



Fondamentaux de l'apprentissage statistique

Sylvain Arlot

► To cite this version:

| Sylvain Arlot. Fondamentaux de l'apprentissage statistique. Myriam Maumy-Bertrand; Gilbert Saporta; Christine Thomas-Agnan. Apprentissage statistique et données massives, Editions Technip, 2018, 9782710811824. hal-01485506

HAL Id: hal-01485506

<https://hal.archives-ouvertes.fr/hal-01485506>

Submitted on 8 Mar 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Fondamentaux de l'apprentissage statistique

Sylvain Arlot

8 Mars 2017

Table des matières

1	Prévision	3
1.1	Formalisation du problème	4
1.2	Prévision idéale	8
1.3	Règles d'apprentissage	10
1.4	Consistance	11
2	Régression et classification	12
2.1	Régression	12
2.2	Classification supervisée	17
2.3	Liens entre régression et classification	24
3	Minimisation du risque empirique	27
3.1	Principe	27
3.2	Exemples en régression	29
3.3	Exemples en classification	31
3.4	Erreur d'approximation, erreur d'estimation	33
3.5	Majoration générale de l'erreur d'estimation	35
3.6	Cas d'un modèle fini	37
3.7	Cas d'un modèle quelconque	40
3.7.1	Symétrisation	40
3.7.2	Classification 0–1 : entropie combinatoire empirique	41
3.7.3	Classe de Vapnik-Chervonenkis	42
3.7.4	Récapitulatif	45
3.7.5	Cas des règles par partition	47
3.7.6	Extensions	47
3.8	Classification zéro-erreur	48
3.9	Choix d'un modèle	49

4 Coûts convexes en classification	58
4.1 Pseudo-classificateurs et Φ -risque	59
4.2 Exemples	60
4.3 Classification idéale pour le Φ -risque	64
4.4 Calibration pour la classification 0–1	66
4.5 Lien entre excès de Φ -risque et excès de risque 0–1	68
5 Moyenne locale	72
5.1 Définition	72
5.2 Consistance : théorème de Stone	74
5.3 Règles par partition	80
5.4 Plus proches voisins	83
5.5 Noyau	86
6 On n'a rien sans rien	89
6.1 À taille d'échantillon fixée	89
6.2 Classification avec \mathcal{X} fini	92
6.3 À loi fixée	93
7 Conclusion : enjeux de l'apprentissage	93
7.1 Minimax	94
7.2 Autres approches	97
8 Annexe : outils probabilistes	98
8.1 Sommes de variables indépendantes bornées	98
8.2 Maximum de variables sous-gaussiennes	101
8.3 Inégalité de Mc Diarmid	102
8.4 Inégalité de symétrisation	102
8.5 Espérance de l'inverse d'une variable binomiale	104
9 Annexe : exercices	105
9.1 Régression et classification	105
9.2 Minimisation du risque empirique	105
9.2.1 Erreur d'approximation et erreur d'estimation	105
9.2.2 Majoration générale de l'erreur d'estimation	108
9.2.3 Cas d'un modèle fini	108
9.2.4 Cas d'un modèle quelconque	109
9.2.5 Choix d'un modèle	112
9.3 Coûts convexes en classification	112
9.4 Moyenne locale	114
9.5 On n'a rien sans rien	114
9.6 Conclusion	115
9.7 Outils probabilistes	116

Ce texte présente les bases de l'apprentissage statistique supervisé, sous un angle mathématique. On décrit d'abord le problème général de prévision (section 1), puis les deux exemples fondamentaux que sont la régression et la classification (ou discrimination) binaire (section 2). Ensuite, on étudie deux grandes familles de règles d'apprentissage. La minimisation du risque empirique (section 3) amène naturellement à la question de la convexification du risque de classification (section 4); puis, les règles par moyenne locale permettent d'énoncer un résultat de consistance universelle (section 5). Enfin, on identifie les limites de l'apprentissage (section 6), pour mieux en dégager les enjeux (section 7). Les outils probabilistes utilisés sont détaillés en section 8. La section 9 rassemble des exercices qui complètent les sections précédentes.

Tout au long de ce texte, l'exemple des règles par partition sert de fil rouge. Il est introduit en section 2 par les exemples 1 et 3. Le lecteur est invité, de manière générale, à réfléchir à ce que chaque résultat énoncé permet de dire sur les règles par partition.

Nous avons pris le parti de donner peu de références bibliographiques. En particulier, la plupart des références historiques (ayant proposé pour la première fois telle ou telle méthode) sont absentes. Le lecteur intéressé est invité à consulter, pour la régression, le livre de Györfi *et al.* (2002), et pour la classification, le livre de Devroye *et al.* (1996) et l'article de survol de Boucheron *et al.* (2005).

Un certain nombre de remarques sont appelées « parenthèses ». Ceci indique qu'on peut les sauter en première lecture; certaines s'adressent au lecteur expérimenté et font appel à des connaissances extérieures ou à des résultats énoncés ultérieurement dans ce texte.

Notations Si x_1, \dots, x_n sont n éléments d'un ensemble \mathcal{X} , on note $x_{1\dots n}$ le n -uplet (x_1, \dots, x_n) . De même, on note $X_{1\dots n}$ pour (X_1, \dots, X_n) .

1 Prévision

Le problème fondamental de l'apprentissage statistique (supervisé) est le problème de prévision¹. Étant donné un ensemble d'observations (X_1, \dots, X_n) , chacune munie d'une étiquette (Y_1, \dots, Y_n) , il s'agit de « prévoir » quelle étiquette Y_{n+1} doit être associée à une nouvelle observation X_{n+1} . De plus, cet apprentissage à partir d'exemples doit pouvoir être réalisé automatiquement, par une machine².

Pour fixer les idées, on peut penser aux problèmes suivants :

- *prévision du taux de pollution* : on dispose de données météorologiques (température à midi, vitesse du vent, etc.), que l'on rassemble dans un vecteur $X_i \in \mathbb{R}^p$, et l'on cherche à prévoir le taux de pollution à l'ozone $Y_i \in \mathbb{R}$ (figure 1). Cornillon et Matzner-Løber (2011) se servent notamment de cet exemple pour illustrer des méthodes de régression linéaire.

1. On dit souvent aussi prédiction pour prévision ; le terme anglais est « prediction ».

2. On parle souvent d'*apprentissage machine*, traduction française de l'anglais « machine learning », même si ce terme recouvre un domaine plus large que l'apprentissage statistique proprement dit.

