en Section 2.1 dans la démonstration du Théorème 2.1, d'après l'inégalité de Hoeffding,

$$\forall t \in \mathbb{S},\ \forall \epsilon > 0, \qquad \mathbb{P}\left(\left|\widehat{\mathcal{R}}_n(t) - \mathcal{R}_P(t)\right| \geq \epsilon\right) \leq 2\exp\left(-2n\epsilon^2\right) \ .$$

Le risque empirique est donc un bon estimateur du risque d'un classifieur.

4.1.1. Définition. Minimiser le risque empirique parmi les classifieurs $t \in S \subset \mathbb{S}$ définit alors une règle de classification, appelée *minimiseur du risque empirique sur* S :

$$\widehat{s}_S(D_n) \in \arg\min_{t\in S}\left\{\widehat{\mathcal{R}}_n(t; D_n)\right\} \ . \tag{4.2}$$

Remarque 4.1 (Existence et unicité du minimiseur du risque empirique). En général, (4.2) ne définit pas toujours bien $\widehat{s}_S(D_n)$:

- le risque empirique n'atteint pas nécessairement son infimum sur S en un unique classifieur. Dans ce cas, on doit choisir une règle pour définir $\widehat{s}_S$ de manière non-ambiguë, l'analyse statistique se faisant pour n'importe quel $\widehat{s}_S \in \mathbb{S}$ satisfaisant (4.2).
- le risque empirique n'atteint pas nécessairement son infimum sur S, auquel cas on se contente d'un minimiseur approché, c'est-à-dire tel que

$$\widehat{\mathcal{R}}_n(\widehat{s}_S(D_n)) \leq \inf_{t\in S_m}\left\{\widehat{\mathcal{R}}_n(t; D_n)\right\} + \rho_n \ , \tag{4.3}$$

avec $\rho_n = 1/n$ par exemple. Le problème de l'unicité se pose alors également. Noter que la complexité algorithmique de la minimisation exacte du risque empirique (4.2) peut être telle que l'on se contente en pratique d'une minimisation approchée (4.3), la valeur de ρ_n dépendant notamment de la puissance de calcul disponible.

L'ensemble $S \subset \mathbb{S}$ de classifieurs considéré est appelé *modèle*. Notons que dans le cas de la classification binaire, on peut toujours écrire

$$S = \{\mathbb{1}_A \text{ t.q. } A \in \mathcal{A}\}$$

où $\mathcal{A}$ est une famille de parties de $\mathcal{X}$. Par extension, $\mathcal{A}$ est également appelé «modèle».

4.1.2. Exemples.

REMARQUE 4.2. La règle $\widehat{s}_{\mathbb{S}}$ est très mauvaise dès que $\mathrm{Card}(\mathcal{X})$ est de l'ordre de la taille n de l'échantillon ou plus grand. En effet, si tous les X_i sont différents, minimiser le risque empirique sur $\mathbb{S}$ revient à attribuer l'étiquette Y_i lorsque $x = X_i$, et n'importe quelle étiquette aux autres $x \in \mathcal{X}$. En particulier, si la loi de X est sans atome, l'un des classifieurs réalisant le minimum dans (4.2) a un risque égal à $1 - \mathcal{R}^\star > \mathcal{R}^\star$.

On peut noter également que la règle du plus proche voisin est l'un des minimiseurs du risque empirique sur $\mathbb{S}$ et n'est pas consistante en général.

EXEMPLE 4.3 (Règle par partition). Pour toute partition $A_1, \ldots, A_K$ de $\mathcal{X}$, on définit le modèle suivant :

$$S_{\text{partition}}(A_1, \ldots, A_K) := \left\{ \sum_{j=1}^{K} a_j \mathbb{1}_{x \in A_j} \text{ t.q. } a_1, \ldots, a_K \in \{0, 1\} \right\} .$$

En particulier, si $\mathcal{X} = [0,1]^\ell$, un modèle particulier d'histogrammes est le modèle associé à la partition de $\mathcal{X}$ en une grille régulière de pas $h \in (0,1)$ avec $h^{-1} \in \mathbb{N}$; son cardinal est $K = h^{-d}$.

Lorsque X n'est pas bornée dans $\mathcal{X} = \mathbb{R}^\ell$, on peut soit considérer une partition dénombrable de $\mathcal{X}$ (en choisissant une étiquette par défaut pour tous les A_j ne contenant aucun des X_i), soit fixer un rayon $R > 0$ et considérer la partition $A_1, \ldots, A_K$ où $A_1, \ldots, A_{K-1}$ partitionne $[-R; R]^\ell$ à l'aide d'une grille régulière de pas h et $A_K = \mathbb{R}^\ell \backslash [-R; R]^\ell$.

Dans tous ces cas, minimiser le risque empirique sur $S_{\text{partition}}(A_1, \ldots, A_K)$ est aisé car cela revient à prendre pour a_j l'étiquette majoritaire parmi les données (X_i, Y_i) telles que $X_i \in A_j$.

EXEMPLE 4.4 (Classification par intervalles). Si $\mathcal{X} = [0,1]$, pour tout $K \geq 1$, on peut considérer le modèle de classification par K intervalles

$$S_{\text{inter}}(K) := \left\{ \sum_{j=1}^{K} a_j \mathbb{1}_{[\alpha_{j-1}, \alpha_j[} \text{ t.q. } 0 = \alpha_0 \leq \alpha_1 \leq \cdots \leq \alpha_K = 1 \text{ et } a_1, \ldots, a_K \in \{0, 1\} \right\} .$$

Il est possible de minimiser le risque empirique sur $S_{\text{inter}}(K)$ à l'aide d'un algorithme de programmation dynamique [**KMNR97**] dont la complexité algorithmique est $\mathcal{O}(n \ln(n))$.

EXEMPLE 4.5 (Rectangles). Si $\mathcal{X} = \mathbb{R}^\ell$ avec $\ell \geq 1$ un entier, alors, le modèle des *rectangles* dans $\mathcal{X}$ est défini par

$$S_{\text{rect}} := \{ x \mapsto \mathbb{1}_{a_i \leq x \leq b_i} \text{ t.q. } a_1 \leq b_1, \ldots, a_\ell \leq b_\ell \} .$$

EXEMPLE 4.6 (Séparation linéaire). Si $\mathcal{X} = \mathbb{R}^\ell$ avec $\ell \geq 1$ un entier, alors, le modèle des *séparateurs linéaires* dans $\mathcal{X}$ est défini par

$$S_{\text{lin}} := \left\{ x \mapsto \mathbb{1}_{a_0 + a^T x > 0} \text{ t.q. } a_0 \in \mathbb{R},\, a \in \mathbb{R}^\ell \right\} \ .$$

Un classifieur $t \in S_{\text{lin}}$ est souvent appelé *perceptron*.

Plus généralement, pour tout $I \subset \{1, \ldots, \ell\}$, on peut définir le modèle des séparateurs linéaires fondés sur les coordonnées de x dont les indices appartiennent à I :

$$S_{\text{lin}}(I) := \left\{ x \mapsto \mathbb{1}_{a_0 + a^T x > 0} \text{ t.q. } a_0 \in \mathbb{R},\, a \in \mathbb{R}^\ell,\, \forall j \notin I,\, a_j = 0 \right\} \ .$$

C'est en particulier nécessaire lorsque $\ell \geq n$, car n points en position générale dans $\mathbb{R}^n$ sont toujours séparables par un hyperplan, indépendamment des valeurs des étiquettes Y_i.

4.1.3. Risque empirique convexifié. Il est difficile de minimiser le risque empirique sur un modèle S général car $\widehat{\mathcal{R}}_n(t)$ n'est pas une fonction convexe de t. C'est pourquoi l'on utilise souvent en classification une autre approche, qui consiste à minimiser une version "convexifiée" du risque.

Pour cela, au lieu de chercher un classifieur $t \in \mathbb{S}$ (application $\mathcal{X} \mapsto \{0, 1\}$), on cherche une application mesurable $f : \mathcal{X} \mapsto \mathbb{R}$ et on lui associe le classifieur $t_f : x \mapsto \mathbb{1}_{f(x) \geq 0}$. La valeur absolue de $f(x)$ quantifie alors le niveau de confiance que l'on a en la classification $t_f(x)$. On note $\mathbb{S}_{\text{conv}}$ l'ensemble des applications mesurables $\mathcal{X} \mapsto \mathbb{R}$. Le risque de $f \in \mathbb{S}_{\text{conv}}$ est alors mesuré par le ϕ-risque

$$\mathcal{R}_{P,\phi}(f) := \mathbb{E}_{X,Y) \sim P}\left[\phi\left(f(X)(1 - 2Y)\right)\right] \ ,$$

pour une fonction décroissante $\phi : \mathbb{R} \mapsto [0; +\infty[$.

On retrouve le risque $\mathcal{R}_P(t_f)$ défini en Section 1.2 et , appelé *risque 0-1*, en prenant $\phi(u) = \phi_{0-1}(u) = \mathbb{1}_{u \leq 0}$. On parle de *risque convexifié* lorsque ϕ est un majorant convexe de ϕ_{0-1} tel que $\lim_{u \to +\infty} \phi(u) = 0$. Des exemples classiques de fonctions ϕ sont :

- la perte charnière (hinge loss) : $\phi_{\text{charniere}}(u) = (1 - u)_+ = \max\{0, 1 - u\}$ utilisée pour les machines à vecteurs de support (SVM).
- la perte logit : $\phi_{\text{logit}}(u) = \ln_2(1 + \exp(u))$ utilisée par le boosting.
- la perte exponentielle : $\phi_{\text{exp}}(u) = \exp(-u)$ utilisée par AdaBoost.
- la perte quadratique : $\phi_{\text{quad}}(u) = (1 - u)^2$.

Lorsque l'objectif est de minimiser le ϕ-risque $\mathcal{R}_{P,\phi}(f)$, il est naturel d'utiliser la minimisation du ϕ-risque empirique

$$\widehat{\mathcal{R}}_{n,\phi}(f; D_n) := \mathcal{R}_{P_n,\phi}(f) = \frac{1}{n} \sum_{i=1}^{n} \phi\left(f(X_i)(1 - 2Y_i)\right)$$

sur un modèle $S \subset \mathbb{S}_{\text{conv}}$, soit

$$\widehat{f}_{S,\phi}(D_n) \in \arg\min_{t \in S} \left\{ \widehat{\mathcal{R}}_{n,\phi}(f; D_n) \right\} \ ,$$

et l'on note $\widehat{s}_{S,\phi}(D_n) = t_{\widehat{f}_{S,\phi}}$ le classifieur associé.

Mais, en général, le risque 0-1 reste le critère le plus naturel à minimiser en classification binaire supervisée. Une question naturelle est de savoir si le classifieur $\widehat{s}_{S,\phi}(D_n)$, qui est plus facile à calculer que $\widehat{s}_S(D_n)$, a un risque 0-1 comparable à celui de $\widehat{s}_S(D_n)$. On peut montrer pour les fonctions ϕ données en exemple ci-dessus que l'excès de risque de $\widehat{s}_{S,\phi}(D_n)$ est petit dès lors que l'excès de ϕ-risque de $\widehat{s}_{S,\phi}(D_n)$ est petit.

4.2. Risque de classification

4.2.1. Décomposition biais-variance du risque. Rappelons la décomposition du risque entre l'erreur d'approximation et erreur d'estimation, introduite en Section 1.7.1.

$$\mathcal{R}_P\left(\widehat{s}_S(D_n)\right) - \mathcal{R}_P^\star = \left[\mathcal{R}_P\left(\widehat{s}_S(D_n)\right) - \inf_{t\in S}\left\{\mathcal{R}_P\left(t\right)\right\}\right] + \left[\inf_{t\in S}\left\{\mathcal{R}_P\left(t\right)\right\} - \mathcal{R}_P^\star\right] \quad . \tag{4.4}$$

Le premier terme de (4.4),

$$\mathcal{R}_P\left(\widehat{s}_S(D_n)\right) - \inf_{t\in S}\left\{\mathcal{R}_P\left(t\right)\right\} \quad , \tag{4.5}$$

est appelé *erreur d'estimation*. Il est en général d'autant plus grand que S est grand.

Le second terme de (4.4),

$$\inf_{t\in S}\left\{\mathcal{R}_P\left(t\right)\right\} - \mathcal{R}_P^\star =: \ell\left(s^\star, S\right) \quad , \tag{4.6}$$

est appelé *erreur d'approximation*. Il est d'autant plus petit que S est grand.

REMARQUE 4.7. Lorsque l'inf du risque sur S est atteint, on note $s_S^\star \in S$ l'un des minimiseurs du risque. Pour simplifier, on supposera dans la suite qu'un tel $s_S^\star$ existe (sans être nécessairement unique), tous les résultats énoncés restant vrais lorsque l'infimum du risque n'est pas atteint dans S.

Dans le reste de ce chapitre, nous supposerons un modèle S fixé et chercherons à déterminer l'ordre de grandeur du risque de $\widehat{s}_S$. Au vu de la décomposition (4.4) du risque, il suffit d'étudier séparément l'erreur d'approximation et l'erreur d'estimation, qui sont de natures différentes.

Un second problème, que nous aborderons au Chapitre 5, est celui de choisir un bon modèle S. La décomposition (4.4) du risque montre qu'il s'agit de réaliser un *compromis entre estimation et approximation*[1]. En effet, l'erreur d'approximation (4.6) est par définition une fonction décroissante de S. À l'inverse, l'erreur d'estimation (4.5) tend à être plus grande lorsque S grandit, car augmenter le nombre de paramètres à estimer augmente l'incertitude globale sur l'estimation de $s_S^\star$.

4.2.2. Erreur d'approximation. D'après la formule de l'excès de risque d'un classifieur (1.4), l'erreur d'approximation d'un modèle S s'écrit

$$\ell\left(s^\star, S\right) = \mathbb{E}\left[\left|2\eta(X) - 1\right| \mathbb{1}_{s_S^\star(X) \neq s^\star(X)}\right] = \mathbb{E}\left[\left|2\eta(X) - 1\right| \left|s_S^\star(X) - s^\star(X)\right|\right] \quad .$$

Dans le cas zéro-erreur, l'erreur d'approximation est donc égale à la distance $L^1(\mu)$ entre $s^\star$ et S, où μ désigne la loi de X. Dans le cas général, cette distance L^1 est pondérée par $|2\eta(X) - 1|$.

Pour obtenir une règle de classification universellement consistante, il est donc important de choisir une de modèles S_n asymptotiquement dense dans $L^1(\mu)$ pour tout mesure de probabilité μ sur $\mathcal{X}$.

C'est par exemple le cas des modèles par partition régulière lorsque le pas h_n de la partition tend vers zéro (et le rayon R_n tend vers l'infini, lorsque X n'est pas bornée *a priori*). En revanche, ce n'est pas le cas des modèles par séparation linéaire, qui ne sont utiles que lorsque l'on a une information *a priori* sur la distribution P des données.