- reconnaissance de caractères manuscrits : les X_i sont des images (chacune représentant un caractère manuscrit) et les étiquettes Y_i indiquent le caractère qui y est écrit ; voir par exemple les célèbres données MNIST (LeCun *et al.*, 1998).
- annotation d’images : les X_i sont des images et les Y_i des attributs associés : présence ou non d’un visage, d’un objet donné, liste des personnes visibles, actions représentées, etc. La figure 2 en donne une illustration ; de nombreux jeux de données classiques sont disponibles pour ce problèmes, par exemple ImageNet. Il s’agit d’un des nombreux problèmes d’apprentissage considérés par le domaine de la vision artificielle.
- détection de courriers électroniques indésirables : chaque X_i est le contenu d’un courrier électronique et Y_i indique s’il s’agit ou non d’un courrier indésirable (pour l’utilisateur).

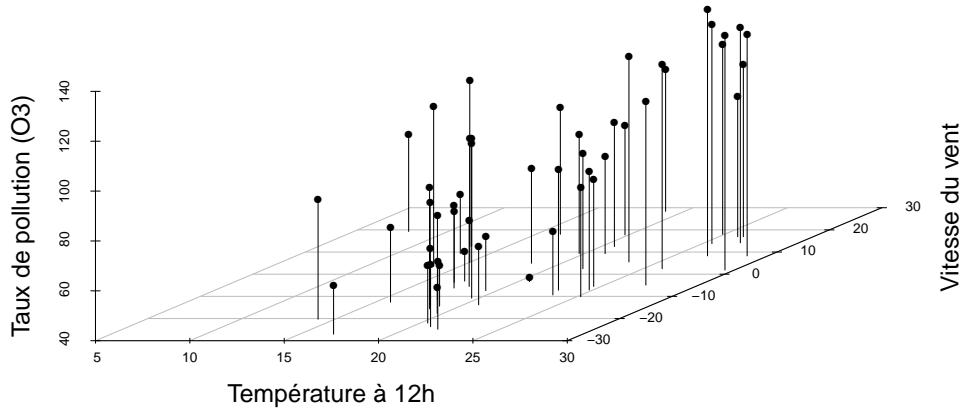


FIGURE 1 – *Taux de pollution à l'ozone en fonction de données météorologiques : un échantillon. Données issues de Cornillon et Matzner-Løber (2011). Il s’agit d’un problème de régression (voir la section 2.1).*

L’objectif de la théorie statistique de l’apprentissage est de dégager des principes théoriques communs à tous ces problèmes.

1.1 Formalisation du problème

On peut formaliser mathématiquement le problème de prévision de la manière suivante.



FIGURE 2 – Illustration du problème de reconnaissance d’« objets » sur une photo (qui est un problème de classification multiétiquette, voir la parenthèse 27 en section 2.2). À gauche : quelques exemples pour trois catégories (papillon, cormoran, tour Eiffel). À droite : une image à étiquetter (on voudrait reconnaître qu’il s’y trouve un cormoran et la tour Eiffel, mais pas de papillon). Note : les trois catégories représentées ici sont présentes dans la base de données Caltech 256 (Griffin et al., 2007), dont les photos reproduites ici ne sont pas extraites.

Données On dispose d’un échantillon $D_n = (X_i, Y_i)_{1 \leq i \leq n}$ (les données) où, pour tout $i \in \{1, \dots, n\}$, $X_i \in \mathcal{X}$ est une³ variable explicative (ou covariable ; le terme anglais usuel est « feature ») et $Y_i \in \mathcal{Y}$ est une variable d’intérêt (ou étiquette ; « label » en anglais). Dans ce texte, on suppose toujours que \mathcal{X} et \mathcal{Y} sont des ensembles mesurables et que $(X, Y), (X_1, Y_1), (X_2, Y_2), \dots$ désigne une suite de variables aléatoires indépendantes et de même loi P . On note P_X la loi de X .

Parenthèse 1 (Modélisation aléatoire) Pourquoi modélise-t-on les observations (X_i, Y_i) par des variables aléatoires, indépendantes et de même loi ?

L’aléa des X_i correspond au fait que l’on échantillonne « au hasard » (uniformément ou pas) parmi une certaine population. Par exemple, pour la reconnaissance de chiffres manuscrits, X_i peut être une image choisie uniformément parmi l’ensemble des images représentant un chiffre écrit à la main par un fran-

3. Souvent, X_i est en fait la collection de plusieurs variables explicatives, que l’on concatène en un vecteur de \mathbb{R}^p . Ceci ne change donc rien formellement : il suffit de prendre $\mathcal{X} = \mathbb{R}^p$.

cophone / un anglophone américain / un individu utilisant les chiffres arabes. Le choix peut se limiter aux chiffres écrits sur une enveloppe (si l'objectif est le tri automatique du courrier), ce qui peut changer la fréquence des différents caractères possibles.

Sachant X_i , la variable Y_i peut être aléatoire pour plusieurs raisons. On peut avoir des erreurs lors de l'acquisition des données (un chiffre « 8 » étiqueté à tort « 6 »). On peut aussi manquer d'information quand on dispose uniquement de X_i . Par exemple, un « 7 » écrit à l'américaine peut ressembler beaucoup à un « 1 » écrit à la française. Et un caractère mal écrit peut simplement être très difficile à lire (il faut deviner ce à quoi pensait celui qui l'a écrit !).

Supposer que (X, Y) et les (X_i, Y_i) sont de même loi P (stationnarité) nécessite qu'elles proviennent d'une même population. Par exemple, pour la reconnaissance de chiffres manuscrits, si l'on apprend avec des chiffres écrits par francophone, on ne teste pas sur des chiffres écrits par un anglophone (qui écrit différemment certains chiffres). Lorsque les données d'apprentissage et de test ne suivent pas la même loi, il faut en tenir compte ; on parle alors d'apprentissage par transfert, apprentissage transductif, adaptation de domaine, ou encore de biais de sélection d'échantillon (« covariate-shift » en anglais). La non-stationnarité peut aussi être remise en question, par exemple quand les données proviennent d'une série temporelle ; lorsque c'est le cas, il faut en tenir compte.

Enfin, l'hypothèse d'indépendance des observations est parfois peu réaliste (notamment si l'on observe une série temporelle, comme dans l'exemple du taux de pollution à l'ozone). Différentes stratégies existent pour s'adapter à des phénomènes de dépendance, nous ne les évoquons pas ici. L'indépendance est aussi rompue dans les cadres de l'apprentissage actif (« active learning » en anglais) et de l'apprentissage par renforcement (« reinforcement learning » en anglais), où les données arrivent l'une après l'autre, et où X_{i+1} est choisi par le statisticien en fonction des observations passées $(X_1, Y_1), \dots, (X_i, Y_i)$; nous n'en parlons pas non plus.