De manière générale, majorer l'erreur d'approximation pour un modèle donné relève de la *théorie de l'approximation*. À l'inverse, majorer l'erreur d'estimation relève de la statistique.

[1] On parle souvent également de "compromis biais-variance", le biais d'un modèle S étant l'erreur d'approximation, et la variance se référant à l'erreur d'estimation (qui provient de l'incertitude que l'on a sur chacun des paramètres du modèles).

4.2.3. Erreur d'estimation. Nous avons vu dans les chapitres précédents (Sections 1.7.2 et 2.1) une majoration générale de l'erreur d'estimation (Lemme 2.1). Rappelons-la ici, en incluant le cas d'un minimiseur approché du risque empirique : pour tout modèle $S \subset \mathbb{S}$ et tout classifieur $\widehat{s}_S(D_n)$ tel que

$$\widehat{\mathcal{R}}_n\left(\widehat{s}_S(D_n); D_n\right) \leq \inf_{t\in S}\left\{\widehat{\mathcal{R}}_n\left(t; D_n\right)\right\} + \rho_n \ ,$$
$$\mathcal{R}_P\left(\widehat{s}_S(D_n)\right) - \inf_{t\in S}\mathcal{R}_P\left(t\right) \leq 2\sup_{t\in S}\left|\widehat{\mathcal{R}}_n\left(t\right) - \mathcal{R}_P\left(t\right)\right| + \rho_n \ . \tag{4.7}$$

Démonstration. Soit $\epsilon > 0$ et $t_\epsilon \in S$ tel que

$$\mathcal{R}_P\left(t_\epsilon\right) \leq \inf_{t\in S}\mathcal{R}_P\left(t\right) + \epsilon \ .$$

Par définition de t_ϵ et de $\widehat{s}$,

$$\begin{aligned}\mathcal{R}_P\left(\widehat{s}\right) - \inf_{t\in S}\mathcal{R}_P\left(t\right) &= \mathcal{R}_P\left(\widehat{s}\right) - \widehat{\mathcal{R}}_n\left(\widehat{s}\right) + \widehat{\mathcal{R}}_n\left(\widehat{s}\right) - \inf_{t\in S}\mathcal{R}_P\left(t\right)\\ &\leq \mathcal{R}_P\left(\widehat{s}\right) - \widehat{\mathcal{R}}_n\left(\widehat{s}\right) + \widehat{\mathcal{R}}_n\left(t_\epsilon\right) - \mathcal{R}_P\left(t_\epsilon\right) + \epsilon\\ &\leq 2\sup_{t\in S}\left|\widehat{\mathcal{R}}_n\left(t\right) - \mathcal{R}_P\left(t\right)\right| + \rho_n + \epsilon \ .\end{aligned}$$

Comme ϵ peut-être pris arbitrairement proche de zéro, on en déduit le résultat. □

Il suffit donc d'obtenir une majoration *uniforme* de l'écart entre le risque et le risque empirique sur S pour obtenir une majoration de l'erreur d'estimation. Nous verrons au Chapitre 5 que cette quantité apparaît également dans le cadre de la sélection de modèles.

Il est à noter que

$$\mathcal{B}_P(S; D_n) := \sup_{t\in S}\left|\widehat{\mathcal{R}}_n\left(t; D_n\right) - \mathcal{R}_P\left(t\right)\right|$$

est une fonction croissante de S. Bien que la majoration (4.7) ne soit pas toujours précise, $\mathcal{B}_P(S; P_n)$ fournit la plupart du temps un bon indicateur de la complexité du modèle S.

Considérons dans un premier temps le cas d'un modèle S fini.

Théorème 4.1. *Soit $S \subset \mathbb{S}$ de cardinal fini. Alors,*

$$\mathbb{P}\left(\mathcal{B}_P(S; D_n) \geq \epsilon\right) \leq 2\,\mathrm{Card}(S)\exp\left(-2n\epsilon^2\right) \tag{4.8}$$

$$\mathbb{E}\left[\mathcal{B}_P(S; D_n)\right] \leq \frac{\sqrt{\ln(\mathrm{Card}(S))} + \sqrt{\pi}}{\sqrt{2n}} \ . \tag{4.9}$$

Démonstration. Pour tout $t \in \mathbb{S}$, l'inégalité de Hoeffding (2.4) appliquée aux variables $\xi_i = n^{-1}\mathbb{1}_{t(X_i)\neq Y_i} \in [0; n^{-1}]$ qui sont indépendantes et d'espérance $n^{-1}\mathcal{R}_P\left(t\right)$, montre que

$$\mathbb{P}\left(\left|\widehat{\mathcal{R}}_n\left(t\right) - \mathcal{R}_P\left(t\right)\right| \geq \epsilon\right) \leq 2\exp\left(-2n\epsilon^2\right) \ .$$

La borne de l'union permet alors d'obtenir (4.8) :

$$\mathbb{P}\left(\sup_{t\in S}\left|\widehat{\mathcal{R}}_n\left(t\right) - \mathcal{R}_P\left(t\right)\right| \geq \epsilon\right) \leq \sum_{t\in S}\mathbb{P}\left(\left|\widehat{\mathcal{R}}_n\left(t\right) - \mathcal{R}_P\left(t\right)\right| \geq \epsilon\right) \leq 2\,\mathrm{Card}(S)\exp\left(-2n\epsilon^2\right) \ .$$

Remarquons que pour tout $u \geq 0$,

$$\begin{aligned}
\mathbb{E}\left[\mathcal{B}_P(S; D_n)\right] &= \int_0^\infty \mathbb{P}\left(\mathcal{B}_P(S; D_n) \geq \epsilon\right) \mathrm{d}\epsilon \\
&\leq u + \int_u^\infty \mathbb{P}\left(\mathcal{B}_P(S; D_n) \geq \epsilon\right) \mathrm{d}\epsilon \\
&\leq u + 2\,\mathrm{Card}(S) \int_u^\infty \exp\left(-2n\epsilon^2\right) \mathrm{d}\epsilon \\
&\leq u + 2\,\mathrm{Card}(S) e^{-2nu^2} \int_0^\infty \exp\left(-2n\epsilon^2\right) \mathrm{d}\epsilon \\
&= u + e^{-2nu^2} \frac{\mathrm{Card}(S)\sqrt{\pi}}{\sqrt{2n}} \quad .
\end{aligned}$$

L'inégalité (4.9) s'en déduit en prenant $u = \sqrt{\ln(\mathrm{Card}(S))/(2n)}$. □

Lorsque le modèle S est infini, la même approche ne peut pas fonctionner car le cardinal de S apparaît dans les bornes sur l'erreur d'estimation. En revanche, on peut remplacer $\mathrm{Card}(S)$ par des mesures moins restrictives de la "capacité" du modèle S.

4.3. Classes de Vapnik-Chervonenkis

Notons que dans le cas de la classification binaire, l'ensemble $\mathbb{S}$ des classifieurs est en bijection avec l'ensemble $\mathfrak{P}(\mathcal{X})$ des parties de $\mathcal{X}$ via l'application $t \mapsto A_t = \{\, x \in \mathcal{X} \text{ t.q. } t(x) = 1 \,\}$. Un modèle $S \subset \mathbb{S}$ peut donc également être vu comme une partie $\mathcal{A}(S) = \{\, A_t \text{ t.q. } t \in S \,\}$ de $\mathfrak{P}(\mathcal{X})$. La notion de classe de Vapnik-Chervonenkis pour une partie $\mathcal{A}$ de $\mathfrak{P}(\mathcal{X})$ peut donc être utilisée pour mesurer la "capacité" d'un modèle S.

4.3.1. Définition.

Définition 4.1 (Classe de Vapnik-Chervonenkis). Si $\mathcal{A}$ est une famille de parties mesurables de $\mathcal{X}$, on note

$$\begin{aligned}
\forall x_1, \ldots, x_n \in \mathcal{X}, \quad N_{\mathcal{A}}(x_1, \ldots, x_n) &:= \mathrm{Card}\left\{\, A \cap \{\, x_1, \ldots, x_n \,\} \text{ t.q. } A \in \mathcal{A} \,\right\} \\
\forall n \in \mathbb{N}, \quad N_{\mathcal{A}}(n) &:= \sup_{x_1, \ldots, x_n \in \mathcal{X}} \left\{\, N_{\mathcal{A}}(x_1, \ldots, x_n) \,\right\} \quad .
\end{aligned}$$

On dit que $\mathcal{A}$ est une *classe de Vapnik-Chervonenkis* ($\mathcal{A}$ est VC) lorsque

$$V_{\mathcal{A}} = \sup\left\{\, n \text{ t.q. } N_{\mathcal{A}}(n) = 2^n \,\right\} < \infty$$

et $V_{\mathcal{A}} = \dim_{\mathrm{VC}} \mathcal{A}$ est appelée la *dimension de Vapnik-Chervonenkis* de $\mathcal{A}$. De la même façon, on dit sous ces conditions que le modèle $S_{\mathcal{A}} = \{\, \mathbb{1}_A \text{ t.q. } A \in \mathcal{A} \,\}$ associé à $\mathcal{A}$ est une *classe de Vapnik-Chervonenkis de dimension* $V_{\mathcal{A}} = V(S_{\mathcal{A}})$.

La dimension de Vapnik-Chervonenkis de $\mathcal{A}$ décrit la capacité de $S_{\mathcal{A}}$ à expliquer des étiquettes quelconques $Y_1, \ldots, Y_n$ qui seraient associées à des points $X_1, \ldots, X_n$. En particulier, si $V(S) \geq n$, il existe une configuration de $X_1, \ldots, X_n$ telle que le risque empirique de $\widehat{s}_S$ est nul, quelles que soient les étiquettes $Y_1, \ldots, Y_n$. Pour que $\widehat{s}_S$ soit capable de généralisation, c'est-à-dire de prédire de nouvelles données, il est donc préférable d'éviter d'utiliser un modèle S telque $\dim_{\mathrm{VC}} S \geq n$.

4.3.2. Exemples. Pour illustrer ce qu'est une classe de Vapnik-Chervonenkis, commençons par deux exemples jouet.

Exemple 4.8. L'ensemble des parties à moins de V éléments de $\mathcal{X}$ est VC de dimension V.

De manière générale, tout ensemble $\mathcal{A} \subset \mathfrak{P}(\mathcal{X})$ fini, est VC de dimension $\dim_{\mathrm{VC}} \mathcal{A} \leq \ln_2(\mathrm{Card}(\mathcal{A}))$.

EXEMPLE 4.9. L'ensemble $\mathcal{A}_c$ des parties convexes de $\mathbb{R}^2$ n'est pas une classe de Vapnik-Chervonenkis car $\forall k$, $N_{\mathcal{A}_c}(k) = 2^k$.

Les exemples suivants montrent que nombre de modèles couramment utilisés sont des classes de Vapnik-Chervonenkis.

EXEMPLE 4.10. Pour toute partition finie $A_1, \ldots, A_K$ de $\mathcal{X}$, le modèle associé $S_{\text{partition}}(A_1, \ldots, A_K)$ défini par l'Exemple 4.3 est VC de dimension K.

EXEMPLE 4.11. Si $\mathcal{A} = \{\,]-\infty; a]$ t.q. $a \in \mathbb{R}\,\}$ et $\mathcal{X} = \mathbb{R}$, alors $N_{\mathcal{A}}(2) = 3 < 2^2$ donc $\mathcal{A}$ est VC de dimension 1.

EXEMPLE 4.12. Pour tout $K \geq 1$, le modèle $S_{\text{inter}}(K)$ de classification par K intervalles défini par l'Exemple 4.4 est VC de dimension K.

REMARQUE 4.13. Les modèles $S_{\text{inter}}(K)$ et $S_{\text{partition}}(([\frac{j-1}{K}; \frac{j}{K}[)_{1\leq j\leq K})$ ont la même dimension de Vapnik-Chervonenkis, alors que le premier est *beaucoup* plus riche que le second. La différence est que $N_{\mathcal{A}_{\text{inter}}(K)}(X_1, \ldots, X_K) = 2^K$ dès que les X_i sont différents, alors que $N_{\mathcal{A}_{\text{partition}}(\cdots)}(X_1, \ldots, X_K) = 2^K$ ne se produit que lorsque les X_i sont dans des éléments de la partition tous différents. La dimension de Vapnik-Chervonenkis est donc beaucoup trop pessimiste pour le cas des partitions fixes.

EXEMPLE 4.14. Pour tout $\ell \geq 1$, le modèle S_{rect} des *rectangles* dans $\mathcal{X} = \mathbb{R}^\ell$ défini par l'Exemple 4.5 est VC de dimension 2ℓ.

EXEMPLE 4.15. L'ensemble des demi-espaces affines de $\mathbb{R}^\ell$ est VC de dimension $\ell + 1$.

Noter que cette partie de $\mathfrak{P}(\mathbb{R}^\ell)$ est associée au modèle S_{lin} des *séparateurs linéaires* (Exemple 4.6). De la même manière, pour tout $I \subset \{1, \ldots, \ell\}$, $S_{\text{lin}}(I)$ défini dans l'Exemple 4.6 est une classe de Vapnik-Chervonenkis de dimension $\text{Card}(I) + 1$.

Plus généralement, on peut montrer le théorème suivant.

THÉORÈME 4.2. *Si $\mathcal{H}$ est un espace vectoriel de fonctions $\mathcal{X} \mapsto \mathbb{R}$ et*

$$S_{\mathcal{H}} = \left\{ x \mapsto \mathbb{1}_{h(x)\geq 0} \text{ t.q. } h \in \mathcal{H} \right\} \ ,$$

alors $S_{\mathcal{H}}$ est une classe de Vapnik-Chervonenkis de dimension $\dim \mathcal{H}$.

Le modèle $S_{\text{lin}}(I)$ considéré par l'Exemple 4.15 correspond à l'espace vectoriel

$$\mathcal{H} := \left\{ x \mapsto a_0 + \sum_{i\in I} a_i x_i \text{ t.q. } a_0 \in \mathbb{R}, (a_i)_{i\in I} \in \mathbb{R}^{\text{Card}\, I} \right\}$$

qui est clairement de dimension $\text{Card}(I) + 1$.

4.3.3. Propriété combinatoire. Une propriété fondamentale des classes de Vapnik-Chervonenkis est le lemme suivant, qui donne une garantie sur la manière dont croît $N_{\mathcal{A}}(n)$ pour $n \geq V$.

LEMME 4.1 (Lemme de Sauer [**VC71, Sau72**]). *Soit $\mathcal{A}$ une classe de Vapnik-Chervonenkis de dimension $V_{\mathcal{A}} < \infty$. Alors, pour tout $n \geq V$,*

$$N_{\mathcal{A}}(n) \leq \sum_{k=1}^{V_{\mathcal{A}}} C_n^k \leq \left(\frac{en}{V}\right)^V \ . \tag{4.10}$$

DÉMONSTRATION. Voir [**DGL96**, Section 13.1]. □

4.3.4. Majoration de l'erreur d'estimation. D'après (4.7), l'erreur d'estimation est majorée par

$$2\mathcal{B}_P(S;D_n) = 2\sup_{t\in S}\left|\widehat{\mathcal{R}}_n\left(t;D_n\right) - \mathcal{R}_P\left(t\right)\right| \quad .$$

Lorsque S est fini, d'après le Théorème 4.1, cette quantité est majorée par 2ϵ avec probabilité au plus $1-2\operatorname{Card}(S)\exp\left(-2n\epsilon^2\right)$, et donc en espérance par $C\sqrt{\ln(e\operatorname{Card}(S))/n}$ pour une constante $C>0$.

Or, étant donnés $X_1,\ldots,X_n \in \mathcal{X}$, seuls $N_{\mathcal{A}(S)}(X_1,\ldots,X_n) \leq N_{\mathcal{A}(S)}(n)$ classifieurs $t\in S$ sont discernables au vu des données. On peut montrer formellement [**DGL96**, Théorème 12.6] que $\mathcal{B}_P(S;D_n)$ se comporte effectivement (à constante multiplicative près) comme si S était de cardinal fini $N_{\mathcal{A}(S)}(n)$.

THÉORÈME 4.3. *Soit $S=S_{\mathcal{A}}$ un modèle. Alors, pour tout $\epsilon>0$ et $n\in\mathbb{N}$,*

$$\mathbb{P}\left(\mathcal{B}_P(S;D_n)>\epsilon\right) \leq 8N_{\mathcal{A}}(n)\exp\left(\frac{-n\epsilon^2}{32}\right) \quad .$$

REMARQUE 4.16. La quantité $N_{\mathcal{A}}(n)$ est pessimiste puisqu'elle correspond à la pire configuration de $X_1,\ldots,X_n$. On pourrait montrer un résultat similaire en remplaçant $N_{\mathcal{A}}(n)$ par $\mathbb{E}\left[N_{\mathcal{A}}(X_1,\ldots,X_n)\right]$.

On déduit de (4.7), du Théorème 4.3 et du Lemme 4.1 la borne suivante sur le risque de $\widehat{s}_S$.

THÉORÈME 4.4. *Il existe une constante $C>0$ telle que pour toute classe de Vapnik-Chervonenkis S de dimension $V(S)$ et toute mesure de probabilité P sur $\mathcal{X}\times\mathcal{Y}$,*

$$\mathbb{E}_{D_n\sim P^{\otimes n}}\left[\ell\left(s^\star,\widehat{s}_S(D_n)\right)\right] \leq \ell\left(s^\star,S\right) + C\sqrt{\frac{V(S)}{n}} \quad . \tag{4.11}$$

REMARQUE 4.17. Les constantes apparaissant dans les bornes théoriques sur le risque telles que (4.11) sont en général pessimistes. Néanmoins, les expériences numériques montrent qu'en général, l'erreur d'estimation évolue avec $V(S)$ et n proportionnellement à $\sqrt{V/n}$. Seules les constantes numériques peuvent être réellement améliorées.

De plus, comme l'indique la section suivante, on peut montrer de façon théorique que la borne (4.11) est optimale à la valeur de la constante C près, même en considérant d'autres estimateurs $\widehat{s}$ que les minimiseurs du risque empirique.

4.4. Risque minimax

Le *risque minimax* donne un moyen de définir l'optimalité d'un classifieur $\widehat{s}$. Si $\mathcal{P}$ est une famille de mesures de probabilité sur $\mathcal{X}\times\mathcal{Y}$, on pose

$$\mathcal{R}_{\min\max}(\mathcal{P}) = \inf_{\widetilde{s}}\max_{P\in\mathcal{P}}\mathbb{E}_P\left[\ell\left(s^\star,\widetilde{s}\right)\right]$$

où l'inf est pris sur tous les estimateurs $\widetilde{s}$.

Comme nous l'avons vu au Chapitre 2 (Théorème 2.3), le risque minimax associé à l'ensemble $\mathcal{P}^\star$ des mesures de probabilité sur $\mathcal{X}\times\mathcal{Y}$ est égal à

$$\mathcal{R}_{\min\max}(\mathcal{P}^\star) = \frac{1}{2}$$

dès lors que $\mathcal{X}$ est infini.

En revanche, si l'on se restreint à l'ensemble $\mathcal{P}(S)$ des distributions telles que $s^\star\in S$ pour un modèle S donné, alors le risque minimax tend vers zéro lorsque n tend vers l'infini si S n'est

pas trop grand, d'après le Théorème 4.4. En utilisant des arguments combinatoires, Devroye et Lugosi [**DL95**] ont ainsi montré qu'il existe un constante $C > 0$ telle que

$$\mathcal{R}_{\min\max}(\mathcal{P}(S)) \geq k_2\sqrt{\frac{V}{n}} \ . \tag{4.12}$$

À une constante près multiplicative près, tout minimiseur $\widehat{s}_S$ du risque empirique sur S est donc optimal au sens minimax parmi les distributions $P \in \mathcal{P}(S)$.

REMARQUE 4.18. Faire l'hypothèse que $P \in \mathcal{P}(S)$ pour un modèle donné permet donc de garantir une convergence du risque de $\widehat{s}_S$ *uniformément sur toutes les distributions P qui satisfont cette hypothèse.* Par ailleurs, au vu des majorations du risque de $\widehat{s}_S$ obtenues précédemment, l'erreur d'approximation mesure ce que l'on risque de perdre lorsque l'hypothèse $P \in \mathcal{P}(S)$ n'est pas vérifiée.

On pourrait penser qu'ici s'achève l'étude du risque des minimiseurs du risque empirique, mais il n'en est rien. Considérer la famille $\mathcal{P}(S)$ est en effet très pessimiste, et il est possible d'obtenir une décroissance du risque de classification beaucoup plus rapide parmi certaines distributions P telles que $s^\star \in S$.

4.5. Cas zéro-erreur et vitesses rapides

Dans le cas zéro-erreur avec $s^\star \in S$, on peut même obtenir une borne plus fine sur l'erreur d'estimation :

THÉORÈME 4.5. *Si $\mathcal{R}_P^\star = 0$ et $s^\star \in S$ fini, alors, pour tout $\epsilon > 0$,*

$$\mathbb{P}\left(\mathcal{R}_P\left(\widehat{s}_S\right) \geq \epsilon\right) \leq \mathrm{Card}(S)\exp\left(-n\epsilon\right) \tag{4.13}$$

$$\mathbb{E}\left[\mathcal{R}_P\left(\widehat{s}_S\right)\right] \leq \frac{1+\ln(\mathrm{Card}(S))}{n} \ . \tag{4.14}$$

DÉMONSTRATION. Comme $\mathcal{R}_P\left(s_S^\star\right) = 0$, alors, avec probabilité 1, $\widehat{\mathcal{R}}_n\left(s_S^\star\right) = 0$. Par conséquent, $\widehat{s}_S$ qui est un minimiseur du risque empirique a un risque empirique nul avec probabilité 1, et donc

$$\begin{aligned}
\mathbb{P}\left(\mathcal{R}_P\left(\widehat{s}_S\right) > \epsilon\right) &\leq \mathbb{P}\left(\exists t \in S,\, \widehat{\mathcal{R}}_n\left(t\right) = 0 \text{ et } \mathcal{R}_P\left(t\right) > \epsilon\right) \\
&\leq \sum_{t\in S,\, \mathcal{R}_P(t)>\epsilon} \mathbb{P}\left(\widehat{\mathcal{R}}_n\left(t\right) = 0\right) \\
&\leq \mathrm{Card}(S)(1-\epsilon)^n \ ,
\end{aligned}$$

d'où l'on déduit (4.13) puisque $1 - x \leq e^{-x}$. Ainsi, pour tout $u \geq 0$,

$$\begin{aligned}
\mathbb{E}\left[\mathcal{R}_P\left(\widehat{s}_S\right)\right] &= \int_0^\infty \mathbb{P}\left(\mathcal{R}_P\left(\widehat{s}_S\right) > \epsilon\right) \mathrm{d}\epsilon \\
&\leq u + \int_u^\infty \mathbb{P}\left(\mathcal{R}_P\left(\widehat{s}_S\right) > \epsilon\right) \mathrm{d}\epsilon \\
&\leq u + \mathrm{Card}(S)\int_u^\infty \exp\left(-n\epsilon\right) \mathrm{d}\epsilon \\
&= u + \frac{\mathrm{Card}(S)e^{-nu}}{n} \ .
\end{aligned}$$

La majoration de l'espérance (4.14) s'en déduit en prenant $u = \ln(\mathrm{Card}(S))/n$. □

Le fait que l'erreur d'estimation décroisse plus vite vers zéro dans le cas zéro-erreur ne se limite pas aux modèles S finis contenant $s^\star$. En effet, on peut montrer qu'il existe des constantes $C_1, C_2 > 0$ telles que si S est une classe de Vapnik de dimension $V(S) \geq 2$, alors

$$\mathbb{E}\left[\mathcal{R}_P\left(\widehat{s}_S\right)\right] \leq \ell\left(s^\star, S\right) + \frac{C_1 V(S)\left[1 + \ln\left(\frac{n}{V(S)}\right)\right]}{n}$$

$$\text{et} \quad \mathbb{E}\left[\mathcal{R}_P\left(\widehat{s}_S\right)\right] \leq \frac{C_2 V(S)}{n} \qquad \text{si} \quad s^\star \in S \ .$$

Au facteur logarithmique près, ces vitesses sont optimales dans le cas zéro-erreur au sens minimax. En effet, pour tout $S \in \mathbb{S}$, notons $\mathcal{P}_z(S)$ l'ensemble des distributions zéro-erreur sur $\mathcal{X} \times \mathcal{Y}$ telles que $s^\star \in S$. Alors, si S est une classe de Vapnik de dimension $V(S)$, pour tout $n \geq V(S) - 1$,

$$\inf_{\widetilde{s}} \sup_{P \in \mathcal{P}_z(S)} \left\{\mathbb{E}\left[\mathcal{R}_P\left(\widehat{s}\right)\right]\right\} \geq \frac{V(S) - 1}{2en}\left(1 - \frac{1}{n}\right) \ .$$

Ceci illustre qu'une "bonne" règle de classification doit aussi être capable d'atteindre une vitesse plus rapide que $n^{-1/2}$ lorsque cela est possible.

Il est à noter que l'erreur d'estimation peut décroître plus vite que $n^{-1/2}$ hors du cas zéro-erreur. Par exemple, Massart et Nédélec **[MN06]** ont montré qu'il existe une constante C_3 telle que pour toute distribution P telle que $|2\eta(X) - 1| \geq h > 0$ p.s. et tout modèle S de dimension de Vapnik $V(S)$,

$$\mathbb{E}\left[\mathcal{R}_P\left(\widehat{s}_S\right)\right] \leq 2\ell\left(s^\star, S\right) + \frac{C_3 V(S)\left[1 + \ln\left(\frac{nh^2}{V(S)}\right)\right]}{nh} \ .$$

On dispose également de minorations du risque minimax dans ce cadre. Notons $\mathcal{P}(h, S)$ désigne l'ensemble des distributions sur $\mathcal{X} \times \mathcal{Y}$ telles que $s^\star \in S$ et $|2\eta(X) - 1| \geq h > 0$ p.s.. Massart et Nédélec **[MN06]** ont montré qu'il existe une constante $C_4 > 0$ telle que pour tout modèle S de dimension de Vapnik $V(S)$,

$$\inf_{\widetilde{s}} \sup_{P \in \mathcal{P}(h,S)} \left\{\mathbb{E}\left[\mathcal{R}_P\left(\widehat{s}\right)\right]\right\} \geq C_4 \min\left\{\frac{V(S)}{nh}; \sqrt{\frac{V(S)}{n}}\right\} \ ,$$

et le facteur supplémentaire $\ln(n/V)$ apparaît dans la borne inférieure pour certaines classes de Vapnik S. Le cas zéro-erreur correspond à $h = 1$.

Plus généralement, toutes les vitesses intermédiaires entre $n^{-1/2}$ et n^{-1} peuvent être obtenues en fonction de la distribution de $\eta(X)$ au voisinage de $1/2$. En effet, les distributions $P \in \mathcal{P}(S)$ pour lesquelles il est le plus difficile de réaliser un bon classifieur sont celles pour lesquelles $\eta(X)$ peut être arbitrairement proche de $1/2$. C'est pourquoi Mammen et Tsybakov **[MT99]** ont introduit la *condition de marge* suivante sur P :

$$\exists C, \alpha, t_0 \geq 0, \ \forall t \in [t_0, 1], \qquad \mathbb{P}\left(0 < |\eta(X) - 1/2| \leq t\right) \leq C t^\alpha \ . \tag{4.15}$$

Le cas précédent correspond à $t_0 = h/2$ et la limite $\alpha = +\infty$.

Massart et Nédélec **[MN06]** ont montréque pour toute distribution P vérifiant (4.15) et tout modèle S de dimension de Vapnik $V(S)$, il existe une constante $C_5(C, \alpha, t_0)$ telle que

$$\mathbb{E}\left[\mathcal{R}_P\left(\widehat{s}_S\right)\right] \leq 2\ell\left(s^\star, S\right) + C_5(C, \alpha, t_0)\left(\frac{V(S)\left[1 + \ln\left(\frac{n}{V(S)}\right)\right]}{n}\right)^{\alpha/(\alpha+1)} \ .$$

On dispose également de minorations du risque minimax dans ce cadre, qui coincident avec la borne supérieure au terme logarithmique près.

4.6. Complexités de Rademacher

Les limites de la notion de classe de Vapnik-Chervonenkis sont, d'une part, le pessimisme des bornes correspondantes sur le risque, et d'autre part, le fait que des modèles de taille "raisonnable" sont trop riches pour être des classes de Vapnik-Chervonenkis. La notion de complexité de Rademacher permet de mesurer finement la capacité d'un modèle S en tenant compte de la distribution P des données. De plus, nous verrons au Chapitre 5 que les complexités de Rademacher peuvent être évaluées avec précision à l'aide des données uniquement.