Sorties Une solution du problème de prévision est une application mesurable $f : \mathcal{X} \rightarrow \mathcal{Y}$, que l'on appelle *prédicteur*. On note $\mathcal{F} = \mathcal{F}(\mathcal{X}, \mathcal{Y})$ l'ensemble des prédicteurs. L'idée est que si l'on a une nouvelle observation $X_{n+1} \in \mathcal{X}$ d'une variable explicative, alors $f(X_{n+1}) \in \mathcal{Y}$ est notre candidat pour « prévoir » la valeur de l'étiquette Y_{n+1} (non observée).

Mesure de qualité On se donne une *fonction de coût* (aussi appelée perte) $c : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$, supposée mesurable. En pratique, on choisit c adaptée au problème posé, telle que $c(y, y')$ est d'autant plus petit que y et y' sont similaires. Dans ce texte, on suppose (pour simplifier) que pour tout $y, y' \in \mathcal{Y}$, $c(y, y') \geq 0$ et $c(y, y) = 0$.

L'objectif est alors de fournir un prédicteur $f \in \mathcal{F}$ tel que $c(f(X_{n+1}), Y_{n+1})$ est petit « en moyenne ». On définit le *risque d'un prédicteur* $f \in \mathcal{F}$ (aussi appelé *erreur de généralisation* ou *erreur de prévision*) par :

$$\mathcal{R}(f) = \mathcal{R}_P(f) = \mathbb{E}[c(f(X), Y)]. \quad (1)$$

Ceci définit, pour toute mesure de probabilité P sur $\mathcal{X} \times \mathcal{Y}$, une fonction (mesurable) $\mathcal{R}_P : \mathcal{F} \rightarrow \mathbb{R} \cup \{+\infty\}$ que l'on cherche à minimiser, sans connaître P (mais seulement un échantillon D_n).

Remarque 2 (Dépendance vis-à-vis de c) Le risque $\mathcal{R}(f)$ dépend de P et de c . Lorsque le contexte le nécessite, on marque cette dépendance explicitement dans la notation, par exemple en l'écrivant $\mathcal{R}_P^c(f)$.

Parenthèse 3 (Risque conditionnel et risque moyen) Le risque $\mathcal{R}(f)$ est parfois appelé *risque conditionnel*. En effet, il est sous-entendu dans la définition (1) que l'espérance est prise *uniquement par rapport à l'aléa de (X, Y)* . Lorsque f est fonction de D_n , f est aléatoire et la définition (1) doit se comprendre ainsi :

$$\mathcal{R}_P(f) = \mathbb{E}\left[c(f(X), Y) \mid D_n\right].$$

Toutefois, cette terminologie peut prêter à confusion car le risque conditionnel désigne souvent

$$\mathbb{E}\left[c(f(X), Y) \mid D_n, X\right],$$

où le conditionnement est pris par rapport à D_n et X . C'est ainsi que l'on parle de « Φ -risque conditionnel » en section 4.3. Dans la suite, on utilise uniquement le terme « risque » pour désigner $\mathcal{R}_P(f)$. Par opposition au risque (conditionnel ou pas), on définit en section 1.3 la notion de risque moyen, qui correspond à prendre l'espérance vis-à-vis de (X, Y) et de D_n .

Parenthèse 4 (Risque infini) L'espérance définissant le risque $\mathcal{R}(f)$ dans (1) n'est pas forcément finie, car il se peut que $c(f(X), Y)$ ne soit pas intégrable. Dans ce cas, le coût c ayant été supposé à valeurs positives ou nulles, la définition (1) reste valide et l'on a $\mathcal{R}(f) = +\infty$.

Parenthèse 5 (Fonctions de risque plus générales) Il n'est pas nécessaire de disposer d'une fonction de coût c pour définir le risque \mathcal{R}_P par l'équation (1). On pourrait très bien considérer le problème de prévision en supposant seulement que pour toute loi de probabilité P sur $\mathcal{X} \times \mathcal{Y}$, une fonction mesurable $\mathcal{R}_P : \mathcal{F} \rightarrow \mathbb{R} \cup \{+\infty\}$ est donnée, et utiliser cette fonction \mathcal{R}_P comme mesure de risque. Par exemple, en régression (voir section 2.1), on peut s'intéresser à l'erreur d'estimation L^p de la fonction de régression $\eta = \eta_P$, avec $p \in [1, +\infty]$. On pose alors $\mathcal{R}_P(f) = \|f - \eta_P\|_{L^p}$, qui ne peut pas s'écrire sous la forme (1). Le carré de l'erreur d'estimation L^2 est une exception, comme expliqué en section 2.1.

Problème de prévision On peut désormais reformuler le problème de prévision comme suit : il s'agit de trouver, à l'aide des données D_n uniquement, un prédicteur $f \in \mathcal{F}$ tel que son risque $\mathcal{R}_P(f)$ est minimal.

Parenthèse 6 (Théorie de la décision) On décrit ici le problème de prévision avec un formalisme inspiré de la théorie de la décision, telle que décrite par Bickel et Doksum (2001, section 1.3). On peut noter que Bickel et Doksum utilisent une terminologie légèrement différente, ce qui leur permet d'englober dans un même cadre la prévision et plusieurs autres problèmes statistiques (estimation, test, « ranking », etc.) : \mathcal{F} est appelé « espace des actions » \mathcal{A} , le risque $\mathcal{R}_P(f)$ — ou plutôt l'excès de risque $\ell(f^*, \cdot)$, tel que défini en section 1.2 — est appelé « fonction de perte » (entre la loi P et l'action f).

1.2 Prévision idéale

Plaçons-nous pour commencer dans une situation idéale : la loi P est connue. Alors, le problème de prévision se résume à un problème d'optimisation : trouver $f \in \mathcal{F}$ tel que $\mathcal{R}_P(f)$ est minimal, où $\mathcal{R}_P : \mathcal{F} \rightarrow \mathbb{R} \cup \{+\infty\}$ est une fonction connue.

On définit alors le *risque de Bayes* :

$$\mathcal{R}^* = \mathcal{R}_P^* := \inf_{f \in \mathcal{F}} \mathcal{R}_P(f). \quad (2)$$

C'est la plus petite valeur de risque envisageable pour un prédicteur. Puisque l'on a supposé c à valeurs positives ou nulles, le risque de Bayes est positif ou nul également. Dans la plupart des cas, une prévision parfaite est impossible, même en connaissant P , et le risque de Bayes est strictement positif (ce qu'illustrent les propositions 1 et 2 en section 2).