Soit $D_n = (X_i, Y_i)_{1\leq i\leq n}$ un n-échantillon i.i.d. de loi P et $D'_n = (X'_i, Y'_i)_{1\leq i\leq n}$ une copie indépendante de ce n-échantillon. On a alors

$$\begin{aligned}
\mathbb{E}\left[\mathcal{B}_P(S; D_n)\right] &= \mathbb{E}_{D_n\sim P^{\otimes n}}\left[\sup_{t\in S}\left|\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{1}_{t(X_i)\neq Y_i} - \mathcal{R}_P\left(t\right)\right)\right|\right] \\
&= \mathbb{E}_{D_n\sim P^{\otimes n}}\left[\sup_{t\in S}\left|\mathbb{E}_{D'_n\sim P^{\otimes n}}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{1}_{t(X_i)\neq Y_i} - \mathbb{1}_{t(X'_i)\neq Y'_i}\right)\right]\right|\right] \\
&\leq \mathbb{E}_{D_n, D'_n\sim P^{\otimes n}}\left[\sup_{t\in S}\left|\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{1}_{t(X_i)\neq Y_i} - \mathbb{1}_{t(X'_i)\neq Y'_i}\right)\right|\right]
\end{aligned}$$

en utilisant l'inégalité de Jensen $\sup_t |\mathbb{E}Z_t| \leq \mathbb{E}\left[\sup_t |Z_t|\right]$. Or, $\left(\mathbb{1}_{t(X_i)\neq Y_i} - \mathbb{1}_{t(X'_i)\neq Y'_i}\right)_{1\leq i\leq n}$ a la même loi que $\left(\epsilon_i\left(\mathbb{1}_{t(X_i)\neq Y_i} - \mathbb{1}_{t(X'_i)\neq Y'_i}\right)\right)_{1\leq i\leq n}$ où $\epsilon_1, \ldots, \epsilon_n$ sont des variables i.i.d. et indépendantes de D_n, D'_n avec $\mathbb{P}\left(\epsilon_1 = 1\right) = \mathbb{P}\left(\epsilon_1 = -1\right) = 1/2$. On a alors

$$\begin{aligned}
\mathbb{E}\left[\mathcal{B}_P(S; D_n)\right] &\leq \mathbb{E}\left[\sup_{t\in S}\left|\frac{1}{n}\sum_{i=1}^{n}\left[\epsilon_i\left(\mathbb{1}_{t(X_i)\neq Y_i} - \mathbb{1}_{t(X'_i)\neq Y'_i}\right)\right]\right|\right] \\
&\leq \mathbb{E}\left[\sup_{t\in S}\left|\frac{1}{n}\sum_{i=1}^{n}\left[\epsilon_i\mathbb{1}_{t(X_i)\neq Y_i}\right]\right|\right] + \mathbb{E}\left[\sup_{t\in S}\left|\frac{1}{n}\sum_{i=1}^{n}\left[-\epsilon_i\mathbb{1}_{t(X'_i)\neq Y'_i}\right]\right|\right] \\
&= 2\mathbb{E}\left[\sup_{t\in S}\left|\frac{1}{n}\sum_{i=1}^{n}\left[\epsilon_i\mathbb{1}_{t(X_i)\neq Y_i}\right]\right|\right] \quad .
\end{aligned}$$

Les ϵ_i étant appelées variables de Rademacher, le majorant obtenu pour $\mathbb{E}\left[\mathcal{B}_P(S; D_n)\right]$ est appelé (au facteur 2 près) *complexité de Rademacher* du modèle S :

$$R_n(S; P) := \mathbb{E}\left[\sup_{t\in S}\left|\frac{1}{n}\sum_{i=1}^{n}\left[\epsilon_i\mathbb{1}_{t(X_i)\neq Y_i}\right]\right|\right] \tag{4.16}$$

Si la démonstration précédente montre que

$$\mathbb{E}\left[\mathcal{B}_P(S; D_n)\right] \leq 2R_n(S; P) \ ,$$

on constate expérimentalement que dans la plupart des cas, le facteur n'est pas nécessaire et que $R_n(S; P)$ est une bonne éstimation de $\mathcal{B}_P(S; D_n)$.

Au-delà des calculs précédents, on peut donner l'interprétation suivante de la complexité de Rademacher. Pour tout échantillon D_n et vecteur de signes $\epsilon_1, \ldots, \epsilon_n$, on définit le sous-échantillon $D_n^- = (X_i, Y_i)_{\epsilon_i=-1}$, sa taille $N_- = \text{Card}\left\{i \text{ t.q. } \epsilon_i = -1\right\}$ et sa mesure empirique

associée $P_n^- = N_-^{-1} \sum_{\epsilon_i=-1} \delta_{(X_i,Y_i)}$. Alors, la complexité de Rademacher (4.16) peut s'écrire

$$\begin{aligned} R_n(S;P) &= \mathbb{E}\left[\sup_{t\in S}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{t(X_i)\neq Y_i} - \frac{2}{n}\sum_{\epsilon_i=-1}\mathbb{1}_{t(X_i)\neq Y_i}\right|\right] \\ &= \mathbb{E}\left[\sup_{t\in S}\left|\mathcal{R}_{P_n}(t) - \frac{2N_-}{n}\mathcal{R}_{P_n^-}(t)\right|\right] \quad . \end{aligned}$$

Compte-tenu du fait que N_- est proche de $n/2$ avec grande probabilité, la complexité de Rademacher peut être vue comme l'espérance des déviations uniformes sur S de l'écart entre le risque empirique sur un échantillon D_n de taille n et le risque empirique sur un sous-échantillon aléatoire de D_n de taille $n/2$. Or, la quantité qu'estime $R_n(S;P)$ est

$$\mathbb{E}\left[\mathcal{B}_P(S;D_n)\right] = \mathbb{E}\left[\sup_{t\in S}\left|\widehat{\mathcal{R}}_n(t;D_n) - \mathcal{R}_P(t)\right|\right] \quad .$$

Pour passer de $\mathcal{B}_P(S;D_n)$ à $R_n(S;P)$, on a donc simplement remplacé le couple de mesures (P,P_n) par le couple (P_n,P_n^-), dans l'idée que la "distance" entre P et P_n est bien évaluée par la "distance" entre P_n et P_n^-. Il s'agit en fait d'un principe beaucoup plus général, appelé *rééchantillonnage*, et qui permet d'obtenir des estimations de quantités telles que $\mathcal{B}_P(S;D_n)$ à l'aide des données uniquement [**Fro07**].

CHAPITRE 5

Calibration d'algorithmes et sélection de modèles

Nous avons vu au cours des deux chapitres précédents deux grandes familles de règles de classification, chacune composée de sous-familles naturelles : k-plus proches voisins, partition sur une grille régulière de pas $h > 0$, minimiseur du risque empirique sur le modèle de classification par K intervalles, *etc.* Dans chaque cas, un ou plusieurs paramètres (k, h, K, ...) restent à calibrer, phase dont l'importance est cruciale pour obtenir au final la meilleure règle de classification possible.

L'objet de ce chapitre est d'étudier les différentes approches envisageables pour calibrer un algorithme de classification, et plus généralement pour choisir parmi un ensemble de règles de classification.

5.1. Problématique de la calibration

Soit $(\widehat{s}_\lambda)_{\lambda\in\Lambda}$ famille de règles de classification et D_n un échantillon. L'objectif de la calibration est de choisir un classifieur parmi $\{\widehat{s}_\lambda(D_n)\}_{\lambda\in\Lambda}$ en n'utilisant que les données. Autrement dit, on cherche $\widehat{\lambda}(D_n) \in \Lambda$ tel que le risque de $\widehat{s}_{\widehat{\lambda}(D_n)}(D_n)$ soit aussi petit que possible.

5.1.1. Calibration idéale, oracle. Le choix idéal, appelé *oracle*, est la valeur du paramètre λ minimisant le risque :

$$\lambda^\star = \lambda^\star(D_n) \in \arg\min_{\lambda\in\Lambda} \{\mathcal{R}_P(\widehat{s}_\lambda(D_n))\} \quad .$$

En pratique, l'oracle est inaccessible car il dépend de la distribution P des données. Un objectif plus réaliste pour $\widehat{\lambda} = \widehat{\lambda}(D_n)$ est qu'il vérifie une *inégalité-oracle*, c'est-à-dire

$$\ell\left(s^\star, \widehat{s}_{\widehat{\lambda}(D_n)}(D_n)\right) \le C \inf_{\lambda\in\Lambda} \{\ell(s^\star, \widehat{s}_\lambda(D_n)) + \epsilon_\lambda\} \tag{5.1}$$

en espérance ou avec grande probabilité. Une telle inégalité est d'autant meilleure que C est proche de 1 et ϵ_λ est petit (idéalement, ϵ_λ doit être négligeable devant $\ell(s^\star, \widehat{s}_\lambda(D_n))\,P$ lorsque $\lambda \approx \lambda^\star$).

L'inégalité (5.1) permet de définir naturellement un critère de qualité pour une méthode de calibration $D_n \mapsto \widehat{\lambda}(D_n)$, à savoir la constante C qui apparaîtrait dans une inégalité-oracle (5.1) lorsque $\epsilon_\lambda = 0$:

$$C_{\text{or}}\left(\widehat{\lambda}(\cdot); D_n\right) := \frac{\ell\left(s^\star, \widehat{s}_{\widehat{\lambda}(D_n)}\right)}{\inf_{\lambda\in\Lambda}\{\ell(s^\star, \widehat{s}_\lambda(D_n))\}} \quad . \tag{5.2}$$

Idéalement, on voudrait que $C_{\text{or}}\left(\widehat{\lambda}(\cdot); D_n\right)$ tende vers 1 lorsque n tend vers l'infini, en espérance ou avec grande probabilité. Cet objectif n'est cependant pas toujours atteignable, les problèmes de calibration les plus difficiles correspondant au cas où la perte du classifieur oracle $\widehat{s}_{\lambda^\star}(D_n)$ est particulièrement faible, soit parce que Λ est grand, soit parce que la famille $(\widehat{s}_\lambda)_{\lambda\in\Lambda}$ est parfaitement adaptée à la distribution P des données.

5.1.2. Exemples. Au vu des différentes familles de règles de classification introduites dans les Chapitres 3 et 4, on peut citer notamment les exemples suivants de problèmes de calibration :

- *Choix de k pour les k plus proches voisins* : $\Lambda = \{1, \dots, n\}$ et pour tout $\lambda \in \Lambda$, $\widehat{s}_\lambda$ est la règle des λ plus proches voisins (une distance d sur $\mathcal{X}$ étant fixée).
- *Choix de k et d'une distance d pour les k plus proches voisins* : $\Lambda = \{1, \dots, n\} \times \{d_1, \dots, d_m\}$, où $d_1, \dots, d_m$ sont des distances sur $\mathcal{X}$, et pour tout $\lambda = (k, d_j) \in \Lambda$, $\widehat{s}_\lambda$ est la règle des k plus proches voisins pour la distance d_j.
- *Choix du pas de la grille pour la règle par partition régulière* : $\Lambda =]0; +\infty]$ et pour tout $\lambda \in \Lambda$, $\widehat{s}_\lambda$ est la règle de classification par partition sur une grille régulière de pas λ (avec $\mathcal{X} = \mathbb{R}^\ell$).
- *Sélection de modèles* parmi une famille $(S_m)_{m \in \mathcal{M}_n}$ de modèles : $\Lambda = \mathcal{M}_n$ et pour tout $\lambda \in \Lambda$, $\widehat{s}_\lambda$ est une règle minimisant le risque empirique sur S_λ. Cet exemple sera étudié en détail à la Section 5.2.
- *Choix du nombre d'intervalles pour la classification par intervalles* : $\Lambda = \{1, \dots, n\}$ et pour tout $\lambda \in \Lambda$, $\widehat{s}_\lambda$ est une règle minimisant le risque empirique sur le modèle $S_{\text{inter}}(\lambda)$ de classification par intervalles (avec $\mathcal{X} = [0, 1]$).
- Et l'on peut bien sûr panacher différentes familles de règles de classification. Par exemple, posons $\Lambda = (\{0, 1\} \times \{1, \dots, n\}) \cup (\{2\} \times]0; +\infty])$ et pour tout $\lambda \in \Lambda$, $\widehat{s}_\lambda$ est la règle des k plus proches voisins si $\lambda = (0, k)$, un minimiseur du risque empirique sur le modèle $S_{\text{inter}}(K)$ si $\lambda = (1, K)$ et $\widehat{s}_\lambda$ est la règle de classification par partition sur une grille fixe de pas h si $\lambda = (2, h)$.

Remarque 5.1 (Calibration d'un paramètre continu). Lorsque l'ensemble Λ des valeurs possibles du paramètre λ est continu, ou plus généralement si Λ est infini, choisir parmi l'ensemble des règles $(\widehat{s}_\lambda)_{\lambda \in \Lambda}$ est susceptible de poser des problèmes algorithmiques. Deux solutions sont possibles.

Si l'on dispose d'un algorithme rapide pour calculer l'ensemble de la trajectoire $\{\widehat{s}_\lambda(D_n)\}_{\lambda \in \Lambda}$, alors il est envisageable de garder un ensemble continu de valeurs possibles pour λ. C'est par exemple le cas des Support Vector Machines, dont le chemin de régularisation $\{\widehat{s}_\lambda(D_n)\}_{\lambda \in \Lambda}$ est linéaire par morceaux.

Sinon, il faut discrétiser Λ, suivant une échelle linéaire ou logarithmique, selon la nature du paramètre λ.

5.1.3. Estimation sans biais du risque. Une approche générale pour le problème de la calibration est la suivante :

(1) pour tout $\lambda \in \Lambda$, construire un estimateur $\widehat{\mathcal{R}}(\widehat{s}_\lambda; D_n)$ non-biaisé (ou presque) du risque de $\widehat{s}_\lambda(D_n)$,
(2) choisir $\widehat{\lambda}(D_n) \in \arg\min_{\lambda \in \Lambda} \left\{ \widehat{\mathcal{R}}(\widehat{s}_\lambda; D_n) \right\}$.

C'est par exemple le cas des méthodes de validation croisée, sur lesquelles nous reviendrons à la Section 5.3 ; voir en particulier la Section 5.3.2 à propos du biais de l'estimateur par validation croisée du risque.

Remarque 5.2 (Estimation biaisée du risque). Il est à noter que l'estimation sans biais du risque peut être réalisée *à constante près*. En effet, si $\widehat{\mathcal{R}}(\widehat{s}_\lambda; D_n)$ estime sans biais $\mathcal{R}_P(\widehat{s}_\lambda(D_n)) + c$ pour un $c \in \mathbb{R}$ indépendant de $\lambda \in \Lambda$ (mais pouvant dépendre de P), alors minimiser $\widehat{\mathcal{R}}(\widehat{s}_\lambda; D_n)$ en $\lambda \in \Lambda$ revient à minimiser $\widehat{\mathcal{R}}(\widehat{s}_\lambda; D_n) - c$ en $\lambda \in \Lambda$, et donc un estimateur sans biais du risque.

Il est en effet important de garder à l'esprit que la qualité de $\widehat{\lambda}$ ne dépend que des *variations du biais* avec λ dans l'estimation du risque de $\widehat{s}_\lambda(D_n)$.

REMARQUE 5.3 (Risque empirique et sur-apprentissage). Nous avons vu à plusieurs reprises dans les chapitres précédents que le risque empirique $\widehat{\mathcal{R}}_n\left(\widehat{s}_\lambda(D_n); D_n\right)$ peut sous-estimer fortement le risque d'un classifieur $\widehat{s}_\lambda(D_n)$ obtenu à l'aide des mêmes données. Par exemple, si les X_i sont tous distincts, le classifieur des plus proches voisins (1-ppv) a toujours un risque empirique nul, alors que son risque ne converge en général pas vers le risque de Bayes lorsque n tend vers l'infini (Section 3.3.2).

Ce biais du risque empirique pour estimer le risque des k-ppv étant d'autant plus fort que le nombre k de voisins est petit, utiliser $\widehat{\mathcal{R}}\left(\widehat{s}_\lambda; D_n\right) = \widehat{\mathcal{R}}_n\left(\widehat{s}_\lambda(D_n); D_n\right)$ pour choisir k conduirait au *sur-apprentissage* (overfitting), c'est-à-dire à choisir un classifieur collant aux données et non un classifieur permettant de bien classifier de nouvelles données.

REMARQUE 5.4 (Taille de Λ). Le sur-apprentissage est également possible lorsque $\left(\widehat{s}_\lambda\right)_{\lambda\in\Lambda}$ est trop riche. Par exemple, si $\Lambda = \mathbb{S}$ et $\forall\lambda \in \Lambda$, $\widehat{s}_\lambda(D_n) = \lambda$, alors $\widehat{\mathcal{R}}_n\left(\widehat{s}_\lambda; D_n\right)$ estime sans biais le risque de $\widehat{s}_\lambda$ (et la validation croisée ne ferait pas mieux). Cependant, minimiser $\widehat{\mathcal{R}}_n\left(\widehat{s}_\lambda; D_n\right)$ en $\lambda \in \Lambda$ conduit au sur-apprentissage, comme vu en Section 4.1.2.

Par contre, si $\mathrm{Card}(\Lambda) \leq Cn^\alpha$, alors minimiser un estimateur sans biais du risque de $\widehat{s}_\lambda$ est très généralement une bonne stratégie, dès lors que n est suffisamment grand. Lorsque $\mathrm{Card}(\Lambda)$ est plus grand (en particulier infini), la qualité de la stratégie de minimisation d'un estimateur sans biais du risque est fonction des dépendances entre les $\widehat{s}_\lambda(D_n)$.

REMARQUE 5.5 (Problème de calibration «simple»). Lorsque Λ ne dépend pas de n et que n est «grand» (en un certain sens), choisir λ minimisant $\widehat{\mathcal{R}}_n\left(\widehat{s}_\lambda(D_n); D_n\right)$ peut être une bonne méthode de calibration, contrairement à ce que pourrait laisser penser la Remarque 5.3.

Considérons par exemple une famille de minimiseurs du risque empirique sur des modèles $\left(S_m\right)_{m\in\mathcal{M}}$ qui sont chacun une classe de Vapnik-Chervonenkis de dimension V_m. Alors, d'après le Théorème 4.3 et le Lemme 4.1, pour chaque $m \in \mathcal{M}$,

$$\left|\mathcal{R}_P\left(\widehat{s}_{S_m}(D_n)\right) - \widehat{\mathcal{R}}_n\left(\widehat{s}_{S_m}(D_n); D_n\right)\right| \leq \mathcal{B}_P(S_m; D_n) \leq C\sqrt{\frac{V_m}{n}}$$

avec grande probabilité. Par conséquent, si $n \gg \max_{m\in\mathcal{M}} V_m$,

$$\widehat{m} \in \arg\min_{m\in\mathcal{M}} \left\{\widehat{\mathcal{R}}_n\left(\widehat{s}_{S_m}(D_n); D_n\right)\right\}$$

choisit avec grande probabilité un modèle minimisant le biais $\ell\left(s^\star, S_m\right)$.

Si $\arg\min_{m\in\mathcal{M}} \ell\left(s^\star, S_m\right)$ est unique, alors, avec grande probabilité, il s'agit du modèle oracle $m^\star$, et $\widehat{m} = m^\star$. Minimiser le risque empirique conduit alors à une inégalité-oracle (5.1) avec constante $C = 1$ et $\epsilon_\lambda = 0$, sur un événement de grande probabilité.

Cet exemple est typique des problèmes de calibration «faciles», où l'on cherche à comparer un ensemble de classifieurs «simples» (en comparaison du grand nombre d'observations n). Les problèmes de calibration «difficiles» correspondent donc aux cas où Λ évolue avec n ou est infini, de telle sorte que certains des classifieurs $\widehat{s}_\lambda(D_n)$ sur-apprennent. L'objet de ce chapitre est de proposer des manières d'éviter le sur-apprentissage.

5.2. Sélection de modèles

La sélection de modèles est un cas particulier de problème de calibration. On suppose donnée une famille de modèles $\left(S_m\right)_{m\in\mathcal{M}_n}$. Pour tout $m \in \mathcal{M}_n$, on note $\widehat{s}_m(D_n) = \widehat{s}_{S_m}(D_n)$ un minimiseur du risque empirique sur S_m. L'objectif est alors de choisir $\widehat{m}(D_n) \in \mathcal{M}_n$ tel que le risque de $\widehat{s}_{\widehat{m}(D_n)}(D_n)$ soit aussi petit que possible.

Les références de cette section sont [**Mas07**] (pour les aspects les plus théoriques), [**HTF01**, Chapitre 7] (livre plus orienté vers la pratique) et [**DGL96**, Chapitre 18] (à propos de la minimisation du risque structurel).

5.2.1. Compromis biais-variance. La problématique de la sélection de modèles repose pour l'essentiel sur la réalisation d'un compromis entre biais (l'erreur d'approximation $\ell\left(s^\star, S_m\right)$) et variance (l'erreur d'estimation $\mathcal{R}_P\left(\widehat{s}_m(D_n)\right) - \mathcal{R}_P\left(s_m^\star(D_n)\right)$), comme nous l'avons évoqué en Section 4.2.1.

Considérons par exemple le cas de la famille de modèles $\left(S_{\text{inter}}(K)\right)_{1\leq K\leq n}$ de classification par intervalles, avec $\mathcal{X} = [0,1]$.

L'erreur d'approximation est alors une fonction décroissante de K. Lorsque $s^\star$ a un nombre fini $K_0 - 1$ de sauts dans $[0,1]$, l'erreur d'approximation est nulle dès que $K \geq K_0$. Pour un classifieur de Bayes $s^\star$ général, l'erreur d'approximation $\ell\left(s^\star, S_{\text{inter}}(K)\right)$ tend vers zéro quand K tend vers l'infini, à une vitesse dépendant de la distribution P.

À l'inverse, l'erreur d'estimation tend à croître avec le nombre de ruptures $K-1$ à placer dans $[0,1]$. Ainsi, la borne supérieure $\mathcal{B}_P(S_{\text{inter}}(K); D_n)$ sur l'erreur d'estimation est une fonction croissante de K, de même que la majoration déterministe $C\sqrt{K/n}$ obtenue au Chapitre 4.

Le risque de classification étant la somme de l'erreur d'approximation et de l'erreur d'estimation, le minimiser nécessite de trouver un compromis entre ces deux termes qui évoluent dans des directions opposées lorsque K varie. Typiquement, pour le meilleur modèle (ici, le meilleur choix de K), *l'erreur d'approximation est du même ordre de grandeur que l'erreur d'estimation.*

5.2.2. Principales méthodes de sélection de modèles. Nous avons présenté en Section 5.1.3 une première idée, qui est d'estimer le risque de $\widehat{s}_m(D_n)$ (par exemple par validation croisée) puis de choisir $\widehat{m}$ qui minimise cet estimateur. Dans le cas de la sélection de modèles, d'autres approches sont possibles, en particulier la pénalisation et la régularisation.

Pénalisation. Une méthode de sélection de modèles par pénalisation est définie comme suit :

$$\widehat{m}(D_n) \in \arg\min_{m\in\mathcal{M}} \left\{ \widehat{\mathcal{R}}_n\left(\widehat{s}_m(D_n)\right) + \text{pen}(m) \right\} \quad , \tag{5.3}$$

où pour tout $m \in \mathcal{M}_n$, $\text{pen}(m)$ est une *pénalité* attribuée au modèle S_m, rendant compte de sa «complexité». La pénalité peut être déterministe ou dépendre des données.

L'objectif étant de trouver $\widehat{m}(D_n)$ tel que $\mathcal{R}_P\left(\widehat{s}_m(D_n)\right)$ est minimal, la meilleure pénalité est la différence entre le risque de $\widehat{s}_m$ et son risque empirique. On l'appelle la *pénalité idéale*

$$\text{pen}_{\text{id}}(m) := \mathcal{R}_P\left(\widehat{s}_{S_m}\right) - \widehat{\mathcal{R}}_n\left(\widehat{s}_{S_m}\right) \quad . \tag{5.4}$$

En général, $\text{pen}_{\text{id}}(m) \geq 0$ avec une très forte probabilité, dans la mesure où *le risque empirique sous-estime le risque.*

Comme $\text{pen}_{\text{id}}(m)$ dépend de la distribution P qui est inconnue, elle ne peut pas être utilisée en pratique. La pénalité idéale reste cependant une bonne référence pour construire une pénalité.

Remarque 5.6. Minimiser le risque empirique pénalisé (5.3) permet de visualiser une forme de compromis entre biais et variance. En effet, $\widehat{\mathcal{R}}_n\left(\widehat{s}_m(D_n)\right) + \text{pen}(m)$ est généralement égal au terme de biais auquel on a retranché une partie du terme de variance, tandis que la pénalité $\text{pen}(m)$ doit être égale au terme de variance, auquel on ajoute la différence entre le biais et le risque empirique. La pénalité est donc proportionnelle au terme de variance, à savoir l'erreur d'estimation.

On distingue les méthodes de pénalisation suivantes :

(1) Pour tout $m \in \mathcal{M}_n$, on utilise comme pénalité *un estimateur* $\mathrm{pen}(m; D_n)$ *de* $\mathrm{pen}_{\mathrm{id}}(m)$. Nous verrons en Section 5.2.5 un exemple de pénalité estimant $\mathrm{pen}_{\mathrm{id}}(m)$.
Utiliser une telle pénalité revient à estimer le risque de chaque modèle, comme à la Section 5.1.3. Par exemple, la validation croisée peut être vue comme une méthode de pénalisation avec

$$\mathrm{pen}_{\mathrm{VC}}(m) := \widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{s}_m; D_n\right) - \widehat{\mathcal{R}}_n\left(\widehat{s}_m(D_n); D_n\right) \ .$$

Mentionnons tout de même deux différences entre la pénalisation par un estimateur de $\mathrm{pen}_{\mathrm{id}}(m)$ et l'approche de la Section 5.1.3. D'une part, il est naturel d'imposer que $\mathrm{pen}(m) \geq 0$ pour tout $m \in \mathcal{M}_n$, ce qui n'est pas nécessairement le cas pour $\mathrm{pen}_{\mathrm{VC}}(m)$. D'autre part, $\mathrm{pen}_{\mathrm{id}}(m)$ est d'un ordre de grandeur plus petit que $\ell\left(s^\star, \widehat{s}_m(D_n)\right)$; par conséquent, si $\mathrm{pen}(m)$ estime $\mathrm{pen}_{\mathrm{id}}(m)$ correctement au premier ordre, alors on estime $\ell\left(s^\star, \widehat{s}_m(D_n)\right)$ correctement au second ordre, ce que toutes les formes de validation croisée ne satisfont pas nécessairement.

(2) *Minimisation du risque structurel* **[VC74, Vap82, Vap98]** : on utilise comme pénalité un estimateur de

$$\mathcal{B}_P(S_m; D_n) := \sup_{t \in S_m} \left| \widehat{\mathcal{R}}_n\left(t; D_n\right) - \mathcal{R}_P\left(t\right) \right| \geq \mathrm{pen}_{\mathrm{id}}(m) \ .$$

Nous en verrons des exemples en Section 5.2.4.

(3) On peut vouloir, pour des raisons liées à la nature du problème de classification considéré, choisir un modèle S_m ou un estimateur qui possède certaines propriétés (par exemple, que $\{\, x \in X \text{ t.q. } \widehat{s}(x; D_n) = 1 \,\}$ ait un petit nombre de composantes connexes). Dans ce cas, l'objectif final étant modifié (de même que le modèle oracle $m^\star$), la pénalité doit être construite en fonction de cette contrainte. Ainsi, dans l'exemple précédent, $\mathrm{pen}(m)$ peut être prise proportionnelle au nombre de composantes connexes, la constante de proportionnalité étant à choisir avec précautions. En particulier, il n'est pas nécessaire que $\mathrm{pen}(m)$ soit reliée à $\mathrm{pen}_{\mathrm{id}}(m)$

Minimisation du risque empirique régularisé. Certaines règles de classification, telles que les SVM[1], consistent à minimiser le risque empirique (ou une version convexifiée du risque empirique) plus un terme «de régularisation» (typiquement proportionnel à une norme) :

$$\widehat{s}(\lambda; D_n) := \arg\min_{t \in \mathbb{S}} \left\{ \widehat{\mathcal{R}}_n\left(t\right) + \lambda \left\|t\right\| \right\} \ .$$

S'il ne s'agit pas *stricto sensu* d'une méthode de sélection de modèles, on peut s'y ramener en remarquant que $\widehat{s}(\lambda; D_n)$ minimise le risque empirique sur la boule $\{\, t \in \mathbb{S} \text{ t.q. } \|t\| \leq \|\widehat{s}(\lambda; D_n)\| \,\}$. Par conséquent, le chemin de régularisation $\left(\widehat{s}(\lambda; D_n)\right)_{\lambda \geq 0}$ coïncide avec $\left(\widehat{s}_r(D_n)\right)_{0 \leq r \leq \infty}$, où $\widehat{s}_r(D_n)$ minimise le risque empirique sur la boule de centre 0 et de rayon r pour la norme $\|\cdot\|$ sur $\mathbb{S}$.

Choisir λ pour un minimiseur du risque empirique régularisé (tel qu'une classifieur SVM) revient donc à choisir un modèle parmi les boules de $\mathbb{S}$ pour la norme $\|\cdot\|$.