Un prédicteur optimal (pour P) est alors un prédicteur $f^* \in \mathcal{F}$ dont le risque est égal au risque de Bayes :

$$f^* \in \operatorname{argmin}_{f \in \mathcal{F}} \{\mathcal{R}_P(f)\}.$$

Un tel prédicteur (lorsqu'il existe) est appelé *prédicteur de Bayes*, ou encore règle de Bayes (en anglais, « Bayes rule ») ou fonction cible.

Remarque 7 (Existence et unicité) Un prédicteur de Bayes n'existe pas toujours. Et s'il existe, il n'est pas forcément unique (voir par exemple les sections 2.1 et 2.2).

À la place du risque, on mesure souvent la qualité d'un prédicteur $f \in \mathcal{F}$ par son *excès de risque* (aussi appelé *risque relatif*), défini par :

$$\ell(f^*, f) = \mathcal{R}_P(f) - \mathcal{R}_P^* \geq 0.$$

L'avantage de la notion d'excès de risque (par rapport au risque) est qu'elle définit une mesure de qualité d'un prédicteur dont le niveau minimal est toujours nul, et correspond à la prévision idéale. On peut donc chercher⁴ un prédicteur f dont l'excès de risque est proche de zéro (au moins lorsque le nombre d'observations disponibles tend vers l'infini).

4. Dans certains cas, avec succès, voir la section 5.

Remarque 8 (Dépendance vis-à-vis de P et c) L'excès de risque d'un prédicteur $f \in \mathcal{F}$ dépend de P (et de c). On peut faire apparaître cette dépendance en le notant $\ell_P(f_P^*, f)$ ou $\ell_P^c(f^*, f)$ si nécessaire.

Parenthèse 9 (Notation $\ell(f^*, f)$) L'excès de risque est bien défini même lorsqu'il n'existe pas de prédicteur de Bayes f^* , ou lorsque celui-ci n'est pas unique. La notation usuelle $\ell(f^*, f)$ relève donc (un peu) de l'abus de notation. Elle fait référence à la situation où f^* existe et est unique, si bien que $\ell(f^*, f)$ mesure une « distance » entre f^* et f . La lettre ℓ utilisée pour désigner l'excès de risque provient du fait que l'excès de risque définit une fonction de perte (« loss », en anglais) sur l'ensemble \mathcal{F} des décisions possibles pour un algorithme de prévision, et que l'on cherche à minimiser sur \mathcal{F} cette fonction de perte (voir la parenthèse 6 en section 1.1).

Parenthèse 10 (Prédicteur de Bayes) Les termes « prédicteur de Bayes » ou « règle de Bayes » trouvent leur origine dans la manière bayésienne de comparer des estimateurs (Bickel et Doksum, 2001, sections 1.3.3 et 3.2). On peut en effet voir le problème de prévision comme un problème d'estimation bayésien. Pour cela, on note $Y = \theta$ et \mathbb{P}_θ la loi conditionnelle de X sachant θ (supposée connue). Étant donné X , on cherche à « estimer » θ (qui est aléatoire, ce qui correspond bien au point de vue bayésien). Le « risque » d'estimation d'un prédicteur f (conditionnellement à la valeur de θ , il s'agit donc d'une quantité différente du risque défini en section 1.1) est alors défini par :

$$\mathbb{E}_{X \sim \mathbb{P}_\theta} [c(\theta, f(X))] = \mathbb{E}[c(\theta, f(X)) | \theta].$$

Le « risque bayésien » de f est défini en prenant une espérance sur θ (avec la loi *a priori* π) :

$$\mathbb{E}_{\theta \sim \pi} \mathbb{E}_{X \sim \mathbb{P}_\theta} [c(\theta, f(X))].$$

Si $\pi = P_Y$ la deuxième marginale de P , le risque bayésien de f coïncide avec le risque $\mathcal{R}_P(f)$ défini en section 1.1. Par conséquent, une règle de décision f qui minimise le « risque bayésien » (appelée « règle de Bayes ») est un prédicteur qui minimise le risque \mathcal{R}_P , et réciproquement. Notons au passage que le risque bayésien — qui est défini pour une règle de décision $f \in \mathcal{F}$ quelconque — est une notion différente du risque de Bayes défini par (2) — qui correspond au risque (bayésien) d'une règle *optimale*.

Bickel et Doksum (2001, section 3.2) montrent comment calculer explicitement la règle de Bayes dans de nombreux cadres. En régression avec le coût quadratique, Bickel et Doksum (2001, section 1.4) démontrent ainsi un résultat similaire à la proposition 1 en section 2.1. En classification avec le coût 0–1, les résultats de Bickel et Doksum (2001, exemple 3.2.2) sont à comparer à la proposition 2 en section 2.2.

1.3 Règles d'apprentissage

Revenons au problème de prévision tel qu'il se pose en pratique : P est inconnue, on dispose seulement d'un échantillon D_n de taille finie $n \in \mathbb{N} \setminus \{0\}$ et l'on doit proposer un prédicteur $f \in \mathcal{F}$. On appelle *règle d'apprentissage*⁵ une solution à ce problème, que l'on définit formellement comme une application mesurable

$$\hat{f}: \bigcup_{n \geq 1} (\mathcal{X} \times \mathcal{Y})^n \rightarrow \mathcal{F}.$$

La section 2 donne des exemples de règles d'apprentissage.

Remarque 11 La taille d'échantillon n n'est pas fixée *a priori* : \hat{f} peut prendre en entrée un échantillon de taille quelconque.

Remarque 12 (Notation $\hat{f}(D_n; x)$) Si \hat{f} est une règle d'apprentissage et D_n un échantillon, $\hat{f}(D_n) \in \mathcal{F}$ est un prédicteur (aléatoire), donc une fonction mesurable $\mathcal{X} \rightarrow \mathcal{Y}$. Pour tout $x \in \mathcal{X}$, on note $\hat{f}(D_n; x)$ la valeur en x de $\hat{f}(D_n)$. Par abus de notation, on écrit souvent \hat{f} au lieu de $\hat{f}(D_n)$, et $\hat{f}(x)$ au lieu de $\hat{f}(D_n; x)$.

Parenthèse 13 (Règle d'apprentissage et estimateur) En statistique, le terme d'estimateur désigne souvent ce que nous appelons une règle d'apprentissage. Les deux notions sont en effet très proches. La raison pour laquelle on préfère utiliser ici « règle d'apprentissage » est que l'objectif de prévision n'est pas d'*estimer* un paramètre de la loi P , mais plutôt de *faire aussi bien que f^** en termes de *risque*. Il n'y a donc pas vraiment de paramètre à estimer en général (sauf dans certains cas, comme en régression avec le coût quadratique, voir la proposition 1 en section 2.1).