De plus, en remarquant que

$$\min_{t \in \mathbb{S}} \left\{ \widehat{\mathcal{R}}_n\left(t\right) + \lambda \left\|t\right\| \right\} = \min_{r \geq 0,\, \|t\| \leq r} \left\{ \widehat{\mathcal{R}}_n\left(t\right) + \lambda r \right\} = \min_{r \geq 0} \left\{ \widehat{\mathcal{R}}_n\left(\widehat{s}_r(D_n)\right) + \lambda r \right\} \ ,$$

on peut montrer que $\widehat{s}(\lambda; D_n) = \widehat{s}_{\widehat{r}}(D_n)$ où $\widehat{r}$ minimise le risque empirique de $\widehat{s}_r(D_n)$ pénalisé par λr.

[1] Support Vector Machines, ou Machines à vecteurs de support, voir **[STC00]**.

Deux fausses bonnes idées. Il est imprudent en général de faire trop confiance aux résultats théoriques pour construire une règle de choix de modèles.

Une première idée pourrait être d'utiliser les bornes supérieures sur le risque obtenues au Chapitre 4. Outre le fait que le biais n'est pas connu (dans les bons cas, il peut être majoré en fonction d'informations *a priori* sur la distribution P), le défaut de cette idée est que le poids relatif de l'erreur d'approximation et de l'erreur d'estimation n'ont aucune raison d'être correcte. Si la borne théorique utilisée est dix fois plus éloignée de la réalité pour l'erreur d'estimation, alors optimiser la borne conduit à donner un poids dix fois trop important au terme de variance dans le compromis «biais-variance».

Une seconde idée (transposable au cadre général de la calibration) serait d'utiliser un résultat théorique indiquant quel est le modèle optimal $S_{m_{\mathrm{opt}}(\alpha_1,\ldots,\alpha_k)}$ avec grande probabilité, les α_j étant des caractéristiques (inconnues) de P. On estimerait ensuite séparément par les α_j par des $\widehat{\alpha}_j$ à l'aide des données. Enfin, on sélectionnerait le modèle $m_{\mathrm{opt}}\left(\widehat{\alpha}_1,\ldots,\widehat{\alpha}_k\right)$.

Les défauts de cette approche de type «plug-in» sont multiples. D'une part, l'optimalité de $m_{\mathrm{opt}}(\alpha_1,\ldots,\alpha_k)$ ne peut êtr garantie qu'en formulant des hypothèses sur P, celles-ci n'étant pas nécessairement satisfaites. D'autre part, les erreurs d'estimation des α_j risquent de se cumuler, selon la forme de $m_{\mathrm{opt}}(\alpha_1,\ldots,\alpha_k)$. Enfin, idéalement, il faudrait construire les $\widehat{\alpha}_j$ de telle sorte que le risque de $\widehat{s}_{m_{\mathrm{opt}}(\alpha_1,\ldots,\alpha_k)}$ soit minimal, objectif en général différent de «$\mathbb{E}\left[\left\|\alpha_j-\widehat{\alpha}_j\right\|^2\right]$ minimal», et particulièrement délicat à atteindre.

5.2.3. Inégalité-oracle pour la pénalisation. Soit $\rho_n > 0$ et $\widehat{m}$ un ρ_n-minimiseur du risque empirique pénalisé par $\mathrm{pen}(m)$, c'est-à-dire

$$\widehat{\mathcal{R}}_n\left(\widehat{s}_{\widehat{m}}(D_n);D_n\right)+\mathrm{pen}(\widehat{m}) \leq \min_{m\in\mathcal{M}_n}\left\{\widehat{\mathcal{R}}_n\left(\widehat{s}_m(D_n);D_n\right)+\mathrm{pen}(m)\right\}+\rho_n \ . \tag{5.5}$$

Alors, en utilisant (5.5) et le fait que pour tout $m\in\mathcal{M}_n$,

$$\widehat{\mathcal{R}}_n\left(\widehat{s}_m(D_n)\right)=\ell\left(s^\star,\widehat{s}_m(D_n)\right)+\mathcal{R}_P\left(s^\star\right)-\mathrm{pen}_{\mathrm{id}}(m) \ ,$$

on obtient pour tout $m\in\mathcal{M}_n$,

$$\begin{aligned}
&\ell\left(s^\star,\widehat{s}_{\widehat{m}}(D_n)\right)+\left(\mathrm{pen}(\widehat{m})-\mathrm{pen}_{\mathrm{id}}(\widehat{m})\right) \leq \ell\left(s^\star,\widehat{s}_m(D_n)\right)+\left(\mathrm{pen}(m)-\mathrm{pen}_{\mathrm{id}}(m)\right) \ , \\
\text{soit}\qquad &\ell\left(s^\star,\widehat{s}_{\widehat{m}}(D_n)\right)+\left(\mathrm{pen}(\widehat{m})-\mathrm{pen}_{\mathrm{id}}(\widehat{m})\right) \leq \inf_{m\in\mathcal{M}_n}\left\{\ell\left(s^\star,\widehat{s}_m(D_n)\right)+\left(\mathrm{pen}(m)-\mathrm{pen}_{\mathrm{id}}(m)\right)\right\} \ .
\end{aligned} \tag{5.6}$$

Trois situations principales sont à distinguer :

(1) Soit $\mathrm{pen}(m)\geq\mathrm{pen}_{\mathrm{id}}(m)$ pour tout $m\in\mathcal{M}_n$, comme c'est en principe le cas pour la minimisation du risque structurelle, ou lorsque $\mathrm{pen}(m)$ estime $C\,\mathrm{pen}_{\mathrm{id}}(m)$ avec $C>1$. Alors, d'après (5.6), l'inégalité-oracle suivante est satisfaite :

$$\ell\left(s^\star,\widehat{s}_{\widehat{m}}(D_n)\right) \leq \inf_{m\in\mathcal{M}_n}\left\{\ell\left(s^\star,\widehat{s}_m(D_n)\right)+\left(\mathrm{pen}(m)-\mathrm{pen}_{\mathrm{id}}(m)\right)\right\} \ .$$

La seule perte (éventuelle) par rapport à (5.1) réside dans le fait que $\mathrm{pen}(m)-\mathrm{pen}_{\mathrm{id}}(m)$ peut être non-négligeable par rapport à $\ell\left(s^\star,\widehat{s}_m(D_n)\right)$, si la pénalité $\mathrm{pen}(m)$ est beaucoup plus grande que $\mathrm{pen}_{\mathrm{id}}(m)$ au voisinage de $m=m^\star$.

(2) Soit $\sup_{m\in\mathcal{M}_n}\left|\mathrm{pen}(m)-\mathrm{pen}_{\mathrm{id}}(m)\right| \leq \epsilon\ell\left(s^\star,\widehat{s}_m(D_n)\right)\in[0;1[$, comme c'est en principe le cas si $\mathrm{pen}(m)$ estime $\mathrm{pen}_{\mathrm{id}}(m)$ pour tout $m\in\mathcal{M}_n$. Alors, (5.6) implique

$$\ell\left(s^\star,\widehat{s}_{\widehat{m}}(D_n)\right) \leq \frac{1+\epsilon}{1-\epsilon}\inf_{m\in\mathcal{M}_n}\left\{\ell\left(s^\star,\widehat{s}_m(D_n)\right)\right\} \ .$$

On a donc une inégalité-oracle (5.1) sans terme de reste, mais avec une constante multiplicative $C = C_\epsilon \geq 1$, d'autant plus proche de 1 que ϵ est proche de 0.

(3) Soit $\mathrm{pen}(m)$ sous-estime $\mathrm{pen}_{\mathrm{id}}(m)$, et (5.6) ne permet pas d'obtenir une inégalité-oracle.

Le raisonnement ci-dessus ne permet en revanche pas de conclure formellement que sous-estimer la pénalité idéale conduit au sur-apprentissage. Pour préciser l'influence de la pénalité $\mathrm{pen}(m)$ sur le risque du modèle $\widehat{m}$ choisi par (5.3), considérons le raisonnement heuristique suivant.

Supposons que $\mathrm{pen}(m)$ estime sans biais (mais avec de petites erreurs d'estimation) $K\,\mathrm{pen}_{\mathrm{id}}(m)$ pour tout modèle $m \in \mathcal{M}_n$, et notons $\widehat{m}(K)$ le modèle sélectionné par (5.3).
Lorsque $K = 0$, $\widehat{m}(K)$ minimise le risque empirique et donc sur-apprend (sauf cas particulier) ; en particulier, l'excès de risque $\ell\,(s^\star, \widehat{s}_{\widehat{m}}\,)$ est grand.
Lorsque K grandit en partant de zéro, $\widehat{m}(K)$ commence par rester parmi les modèles les plus complexes. En effet, on constate le plus souvent que $\widehat{\mathcal{R}}_n\,(\,\widehat{s}_m\,) \approx \mathcal{R}_P\,(\,s_m^\star\,) - K_{\min}\,\mathrm{pen}_{\mathrm{id}}(m)$ pour une constante $K_{\min} > 0$, si bien que $\widehat{\mathcal{R}}_n\,(\,\widehat{s}_m\,) + K\,\mathrm{pen}_{\mathrm{id}}(m)$ décroît avec la complexité de m tant que $K \leq K_{\min}$.
Dès que $K > K_{\min}$, le sur-apprentissage cesse car $\widehat{\mathcal{R}}_n\,(\,\widehat{s}_m\,) + \mathrm{pen}(m) \approx \mathcal{R}_P\,(\,s_m^\star\,) + (K - K_{\min})\,\mathrm{pen}_{\mathrm{id}}(m)$ croît à partir d'un certain niveau de complexité pour S_m.
Tant que $K_{\min} < K < K_{\min} + 1$, $\widehat{m}$ tend à être plus complexe que l'oracle $m^\star$.
Lorsque $K \approx K_{\min} + 1$, $\widehat{m}$ tend à être dans un voisinage de l'oracle $m^\star$, mais l'incertitude sur l'estimation de $\mathrm{pen}_{\mathrm{id}}(m)$ par $\mathrm{pen}(m)$ peut induire la sélection de modèles significativement plus complexes que l'oracle avec une probabilité non-négligeable. Un tel «sur-apprentissage par accident» ne peut être exclu que si le rapport signal sur bruit est suffisamment grand, ou bien en prenant K légèrement plus grande que $K_{\min} + 1$.
Enfin, lorsque $K > K_{\min} + 1$, le choix $\widehat{m}$ est un peu trop conservatif (c'est-à-dire, $\widehat{m}$ tend à être moins complexe que l'oracle $m^\star$), ce qui ne fait augmenter l'excès de risque que d'un facteur multiplicatif majoré par $K - K_{\min}$.

Pour résumer, nous avons pu confirmer le raisonnement fondé sur la capacité à obtenir une inégalité-oracle à partir de (5.6) :

- en-dessous d'un niveau minimal de pénalisation, $\widehat{m}$ sur-apprend ;
- lorsque $\mathrm{pen} \approx \mathrm{pen}_{\mathrm{id}}$, $\widehat{m} \approx m^\star$ en général, mais le risque de sur-apprendre légèrement persiste si l'estimation de $\mathrm{pen}_{\mathrm{id}}(m)$ par $\mathrm{pen}(m)$ est incertaine ;
- enfin, lorsque $\mathrm{pen}(m) > \mathrm{pen}_{\mathrm{id}}(m)$, $\widehat{m}$ est un choix plus conservatif que $m^\star$, ce qui a des conséquence sur le risque moins critiques que le sur-apprentissage, tant que $\mathrm{pen}(m)$ n'est pas beaucoup plus grande que $\mathrm{pen}_{\mathrm{id}}(m)$.

5.2.4. Minimisation du risque structurel. L'idée de la Minimisation du risque structurel **[VC74, Vap82, Vap98]** est d'utiliser comme pénalité un estimateur de

$$\mathcal{B}_P(S_m; D_n) := \sup_{t \in S_m} \left| \widehat{\mathcal{R}}_n\,(t; D_n\,) - \mathcal{R}_P\,(t\,) \right| \geq \mathrm{pen}_{\mathrm{id}}(m) \ .$$

Dans un premier temps, on peut se contenter de pénaliser par une borne supérieure théorique sur $\mathcal{B}_P(S_m; D_n)$, en prenant toutefois soin de ne pas utiliser les valeurs des constantes multiplicatives données par la théorie, mais plutôt d'estimer ces constantes par simulation ou par une

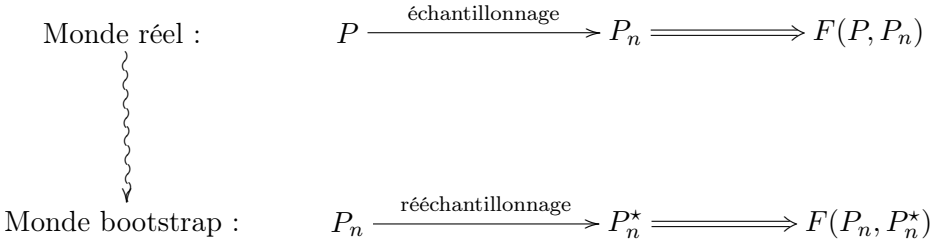

FIG. 5.1. L'heuristique de rééchantillonnage, selon Efron [**Efr79**]. Schéma inspiré de [**Efr03**, Figure 1].

méthode de calibration. Les résultats prouvés au Chapitre 4, conduisent à définir pénalités suivantes :

$$\mathrm{pen}(m) = C\sqrt{\frac{\ln(\mathrm{Card}(S_m))}{n}} \qquad \text{si } S_m \text{ est fini,}$$

$$\mathrm{pen}(m) = C\sqrt{\frac{K}{n}} \qquad \text{si } S_m = S_{\mathrm{inter}}(K) \text{ ou } S_{\mathrm{partition}}(A_1, \ldots, A_K),$$

$$\mathrm{pen}(m) = C\sqrt{\frac{\mathrm{Card}(I)+1}{n}} \qquad \text{si } S_m = S_{\mathrm{lin}}(I) \text{ avec } I \subset \{1, \ldots, n\},$$

$$\mathrm{pen}(m) = C\sqrt{\frac{V_m}{n}} \qquad \text{si } S_m \text{ est une classe de Vapnik-Chervonenkis de dimension } V_m.$$

Ces bornes supérieures sur $\mathcal{B}_P(S_m; D_n)$ étant souvent pessimistes (voir notamment la Remarque 4.13 en Section 4.3.2 à ce sujet), un premier raffinement consiste à estimer directement $\mathcal{B}_P(S_m; D_n)$.

Pour ce faire, on peut utiliser le *rééchantillonnage*, heuristique introduite par Efron [**Efr79, Efr82, ET93**] et dont le principe est le suivant. On désire estimer une quantité $F(P, P_n)$ dépendant des données D_n (via la mesure empirique $P_n = n^{-1}\sum_{i=1}^n \delta_{(X_i,Y_i)}$) et de leur distribution P ; ici, $F(P, P_n) = \mathcal{B}_P(S_m; D_n)$. Pour cela, il est nécessaire d'«estimer» la relation qui existe entre P et P_n, dont on sait seulement que P_n a été obtenue à partir de P par un processus d'échantillonnage i.i.d. L'idée du rééchantillonnage est de construire un «Monde bootstrap», miroir du «Monde réel», où P est remplacée par P_n, et le processus d'échantillonnage par un processus dit de «rééchantillonnage» (voir Figure 5.1). La paire (P, P_n) est donc remplacée par $(P_n, P_n^\star)$, où $P_n^\star$ est la mesure empirique du rééchantillon, et $F(P, P_n)$ est estimée par $F(P_n, P_n^\star)$, L'aléa de rééchantillonnage étant arbitraire, on peut s'en affranchir en estimant $F(P, P_n)$ par $\mathbb{E}\,[F(P_n, P_n^\star) \mid D_n]$, l'espérance ne portant que sur la loi du rééchantillonnage.

Lorsque le rééchantillon est obtenu à partir de D_n en choisissant n fois indépendamment un élément de $\{(X_i, Y_i) \text{ t.q. } 1 \le i \le n\}$ uniformément[2], on parle de «bootstrap». Une autre façon de rééchantillonner (parmi de nombreuses autres), est de choisir indépendamment pour chaque i de conserver (X_i, Y_i) dans le rééchantillon avec probabilité $1/2$.

Autrement dit, on considère $\epsilon_1, \ldots, \epsilon_n$ des variables indépendantes et de même loi satisfaisant $\mathbb{P}(\epsilon_i = 1) = \mathbb{P}(\epsilon_i = -1) = 1/2$, appelées variables de Rademacher, et

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

[2]les (X_i, Y_i) apparaissant plusieurs fois dans le rééchantillon se voyant attribuer un poids correspondant

est (à un facteur de normalisation proche de 1 près) la mesure empirique du rééchantillon

$$D_n^\star := (X_i, Y_i)_{\{\, i \text{ t.q. } \epsilon_i=1 \,\}} \ .$$

Si l'on applique l'idée du rééchantillonnage i.i.d. Rademacher à l'estimation de $\mathcal{B}_P(S_m; D_n)$, on obtient la complexité de Rademacher conditionnelle :

$$R_n^\epsilon(S_m; D_n) := \mathbb{E}\left[\sup_{t\in S} \left| \frac{1}{n} \sum_{i=1}^{n} \left[\epsilon_i \mathbb{1}_{t(X_i)\neq Y_i} \right] \right| \,\middle|\, D_n \right] = \mathbb{E}\left[\sup_{t\in S} \left| \mathcal{R}_{P_n}(t) - \mathcal{R}_{P_n^\star}(t) \right| \,\middle|\, D_n \right] \ .$$

En sus de l'heuristique de rééchantillonnage qui garantit que $R_n^\epsilon(S_m; D_n)$ devrait être proche de $\mathcal{B}_P(S_m; D_n)$ avec grande probabilité, on peut montrer (voir Section 4.6) que la complexité de Rademacher conditionnelle se compare à $\mathcal{B}_P(S_m; D_n)$ en espérance :

$$\mathbb{E}\left[\mathcal{B}_P(S_m; D_n)\right] \leq 2\mathbb{E}\left[R_n^\epsilon(S_m; D_n)\right] \ .$$

De plus, on constate expérimentalement que sauf cas exceptionnel, ce facteur deux est inutile, et que $R_n^\epsilon(S_m; D_n)$ est une bonne éstimation de $\mathcal{B}_P(S_m; D_n)$ [**Loz00**]. On peut donc utiliser $R_n^\epsilon(S_m; D_n)$ comme pénalité dans le cadre de la minimisation du risque structurel.

Notons que d'autres méthodes de rééchantillonnage peuvent être utilisées pour estimer $\mathcal{B}_P(S_m; D_n)$, par exemple le bootstrap, conduisant à une famille de pénalités qui satisfont des inégalités de type[3] oracle [**Fro07**].

5.2.5. Pénalités plus fines. Le principe de minimisation du risque structurel trouve cependant sa limite lorsque l'erreur d'estimation décroît comme $n^{-\alpha}$ avec $\alpha \in]1/2; 1]$, comme évoqué en Section 4.5. En effet, la borne $\mathcal{B}_P(S; D_n)$ sur l'erreur d'estimation est elle-même toujours de l'ordre de $n^{-1/2}$.

C'est pourquoi il est préférable d'estimer directement $\mathrm{pen}_{\mathrm{id}}(m)$ pour construire une pénalité. Ceci peut notamment être fait par les différentes méthodes de rééchantillonnage [**Efr83, Arl08**]. En utilisant le rééchantillonnage i.i.d. Rademacher, on obtient la *pénalité Rademacher*

$$\mathrm{pen}_{\mathrm{Rad}}(m) := \mathbb{E}\left[\mathcal{R}_{P_n}\left(\widehat{s}_m(D_n^\star)\right) - \mathcal{R}_{P_n^\star}\left(\widehat{s}_m(D_n^\star)\right) \,\middle|\, D_n \right] \ .$$

On peut se demander quelle pénalité choisir parmi les différentes possibilités évoquées dans cette section. Schématiquement, les pénalités calculées par rééchantillonnage ont l'avantage d'être plus précises car elles essaient de s'adapter à la distribution des données. À l'inverse, les pénalités fondées sur la dimension de Vapnik-Chervonenkis des modèles sont adaptées au pire des cas, et donc souvent trop pessimistes.

Le gain en précision des différentes pénalités par rééchantillonnage a cependant un prix, qui est le temps de calcul nécessaire pour les évaluer avec une précision suffisante. Pour choisir une pénalité, il faut donc faire un compromis entre performance statistique et complexité algorithmique, le poids relatif de chacun de ces facteurs dépendant fortement du problème de classification considéré.

5.3. Calibration par validation croisée

Pour un problème de calibration général, en l'absence d'information sur le problème de classification considéré, l'idée la plus naturelle est d'utiliser l'une des formes de la validation croisée.

[3]On parle d'inégalité «de type oracle» lorsque le terme additionnel ϵ_λ est susceptible d'être prépondérant dans le membre de droite de (5.1), en fonction de caractéristiques de la distribution de P.

5.3.1. Définitions. Rappelons pour commencer quelques définitions vues au Chapitre 1.

Soit $I^{(e)} \subset \{1,\dots,n\}$ de cardinal $1 \leq n_e \leq n-1$. L'estimateur par *validation* du risque de $\widehat{s}(D_n)$ (Définition 1.13) est défini par

$$\widehat{\mathcal{R}}^{\mathrm{val}}\left(\widehat{s}; D_n; I^{(e)}\right) = \widehat{\mathcal{R}}^{\mathrm{val}}(\widehat{s}; D_n) := \frac{1}{n_v} \sum_{i \in D_n^{(v)}} \mathbb{1}_{\widehat{s}\left(X_i; D_n^{(e)}\right) \neq Y_i} \ , \tag{5.7}$$

où l'on a posé $I^{(v)} = \{1,\dots,n\} \setminus I^{(e)}$, $card(I^{(v)}) = n_v$, $D_n^{(e)} := (X_i, Y_i)_{i \in I^{(e)}}$ est l'*échantillon d'entraînement* et $D_n^{(v)} := (X_i, Y_i)_{i \in I^{(v)}}$ est l'*échantillon de validation*.

Étant donnés un entier $B \geq 1$ et une suite $I_1^{(e)}, \dots, I_B^{(e)}$ (aléatoire ou non) de sous-ensembles stricts de $\{1,\dots,n\}$ non-vides, on définit alors l'estimateur par *validation croisée* du risque de $\widehat{s}(D_n)$ (Définition 1.15) :

$$\widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{s}; D_n; \left(I_j^{(e)}\right)_{1\leq j\leq B}\right) = \widehat{\mathcal{R}}^{\mathrm{vc}}(\widehat{s}; D_n) := \frac{1}{B}\sum_{j=1}^{B} \widehat{\mathcal{R}}^{\mathrm{val}}\left(\widehat{s}; D_n; I_j^{(e)}\right) \ , \tag{5.8}$$

où pour tout j, $I_j^{(v)}$ désigne le complémentaire de $I_j^{(e)}$ et $D_{n,j}^{(e)}$ l'échantillon d'entraînement associé à $I_j^{(e)}$.

Étant donnée une famille $(\widehat{s}_\lambda)_{\lambda\in\Lambda}$ de classifieurs, on «*choisit λ par validation croisée*» en choisissant

$$\widehat{\lambda}^{\mathrm{vc}} = \widehat{\lambda}^{\mathrm{vc}}\left(D_n; \left(I_j^{(e)}\right)_{1\leq j\leq B}\right) \in \arg\min_{\lambda\in\Lambda}\left\{\widehat{\mathcal{R}}^{\mathrm{val}}\left(\widehat{s}; D_n; \left(I_j^{(e)}\right)_{1\leq j\leq B}\right)\right\} \ . \tag{5.9}$$

Considérons dans un premier temps une règle de classification $\widehat{s}_\lambda$ fixée, et cherchons à évaluer la qualité des différentes méthodes de validation croisée pour estimer le risque $\mathcal{R}_P\left(\widehat{s}_\lambda(D_n)\right)$.

5.3.2. Biais. Considérons tout d'abord le *biais* de $\widehat{\mathcal{R}}^{\mathrm{vc}}(\widehat{s}_\lambda; D_n; (I_j^{(e)})_{1\leq j\leq B})$ comme estimateur de $\mathcal{R}_P\left(\widehat{s}_\lambda(D_n)\right)$, c'est-à-dire

$$\mathbb{E}_{D_n\sim P^{\otimes n}}\left[\widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{s}_\lambda; D_n; \left(I_j^{(e)}\right)_{1\leq j\leq B}\right) - \mathcal{R}_P\left(\widehat{s}_\lambda(D_n)\right)\right] \ .$$

Supposons pour cela que pour tout $1 \leq j \leq B$, $\mathrm{Card}(I_j^{(e)}) = n_e$ fixé. Alors,

$$\begin{aligned}
\mathbb{E}\left[\widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{s}_\lambda; D_n; \left(I_j^{(e)}\right)_{1\leq j\leq B}\right)\right] &= \mathbb{E}\left[\widehat{\mathcal{R}}^{\mathrm{val}}\left(\widehat{s}; D_n; I^{(e)}\right)\right] \\
&= \frac{1}{n_v}\sum_{i\in D_n^{(v)}} \mathbb{E}_{D_n^{(e)}}\mathbb{P}_{(X_i,Y_i)}\left(\widehat{s}\left(X_i; D_n^{(e)}\right) \neq Y_i\right) \\
&= \mathbb{E}_{D_n^{(e)}}\left[\mathcal{R}_P\left(\widehat{s}\left(D_n^{(e)}\right)\right)\right] \ ,
\end{aligned} \tag{5.10}$$

en utilisant le fait que l'échantillon d'entraînement $D_n^{(e)}$ est indépendant de l'échantillon de validation. Si l'on note

$$F(n) = \mathbb{E}_{D_n\sim P^{\otimes n}}\left[\mathcal{R}_P\left(\widehat{s}\left(D_n\right)\right)\right] \ ,$$

l'estimateur par validation croisée du risque vaut donc $F(n_e)$ en espérance, tandis qu'il cherche à estimer le risque dont l'espérance est $F(n)$.

Le biais de l'estimateur validation croisée du risque dépend donc uniquement de la taille de l'échantillon d'entraînement n_e. Comme $F(n)$ tend en général à décroître avec n (ce qui n'est

pas toujours vrai, voir Remarque 5.8) et comme $n_e < n$ par construction, la validation croisée surestime légèrement le risque de $\widehat{s}_\lambda(D_n)$, et ceci d'autant plus que n_e est petit.

Si l'on revient au problème de calibration, ce biais a pour conséquence de faire un choix $\widehat{\lambda}$ légèrement plus conservatif que l'oracle $\lambda^\star$, et ceci d'autant plus que n_e est petit. En effet, considérons le cas de la sélection de modèles : l'erreur d'approximation ne dépendant pas de n, estimer $F(n_e)$ au lieu de $F(n)$ revient à surestimer l'erreur d'estimation. Plus précisément, comme l'erreur d'estimation varie souvent en $n^{-\alpha}$ avec $\alpha \in [1/2; 1]$, remplacer n par n_e conduit à surestimer le poids de l'erreur d'estimation d'un facteur $(n/n_e)^\alpha$.

La conséquence est que si $n_e \sim n$, alors, asymptotiquement, le risque du classifieur choisi par validation croisée tend à être de l'ordre de celui de l'oracle. Par contre, si $n_e \sim \mu n$ avec $\mu < 1$, ou si $n_e \ll n$, alors le classifieur choisi par validation croisée ne peut pas satisfaire une inégalité-oracle (5.1) avec une constante C tendant vers 1 lorsque n tend vers l'infini.

Remarque 5.7 (Le biais peut être bénéfique). Lorsque le rapport signal sur bruit est faible, comme nous l'avons noté en Section 5.2.3, il peut être avantageux de surestimer légèrement la composante du risque correspondant à l'erreur d'estimation. Lorsque l'on utilise la validation croisée, ceci se traduit en un avantage éventuel à choisir n_e éloigné de n, dans la mesure où un choix de $\widehat{\lambda}$ plus conservatif permet de se prémunir contre tout risque de sur-apprentissage.

Remarque 5.8 (Règles intelligentes). Une règle de classification $\widehat{s}$ est dite *intelligente* lorsque pour toute distribution P sur $\mathcal{X} \times \mathcal{Y}$, l'espérance de son risque

$$F(n) := \mathbb{E}_{D_n \sim P^{\otimes n}} \left[\mathcal{R}_P \left(\widehat{s}\left(D_n \right) \right) \right]$$

est une fonction décroissante de la taille n de l'échantillon.

Certaines règles idiotes sont «intelligentes», en particulier la règle des n-ppv, qui ignore les X_i et attribue à tout point $x \in \mathcal{X}$ l'étiquette majoritaire parmi les Y_i. Toute règle de classification par partition est intelligente.

En revanche, la règle des 1-ppv n'est pas intelligente. En effet, considérons $\mathcal{X} = [0, 1]$ et P telle que $(X, Y) = (0, 1)$ avec probabilité p, et $(X, Y) = (Z, 0)$ avec probabilité $1 - p$, où Z est une variable aléatoire uniforme sur $[-1000, 1000]$. On a alors $F(1) = 2p(1 - p)$ et

$$\begin{aligned} F(2) &= 2p(1-p)^2 \left(\frac{1}{2} + \frac{\mathbb{E}\,|Z|}{4000} \right) + p^2(1-p) + (1-p)^2 p \\ &= 2p(1-p) \left(\frac{5(1-p)}{8} + \frac{1}{2} \right) \ , \end{aligned}$$

qui est strictement plus grand que $F(1)$ pour tout $p \in (0, 1/5)$. Notons qu'il est conjecturé qu'aucune règle de classification ne peut être à la fois universellement consistante et intelligente. Pour en savoir plus, voir [**DGL96**, Section 6.8].