On rappelle que le *risque* de $\hat{f}(D_n)$, défini en section 1.1, s'écrit

$$\mathcal{R}_P(\hat{f}(D_n)) = \mathbb{E}[c(\hat{f}(D_n; X), Y) | D_n]$$

où $(X, Y) \sim P$ est indépendant de D_n . C'est une variable aléatoire (l'erreur moyenne commise par $\hat{f}(D_n)$ sur une *nouvelle observation*).

On définit également le *risque moyen* de \hat{f} (avec n observations indépendantes et de même loi P) par

$$\mathbb{E}[\mathcal{R}_P(\hat{f}(D_n))] = \mathbb{E}[c(\hat{f}(D_n; X), Y)]$$

5. Il n'y a pas, à notre connaissance, de terminologie unanimement choisie pour ce que l'on appelle ici « règle d'apprentissage » (traduction de l'anglais learning rule, utilisé par Devroye *et al.*, 1996). Les termes « méthode d'apprentissage » et « algorithme d'apprentissage » peuvent également être utilisés. On pourrait également penser à « estimateur », une notion très proche, mais la parenthèse 13 explique pourquoi on évite de confondre les deux notions. Signalons enfin que le terme « règle » est parfois utilisé en un sens plus restreint en intelligence artificielle (par exemple, dans l'expression « rule-based machine learning ») ; ici, on n'impose aucune forme particulière à \hat{f} , hormis sa mesurabilité.

où l'on rappelle que D_n est un échantillon de n variables indépendantes et de même loi P .

Parenthèse 14 (Théorie de la décision, suite) À la suite de la parenthèse 6 en section 1.1, on peut signaler que Bickel et Doksum (2001, section 1.3) utilisent le terme « procédure de décision » pour ce que l'on appelle ici une règle d'apprentissage, et nomment simplement « risque » ce que l'on appelle ici le risque moyen.

1.4 Consistance

Pour qu'une règle d'apprentissage soit considérée comme bonne, un critère naturel est de demander à ce qu'elle fasse aussi bien que si P était connue (c'est-à-dire, atteindre le risque de Bayes) lorsque le nombre d'observations disponibles tend vers l'infini. On parle de *consistance*, une notion que l'on peut formaliser de différentes manières.

Définition 1 (Consistance) Soit \hat{f} une règle d'apprentissage, P une loi de probabilité sur $\mathcal{X} \times \mathcal{Y}$ et, pour tout entier $n \geq 1$, D_n un échantillon de n variables indépendantes et de même loi P . On dit que :

(i) \hat{f} est *faiblement consistante* pour P lorsque

$$\mathbb{E}\left[\mathcal{R}_P(\hat{f}(D_n))\right] \xrightarrow[n \rightarrow +\infty]{\quad} \mathcal{R}_P^*.$$

(ii) \hat{f} est *fortement consistante* pour P lorsque

$$\mathcal{R}_P(\hat{f}(D_n)) \xrightarrow[n \rightarrow +\infty]{\text{p.s.}} \mathcal{R}_P^*.$$

Remarque 15 (Excès de risque et consistance) De manière équivalente, \hat{f} est consistante pour P lorsque l'excès de risque de $\hat{f}(D_n)$ tend vers zéro. On parle de consistance faible lorsque ceci a lieu pour la convergence L^1 (en espérance). On parle de consistance forte lorsque ceci a lieu pour la convergence presque sûre.

Remarque 16 (Lien entre consistance faible et forte) Si le coût c définissant le risque est borné, alors le risque \mathcal{R}_P est borné et la consistance faible équivaut à la convergence en probabilité de $\mathcal{R}_P(\hat{f}(D_n))$ vers \mathcal{R}_P^* . En particulier, lorsque c est borné, la consistance forte de \hat{f} entraîne la consistance faible de \hat{f} (d'où la terminologie faible/forte).

Parenthèse 17 (Autres définitions de la consistance) Deux définitions de la consistance, légèrement différentes de la définition 1, sont proposées par Vapnik (2000, chapitre 2) dans le cas d'une règle \hat{f}_S minimisant le risque empirique sur $S \subset \mathcal{F}$ (comme défini en section 3) ; voir à ce sujet la parenthèse 68.

La loi P est inconnue en pratique. Une bonne manière de s'assurer que \hat{f} est consistante pour la loi P qui a généré les données est donc de démontrer que f est (faiblement ou fortement) consistante *pour toute loi de probabilité* P sur $\mathcal{X} \times \mathcal{Y}$. On dit alors que P est *universellement (faiblement ou fortement) consistante*.

Remarque 18 (Consistance et vitesse de convergence) Savoir que \hat{f} est consistante garantit que le risque de $\hat{f}(D_n)$ tend vers le risque de Bayes quand n tend vers l'infini. En revanche, ceci ne dit rien sur la *vitesse de convergence* du risque (souvent appelée vitesse d'apprentissage), qui peut être très lente. En particulier, si \hat{f} est universellement consistante, la vitesse de convergence de $\mathcal{R}_P(\hat{f}(D_n))$ peut dépendre de P (et donc être inconnue de l'utilisateur). La section 6 démontre qu'il est impossible, sauf dans une situation « simple », d'avoir une garantie sur cette vitesse de convergence sans disposer d'informations *a priori* sur P .

2 Régression et classification

Le problème général de prévision inclut deux cadres fondamentaux : la régression et la classification binaire supervisée.

2.1 Régression

Lorsque la variable d'intérêt Y est continue et univariée (c'est-à-dire, $\mathcal{Y} = \mathbb{R}$ ou un intervalle de \mathbb{R}), on parle de *régression*. Le lecteur intéressé par plus de détails sur la régression (non-paramétrique) est invité à consulter le livre de Györfi *et al.* (2002).

Remarque 19 (Régression multivariée) Lorsque $\mathcal{Y} = \mathbb{R}^d$, on parle de régression multivariée (ou régression multiple, ou régression multi-tâches). Ce problème peut s'aborder de manière similaire : il s'agit de résoudre (simultanément) d problèmes de régression univariée. Des techniques spécifiques permettent alors d'exploiter les similarités entre ces d problèmes, voir par exemple le livre de Giraud (2014, chapitre 6) et la thèse de Solnon (2013). Dans cette section, on ne traite que du cas univarié, que l'on nomme « régression » sans plus de précision.

Sous l'hypothèse que Y admet une espérance, on définit la *fonction de régression* $\eta : \mathcal{X} \rightarrow \mathbb{R}$ par

$$\eta(X) := \mathbb{E}[Y | X] \tag{3}$$

presque sûrement. On pose alors $\varepsilon := Y - \eta(X)$, ce qui permet d'écrire Y comme la somme d'une fonction de X et d'un bruit :

$$Y = \eta(X) + \varepsilon \quad \text{avec} \quad \mathbb{E}[\varepsilon | X] = 0 \quad \text{presque sûrement.}$$

Parenthèse 20 (Unicité de la fonction de régression) L'équation (3) définit en fait η à un ensemble de mesure nulle pour P_X près. Pour ce qui nous

intéresse dans la suite, ce niveau d'indétermination de η ne pose pas problème.
On peut considérer que l'on choisit (une fois pour toutes) une fonction $\eta \in \mathcal{F}$ qui vérifie (3) et que l'on nomme « la » fonction de régression.

Règles par partition en régression

Un exemple de règle de régression, que nous considérons à plusieurs reprises dans ce texte, est donné par les *règles par partition*, aussi appelées règles par histogrammes (en prévision en général) ou régressogrammes (pour la régression uniquement).

Exemple 1 (Règle de régression par partition) Soit \mathcal{A} une partition⁶ mesurable de \mathcal{X} , finie ou dénombrable. Pour tout $x \in \mathcal{X}$, on note $\mathcal{A}(x)$ l'unique élément de la partition \mathcal{A} qui contient x . On pose alors, pour tout entier $n \geq 1$ et tout $x \in \mathcal{X}$ et $(x_i, y_i)_{1 \leq i \leq n} \in (\mathcal{X} \times \mathcal{Y})^n$:

$$\hat{f}_{\mathcal{A}}^{\text{p-r}}((x_i, y_i)_{1 \leq i \leq n}; x) := \frac{1}{N_{\mathcal{A}(x)}(x_{1 \dots n})} \sum_{i=1}^n y_i \mathbb{1}_{x_i \in \mathcal{A}(x)},$$

en notant

$$N_A(x_{1 \dots n}) = \text{Card}\{i \in \{1, \dots, n\} / x_i \in A\}$$

pour tout $A \in \mathcal{A}$ et avec la convention $\frac{0}{0} = 0$ si $N_{\mathcal{A}(x)}(x_{1 \dots n}) = 0$. Ceci définit la *règle de régression par partition* associée à la partition \mathcal{A} , notée $\hat{f}_{\mathcal{A}}^{\text{p-r}}$.

Un des intérêts des règles par partition est qu'elles peuvent être définies sans aucune structure sur \mathcal{X} (par exemple, pas besoin de munir d'une distance) : il suffit de disposer d'une partition de \mathcal{X} . Bien sûr, lorsque \mathcal{X} possède une structure, un choix judicieux pour \mathcal{A} doit en tenir compte, comme le montre la section 5.

Remarque 21 (Partition variant avec n) La partition \mathcal{A} peut aussi dépendre de la taille n de l'échantillon dans l'exemple 1. On obtient alors $\hat{f}_{(\mathcal{A}_n)}^{\text{p-r}}$ la règle de régression par partition associée à la suite de partitions $(\mathcal{A}_n)_{n \geq 1}$.

Remarque 22 (Convention pour les cellules vides) Lorsque la cellule $\mathcal{A}(x)$ ne contient aucun point de l'échantillon, la valeur de $\hat{f}_{\mathcal{A}}^{\text{p-r}}(D_n; x)$ est fixée à 0. On aurait pu prendre une autre valeur fixe dans \mathcal{Y} , éventuellement fonction des observations (par exemple $\bar{Y} := n^{-1} \sum_{i=1}^n Y_i$). En pratique, on pourrait aussi utiliser les valeurs de $\hat{f}_{\mathcal{A}}^{\text{p-r}}$ dans les cellules « voisines » de $\mathcal{A}(x)$, à condition de pouvoir donner un sens à « voisines ». Du point de vue théorique, en général, ce choix n'a quasiment pas d'incidence lorsque n est assez grand, hormis quelques détails techniques.

6. Une partition de \mathcal{X} est une collection de sous-ensembles *disjoints* et non-vides de \mathcal{X} , dont la réunion est l'ensemble \mathcal{X} tout entier. On dit qu'une partition est *mesurable* lorsque tous ses éléments sont des parties mesurables de \mathcal{X} . Un élément d'une partition est parfois appelé case ou cellule.

Parenthèse 23 (Ensemble \mathcal{Y} plus général) L'exemple 1 s'étend sans difficulté au cas où \mathcal{Y} est un espace vectoriel (par exemple, \mathbb{R}^d) ou bien une partie convexe d'un espace vectoriel.

Lorsque $\mathcal{X} = \mathbb{R}^p$, un choix naturel est de considérer une partition cubique (ou régulière), que l'on définit ainsi.

Exemple 2 (Règle de régression par partition cubique) Si $\mathcal{X} = \mathbb{R}^p$ pour un entier $p \geq 1$, pour tout $h > 0$, on définit la *partition cubique de pas h* par :

$$\mathcal{A}^{\text{cub}}(h) := \left(\prod_{i=1}^p [hk_i, h(k_i + 1)) \right)_{(k_1, \dots, k_p) \in \mathbb{Z}^p}.$$

La règle par partition associée à $\mathcal{A}^{\text{cub}}(h)$, notée $\hat{f}_h^{\text{cub-r}}$, est appelée *règle par partition cubique de pas h* .

Remarque 24 (Pas h variant avec n) Le pas h peut varier avec la taille n de l'échantillon dans l'exemple 2. On obtient alors $\hat{f}_{(h_n)}^{\text{cub-r}}$ la règle de régression par partition cubique de pas $(h_n)_{n \geq 1}$.

Remarque 25 (Partition cubique sur une partie de \mathbb{R}^p) On peut étendre l'exemple 2 au cas où \mathcal{X} est un sous-ensemble quelconque de \mathbb{R}^p , en remplaçant $\mathcal{A}^{\text{cub}}(h)$ par

$$\mathcal{A}_{\mathcal{X}}^{\text{cub}}(h) := \left(\mathcal{X} \cap \prod_{i=1}^p [hk_i, h(k_i + 1)) \right)_{(k_1, \dots, k_p) \in \mathbb{Z}^p},$$

dont on retire les intersections vides.

Une autre forme classique de règle par partition en régression est celle des arbres de décision.

Coût quadratique

La fonction de coût la plus classique en régression est le *coût quadratique*, aussi appelé *coût des moindres carrés* :

$$\forall y, y' \in \mathbb{R}, \quad c(y, y') = (y - y')^2.$$

On suppose alors que

$$\mathbb{E}[Y^2] < +\infty.$$

Le risque associé

$$\forall f \in \mathcal{F}, \quad \mathcal{R}_P(f) := \mathbb{E}\left[\left(f(X) - Y\right)^2\right]$$

est appelé *risque quadratique* ou *risque des moindres carrés*. La proposition suivante montre qu'il est minimal lorsque $f = \eta$ la fonction de régression, et que l'excès de risque correspondant est égal au carré de l'erreur d'estimation $L^2(P_X)$ de η .