5.3.3. Variabilité. Comme l'indique la Section 5.3.2 ci-dessus, au premier ordre, seul le biais compte pour que le classifieur choisi par validation croisée soit asymptotiquement optimale.

En revanche, pour une taille d'échantillon n fixée, on ne peut pas toujours considérer que le rapport signal sur bruit est fort. Pour évaluer les performances des différentes méthodes de validation croisée, il faut donc tenir compte de la variabilité de l'estimation du risque, mesurée par la variance

$$V\left(\left(I_j^{(e)} \right)_{1 \le j \le B} \right) := \mathrm{var}_{D_n \sim P^{\otimes n}} \left[\widehat{\mathcal{R}}^{\mathrm{vc}} \left(\widehat{s}_\lambda; D_n; \left(I_j^{(e)} \right)_{1 \le j \le B} \right) \right] \ .$$

S'il est difficile d'évaluer cette quantité en général, nous pouvons tout de même faire les observations suivantes :

(1) $V\left(\left(I_j^{(e)}\right)_{1\leq j\leq B}\right)$ est une fonction décroissante de B : plus on considère d'échantillons d'entraînement distincts, plus l'estimation du risque est précise.

(2) $V\left(\left(I_j^{(e)}\right)_{1\leq j\leq B}\right)$ décroît d'autant plus vite avec B que les quantités $\widehat{\mathcal{R}}^{\mathrm{val}}\left(\widehat{s}; D_n; \left(I_j^{(e)}\right)_{1\leq j\leq B}\right)$ dépendent peu les unes des autres.

(3) Première conséquence de (2), lorsque n_e est trop proche de n, on peut s'attendre à ce que ces dépendances soient fortes et nécessitent de prendre B grand (de l'ordre de n) pour atteindre un niveau de variabilité acceptable. Ceci expliquerait que dans de nombreux cas, l'estimateur leave-one-out est expérimentalement plus variable que l'estimateur «V-fold» pour $V < n$ [**HTF01**, Chapitre 7].

(4) Deuxième conséquence de (2), à n_e et B fixés, on peut penser que $V\left(\left(I_j^{(e)}\right)_{1\leq j\leq B}\right)$ est d'autant plus petit que les échantillons d'entraînement sont choisis de manière «équilibrée».
En particulier, mieux vaut que chaque point apparaisse environ le même nombre de fois dans l'échantillon d'entraînement (heuristique qui sous-tend l'idée du «V-fold»).
Il est également souvent préconisé de choisir les $I_j^{(e)}$ tels que chaque découpage entraînement/validation respecte approximativement la structuration spatiale des données.

5.3.4. Choix d'une méthode de validation croisée. Au final, choisir une méthode de validation croisée, en particulier n_e et B, nécessite de faire un compromis entre la précision de l'estimation du risque et le temps de calcul. En effet, pour augmenter la précision, on doit choisir n_e proche de n et B aussi grand que possible (d'autant plus que n_e est proche de n), tout en prenant garde à des phénomènes d'augmentation forte de la variabilité lorsque n_e est trop proche de n. À l'inverse, le temps de calcul est proportionnel à B, et impose souvent des contraintes telles que $B \leq 10$.

Il est difficile de tirer des conclusions générales pour comparer différentes méthodes de calibration, excepté le fait que la validation simple fournit un estimateur du risque trop variable car elle repose sur un choix arbitraire de découpage entre échantillons d'entraînement et de validation. D'un point de vue pratique, procéder à des simulations dans un cadre adapté au problème que l'on souhaite résoudre peut permettre de trancher cette question.

Notons qu'il n'est pas surprenant qu'il n'existe pas une méthode de calibration clairement meilleure que les autres, au vu du Théorème 2.3. En effet, $D_n \mapsto \widehat{s}_{\widehat{\lambda}(D_n)}(D_n)$ reste une règle de classification, et ne peut donc pas être universellement meilleure que toutes les autres règles de classification.

5.4. Remarques conclusives

5.4.1. Interprétation du modèle sélectionné $\widehat{m}$. Il faut faire attention à ne pas surinterpréter le modèle sélectionné $\widehat{m}(D_n)$, même si une inégalité-oracle (5.1) est satisfaite pour cette procédure de choix de modèles. Prenons pour simplifier l'exposé le cas de la famille des modèles $S_K = S_{\mathrm{inter}}(K)$ de classification par intervalles, pour lequel on voudrait choisir un $\widehat{K} \in \{1, \ldots, n\}$.

D'une part, ce n'est pas parce que le risque de $\widehat{s}_{\widehat{K}}$ est proche de celui de l'estimateur oracle $\widehat{s}_{K^\star}$ que l'on a nécessairement $\widehat{K} = K^\star$ avec grande probabilité, ni même $\left|\widehat{K} - K^\star\right| \leq c$ fixée.

D'autre part, on n'a pas nécessairement $s^\star \in S_{K^\star(D_n)}$ car l'oracle $K^\star(D_n)$ dépend des données. Par exemple, même si $s^\star \in S_{K_0}$ pour un certain $K_0 \leq n$, si le bruit est trop fort ou le nombre

d'observations trop faibles en comparaison de K_0, le nombre $K^\star$ d'intervalles optimal pour la classification de nouvelles données à partir de D_n est en général beaucoup plus petit que K_0.

Enfin, lorsque $s^\star \in S_{K_0}$ pour un $K_0 \ll n$, on pourrait légitimement souhaiter choisir un $\widehat{K}$ tel que

$$\mathbb{P}_{D_n \sim P^{\otimes n}} \left(\widehat{K}(D_n) = K_0 \right) \xrightarrow[n\to\infty]{} 1 \ .$$

Cet objectif, appelé *identification*, est en général distinct de la minimisation du risque de classification et requiert l'emploi de méthodes spécifiques [**Yan05, Yan07**].

5.4.2. Interprétation de l'estimation du risque pour le modèle sélectionné. Lorsque l'on estime $\widehat{\lambda}$ en minimisant un estimateur $\widehat{\mathcal{R}}(\widehat{s}_\lambda; D_n)$ du risque de $\widehat{s}_\lambda(D_n)$, par exemple un estimateur par validation croisée, alors on ne peut pas utiliser $\widehat{\mathcal{R}}(\widehat{s}_{\widehat{\lambda}(D_n)}; D_n)$ pour évaluer le risque du classifieur final $\widehat{s}_{\widehat{\lambda}(D_n)}$. Cette valeur a en effet tendance à *sous-estimer largement le risque de classification sur de nouvelles données.*

Considérons le cas où $\widehat{s}_\lambda = \lambda$ pour $\lambda \in S$ un modèle, et $\widehat{\mathcal{R}}(\widehat{s}_\lambda; D_n) = \widehat{\mathcal{R}}_n(\lambda; D_n)$ qui est un estimateur sans biais du risque de $\widehat{s}_\lambda(D_n)$ (la conclusion serait la même avec l'estimateur par validation croisée du risque). Alors, comme nous l'avons vu en Section 5.2, le risque empirique du minimiseur du risque empirique sous-estime son risque ; c'est pour cette raison qu'il est nécessaire de lui ajouter une pénalité, ou de le tester sur de *nouvelles* données pour obtenir un estimateur non biaisé du risque.

Si l'on veut disposer d'une estimation *a priori* du risque du classifieur $\widehat{s}_{\widehat{\lambda}}$ que l'on va utiliser sur de nouvelles données, par exemple dans le cadre d'une application industrielle, alors on doit procéder de la manière suivante.

Tout d'abord, couper au préalable l'ensemble des observations en deux. La première partie D_n est utilisée pour constuire un classifieur, éventuellement en utilisant une procédure de calibration, et donc en découpant à nouveau D_n en $D_n^{(e)}$ et $D_n^{(v)}$ s'il s'agit de validation croisée. La seconde partie, notée $D_n^{(t)}$ et appelée *échantillon test*, est réservée pour l'évaluation du risque de $\widehat{s}_{\widehat{\lambda}(D_n)}$, via la quantité

$$\mathcal{R}_{P_n^{(t)}} \left(\widehat{s}_{\widehat{\lambda}(D_n)} \right)$$

où $P_n^{(t)}$ est la mesure empirique associée à $D_n^{(t)}$. Cette quantité estime sans biais le risque $\mathcal{R}_P \left(\widehat{s}_{\widehat{\lambda}(D_n)} \right)$ dès lors que D_n et $D_n^{(t)}$ sont indépendants.

5.4.3. Pas de classification complètement automatique. On peut trouver décevant que la conclusion de l'étude des différentes méthodes de calibration soit qu'il reste à les calibrer, via le choix d'une constante devant la pénalité, des paramètres B et n_e de la validation croisée, ou encore de la manière de choisir les échantillons d'entraînement («V-fold» ou apprentissage-test répété ?).

Cependant, comme nous l'avons déjà indiqué, l'existence de No Free Lunch Theorems indique que l'on n'a pas d'espoir de s'affranchir de tout choix de paramètre par l'utilisateur final. Plus encore, *une règle de classification qui serait présentée comme complètement automatique doit immédiatement être considérée comme suspecte*, car cela signifie qu'au moins un paramètre a été fixé arbitrairement quelque part à l'intérieur de la méthode. Si ce choix n'a pas été fait sur la base d'une étude sérieuse et pour le problème précis de classification que l'on souhaite résoudre, alors il y a fort à parier qu'une telle règle de classification pourrait être grandement améliorée.

L'intérêt d'une méthode de calibration reste toutefois dans le fait que la performance dépend en général faiblement du choix des paramètres. Par exemple, toutes les valeurs de V pour le «V-fold» ont *a priori* des performances du même ordre de grandeur, ce qui est loin d'être le cas

pour les k-ppv lorsque k varie de 1 à n. De plus, l'influence des différents paramètres n_e, B, ou la constante devant une pénalité, peut s'interpréter plus simplement que celle d'un paramètre λ quelconque, notamment lorsque des règles de classification très différentes sont comparées entre elles.

5.4.4. Pénalisation ou validation croisée ? Dans le cas du problème de sélection de modèles, pénalisation et validation croisée sont possibles. Quelle méthode doit-on privilégier ?

Nous avons déjà recensé plusieurs avantages de la pénalisation en Section 5.2. On peut y ajouter la possibilité d'utiliser simplement une information *a priori* sur le problème considéré (par exemple, la forme de pénalité à utiliser), ainsi que la faible complexité algorithmique lorsque la forme de la pénalité est donnée par avance.

À l'inverse, les méthodes de validation croisée ne reposent que sur l'hypothèse d'indépendance des données entre elles (ce que font implicitement les différentes pénalités que nous avons définies). Elles sont donc *a priori* plus robustes, au prix d'une complexité algorithmique supérieure.

Un compromis entre ces deux approches pourrait être l'utilisation des pénalités par rééchantillonnages, qui ne font pas plus d'hypothèse que la validation croisée tout en possédant la flexibilité et l'interprétabilité des méthodes de pénalisation. Leur défaut reste cependant dans la faiblesse des résultats théoriques les concernant.

Par contre, le calcul de l'espérance de l'estimateur par validation croisée du risque fait en Section 5.3.2 est valable en toute généralité, ce qui permet d'accorder une grande confiance aux méthodes de validation croisée.

Bibliographie

[Arl08] Sylvain Arlot. Model selection by resampling penalization, March 2008. oai :hal.archives-ouvertes.fr :hal-00262478_v1.

[BFOS84] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and regression trees*. Wadsworth Statistics/Probability Series. Wadsworth Advanced Books and Software, Belmont, CA, 1984.

[DGL96] 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.

[DL95] L. Devroye and G. Lugosi. Lower bounds in pattern recognition. *Pattern recognition*, 28 :1011–1018, 1995.

[Efr79] Bradley Efron. Bootstrap methods : another look at the jackknife. *Ann. Statist.*, 7(1) :1–26, 1979.

[Efr82] Bradley Efron. *The Jackknife, the Bootstrap and Other Resampling Plans*, volume 38 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, Pa., 1982.

[Efr83] Bradley Efron. Estimating the error rate of a prediction rule : improvement on cross-validation. *J. Amer. Statist. Assoc.*, 78(382) :316–331, 1983.

[Efr03] Bradley Efron. Second thoughts on the bootstrap. *Statist. Sci.*, 18(2) :135–140, 2003. Silver anniversary of the bootstrap.

[ET93] Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, New York, 1993.

[Fro07] Magalie Fromont. Model selection by bootstrap penalization for classification. *Mach. Learn.*, 66(2–3) :165–207, 2007.

[HTF01] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer-Verlag, New York, 2001. Data mining, inference, and prediction.

[KMNR97] Michael Kearns, Yishay Mansour, Andrew Y. Ng, and Dana Ron. An experimental and theoretical comparison of model selection methods. *Mach. Learn.*, 7 :27–50, 1997.

[Loz00] Fernando Lozano. Model selection using rademacher penalization. In *Proceedings of the 2nd ICSC Symp. on Neural Computation (NC2000). Berlin, Germany*. ICSC Academic Press, 2000.

[Mas07] Pascal Massart. *Concentration Inequalities and Model Selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[MN06] Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *Ann. Statist.*, 34(5) :2326–2366, 2006.

[MT99] Enno Mammen and Alexandre B. Tsybakov. Smooth discrimination analysis. *Ann. Statist.*, 27(6) :1808–1829, 1999.

[Sau72] N. Sauer. On the density of families of sets. *J. Combinatorial Theory Ser. A*, 13 :145–147, 1972.

[STC00] John Shawe-Taylor and Nello Cristianini. *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.

[Sto77] Charles J. Stone. Consistent nonparametric regression. *Ann. Statist.*, 5(4) :595–645, 1977. With discussion and a reply by the author.

[Vap82] Vladimir Vapnik. *Estimation of dependences based on empirical data*. Springer Series in Statistics. Springer-Verlag, New York, 1982. Translated from the Russian by Samuel Kotz.

[Vap98] Vladimir N. Vapnik. *Statistical learning theory*. Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons Inc., New York, 1998. A Wiley-Interscience Publication.

[VC71] Vladimir N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2) :264–280, 1971.

[VC74] Vladimir N. Vapnik and Alexey Ya. Chervonenkis. *Teoriya Raspoznavaniya Obrazov. Statisticheskie Problemy Obucheniya*. Izdat. "Nauka", Moscow, 1974. Theory of Pattern Recognition (In Russian).

[Yan05] Yuhong Yang. Can the strengths of AIC and BIC be shared ? A conflict between model indentification and regression estimation. *Biometrika*, 92(4) :937–950, 2005.

[Yan07] Yuhong Yang. Consistency of cross validation for comparing regression procedures. *Ann. Statist.*, 35(6) :2450–2473, 2007.