Proposition 1 On suppose $\mathcal{Y} = \mathbb{R}$, $\mathbb{E}[(Y - \eta(X))^2] < +\infty$ et l'on considère le coût quadratique $c : (y, y') \mapsto (y - y')^2$. On a alors :

- (i) La fonction de régression η est un prédicteur de Bayes.
- (ii) Un prédicteur f est un prédicteur de Bayes si et seulement si

$$f(X) = \eta(X) \quad \text{presque sûrement.}$$

- (iii) Le risque de Bayes vaut :

$$\mathcal{R}_P^* = \mathbb{E}\left[\left(Y - \eta(X)\right)^2\right] = \mathbb{E}\left[\text{var}(Y | X)\right] = \mathbb{E}[\varepsilon^2].$$

- (iv) L'excès de risque de tout prédicteur $f \in \mathcal{F}$ s'écrit :

$$\ell(f^*, f) = \mathbb{E}\left[\left(f(X) - \eta(X)\right)^2\right] = \|f - \eta\|_{L^2(P_X)}^2.$$

Démonstration Pour tout $f \in \mathcal{F}$, on écrit :

$$\mathcal{R}_P(f) = \mathbb{E}\left[\left(f(X) - \eta(X) - \varepsilon\right)^2\right] = \mathbb{E}\left[\left(f(X) - \eta(X)\right)^2\right] + \mathbb{E}[\varepsilon^2].$$

Par conséquent,

$$\mathcal{R}_P(f) \geq \mathbb{E}[\varepsilon^2] = \mathcal{R}_P(\eta)$$

ce qui démontre (i), (iii) et (iv). Enfin, $\mathcal{R}_P(f) = \mathcal{R}_P(\eta)$ équivaut à

$$\mathbb{E}\left[\left(f(X) - \eta(X)\right)^2\right] = 0,$$

c'est-à-dire, $f(X) = \eta(X)$ presque sûrement, ce qui démontre (ii). \square

Parenthèse 26 (Sur la proposition 1) La proposition 1 est cohérente avec le fait que η est définie à un ensemble de mesure nulle près pour P_X (voir la parenthèse 20) : son énoncé reste valable quel que soit le « représentant » choisi pour η . Dans la suite, en régression, on utilise indifféremment les notations f^* et η .

L'hypothèse $\mathbb{E}[Y^2] < +\infty$ n'est pas nécessaire. Il suffit en fait de supposer que $\mathbb{E}[(Y - \eta(X))^2] < +\infty$ (hypothèse minimale, puisque sans elle le risque de tout prédicteur est infini). La proposition 1 reste valable sous cette hypothèse plus faible, étant sous-entendu au point (iv) que si $f - \eta \notin L^2(P_X)$, alors le risque (et donc l'excès de risque) de f est infini.

Autres fonctions de coût

D'autres fonctions de coût peuvent être considérées sur $\mathcal{Y} = \mathbb{R}$. Par exemple, on peut utiliser le *coût valeur absolue* :

$$\forall y, y' \in \mathbb{R}, \quad c(y, y') = |y - y'|,$$

pour lequel on peut démontrer un équivalent de la proposition 1 (voir les exercices 1 et 2).

Plus généralement, pour tout $p > 0$, on peut considérer le *coût* L^p :

$$\forall y, y' \in \mathbb{R}, \quad c(y, y') = |y - y'|^p.$$

Lorsque $p = 1$, on retrouve le coût valeur absolue ; lorsque $p = 2$, on retrouve le coût quadratique.

En fonction de l'objectif visé, il peut être utile d'utiliser d'autres fonctions de coût en régression. Les exemples suivants (de même que les coûts L^p) sont tous de la forme

$$c(y, y') = \psi(y - y')$$

pour une fonction $\psi : \mathbb{R} \rightarrow [0, +\infty[$ qui atteint sa valeur minimale (zéro) en zéro.

- *Coût « intervalle de prévision ».* On pose $\psi : u \in \mathbb{R} \mapsto \mathbf{1}_{|u| > \alpha}$, pour un paramètre $\alpha > 0$ à choisir. Ceci correspond à l'objectif d'effectuer une prévision de Y correcte à α près.
- *Coût quadratique seuillé*⁷. On pose $\psi : u \in \mathbb{R} \mapsto \min\{u^2, \alpha^2\}$ avec $\alpha > 0$. Par rapport au coût quadratique, les erreurs de plus de α (en valeur absolue) sont toutes pénalisées comme une erreur de α exactement.
- *Fonction de Huber.* On pose :

$$\psi : u \in \mathbb{R} \mapsto \begin{cases} u^2 & \text{si } |u| \leq \alpha \\ 2\alpha|u| - \alpha^2 & \text{sinon.} \end{cases}$$

Cette fonction de coût est quadratique pour les erreurs inférieures à α (en valeur absolue), et linéaire ensuite.

- *Fonction de Tukey-biweight.* On pose :

$$\psi : u \in \mathbb{R} \mapsto \begin{cases} 1 - \left(1 - \frac{u^2}{\alpha^2}\right)^2 & \text{si } |u| \leq \alpha \\ 1 & \text{sinon.} \end{cases}$$

Par rapport au coût quadratique, les coûts valeur absolue, quadratique seuillé, Huber et Tukey-Biweight ont l'avantage d'être peu sensibles à la présence de valeurs aberrantes (dans le cadre de procédures de minimisation du risque empirique, comme défini en section 3). Les fonctions de Huber et de Tukey-Biweight sont ainsi très classiques en statistique robuste (Droesbeke *et al.*, 2015, section 6.3.1).

Toutes les fonctions ψ ci-dessus sont symétriques : $\psi(-u) = \psi(u)$ pour tout $u \in \mathbb{R}$. Lorsque surestimer ou sous-estimer la variable d'intérêt y n'ont pas le même impact, il est utile de considérer des fonctions ψ asymétriques. Par exemple, si l'on souhaite avant tout éviter de sur-estimer la variable d'intérêt, on peut modifier le coût « intervalle de prévision » en posant $c(y, y') = \mathbf{1}_{y-y' > \alpha}$, avec $\alpha > 0$.

7. Le coût quadratique seuillé est appelé en anglais « truncated quadratic loss », mais il ne faut pas le confondre avec le coût quadratique tronqué défini en section 4.2.

2.2 Classification supervisée

Lorsque la variable d'intérêt Y ne prend qu'un nombre fini de valeurs (c'est-à-dire, lorsque \mathcal{Y} est fini), on parle de *classification supervisée*, aussi appelée *discrimination* ou *classement*⁸. On parle alors souvent de *classifieur* plutôt que de prédicteur, de *classifieur de Bayes* plutôt que de prédicteur de Bayes, et de *règle de classification* plutôt que de règle d'apprentissage. Le lecteur intéressé par plus de détails sur la classification est invité à consulter le livre de Devroye *et al.* (1996) et l'article de survol de Boucheron *et al.* (2005).

Dans ce cours, on se focalise sur le cas où seulement deux valeurs sont possibles pour Y . On parle alors de *classification binaire supervisée* ou *discrimination binaire*; on nomme ici ce cadre « *classification* » par abus de langage⁹. Le cas général (dit *classification multiclasse*) présente des difficultés supplémentaires, mais les enjeux fondamentaux du problème sont les mêmes que dans le cas binaire.

Parenthèse 27 (Classification multiclasse) On peut toujours formuler le problème de classification multiclasse comme une succession de problèmes à deux classes. Voici deux idées naïves et deux méthodes couramment utilisées pour cela. Posons $\mathcal{Y} = \{0, \dots, M - 1\}$ par convention. On peut résoudre les problèmes binaires « $Y = i$ » vs. « $Y < i$ » pour tout $i \in \{1, \dots, M - 1\}$ et attribuer une étiquette $f(x)$ à $x \in \mathcal{X}$ en se demandant successivement s'il faut attribuer à x l'étiquette $M - 1$, puis (si la réponse est non) l'étiquette $M - 2$, et ainsi de suite jusqu'à l'étiquette 1 (le choix 0 étant le choix par défaut). Une meilleure idée est de procéder d'une manière similaire en plaçant les éléments de \mathcal{Y} dans une hiérarchie binaire, et en considérant successivement les problèmes binaires « l'étiquette appartient-elle au groupe 1 ou au groupe 2 ? », et ainsi de suite en continuant à l'intérieur du groupe prévu. Ces deux premières stratégies souffrent cependant très souvent de l'accumulation d'erreurs de classification. En pratique, la méthode « un contre tous » (en anglais, « one-versus-all », OvA, ou « one-versus-rest », OvR) est plus efficace (et largement utilisée). Pour chaque classe $i \in \mathcal{Y}$, on construit un pseudo-classifieur (voir la section 4) pour le problème « $Y = i$ » vs. « $Y \neq i$ ». On choisit *in fine* l'étiquette i correspondant au pseudo-classifieur de plus grande valeur. On peut aussi utiliser la stratégie « chacun contre chacun » (en anglais, « round robin classification » ou « one-vs-one » ; Fürnkranz, 2002) : pour tout $i, j \in \mathcal{Y}, i \neq j$, on apprend un classifieur discriminant entre $Y = i$ et $Y = j$. Puis, pour étiquetter $x \in \mathcal{X}$, on lui applique tous les classificateurs « i contre j » et l'on retient l'étiquette qui a gagné le plus grand nombre de matches binaires. Notons enfin que si M est grand ou si les effectifs des classes sont déséquilibrés, une approche spécifique (et directe) est indispensable (Guermeur, 2007).

8. Le terme « *classement* », pour désigner le problème de classification supervisée, peut prêter à confusion car c'est aussi une traduction naturelle de « *ranking* » (ou « *learning to rank* », « *ordonnancement* »), un problème où l'on cherche à ordonner une liste d'objets (décrits par des attributs). On choisit ici de n'utiliser le terme *classement* pour désigner aucun de ces deux problèmes.

9. Attention ! Le terme anglais « *classification* » correspond à ce que l'on nomme en français « *classification supervisée* » ou *discrimination*. En français, « *classification* » seul désigne souvent le problème de classification *non supervisée* (« *clustering* » en anglais), contrairement à la convention prise dans ce texte.

Un problème relié important est la classification multiétiquette (« multi-label classification » en anglais), où l'étiquette $y \in \mathcal{Y}$ est un sous-ensemble d'un ensemble \mathcal{E} d'étiquettes possibles (c'est le cas du problème d'annotation d'images, qui est évoqué au début de la section 1 et illustré par la figure 2). On peut le voir comme un problème de classification multiclasse, où \mathcal{Y} est l'ensemble des parties de \mathcal{E} ; c'est en général une mauvaise idée. Ou bien on peut le voir comme un ensemble de $\text{Card}(\mathcal{E})$ problèmes de classification binaire et les résoudre indépendamment. Mais il est souvent judicieux d'utiliser une stratégie spécifiquement conçue pour le problème multiétiquette (Zhang et Zhou, 2014), prenant en compte les dépendances entre étiquettes (par exemple, à l'aide de modèles graphiques); on est alors dans un cas particulier d'apprentissage « multi-tâches ».

Parenthèse 28 (Prévision structurée) On peut généraliser la classification supervisée au-delà du problème multiclasse, avec ce que l'on appelle la prévision structurée (« structured prediction » ou « structured output learning » en anglais). Il s'agit du problème de prévision avec un ensemble \mathcal{Y} qui n'est pas un espace vectoriel. On obtient la classification multiclasse lorsque \mathcal{Y} est fini, mais il est également classique de considérer des ensembles \mathcal{Y} infinis et « structurés » : un ensemble de graphes, d'histogrammes, de chaînes de caractères, etc. Des algorithmes efficaces existent pour traiter ce problème lorsque \mathcal{Y} est un ensemble structuré quelconque (Nowozin *et al.*, 2014), notamment à l'aide de variantes des SVM.

On suppose donc désormais que $\mathcal{Y} = \{0, 1\}$. Par exemple, l'étiquette Y peut encoder la présence (1) ou l'absence (0) d'un papillon au sein d'une image représentée par X .

Remarque 29 (Étiquettes ± 1) Une autre convention pour \mathcal{Y} (moins naturelle dans ce texte, sauf en section 4) est souvent utilisée en classification binaire : poser $\mathcal{Y} = \{-1, 1\}$. Ceci est bien sûr totalement équivalent à la convention $\mathcal{Y} = \{0, 1\}$, via la transformation $y \in \{0, 1\} \mapsto 2y - 1 \in \{-1, 1\}$ ou sa réciproque $z \in \{-1, 1\} \mapsto (z + 1)/2 = \mathbb{1}_{z \geq 0} \in \{0, 1\}$. Il faut seulement faire attention à ne pas utiliser une formule valable avec l'une de ces conventions quand on a choisi l'autre.

Remarque 30 (Classifieur et parties de \mathcal{X}) En classification binaire, on a une bijection entre l'ensemble des classificateurs \mathcal{F} et l'ensemble des parties mesurables de \mathcal{X} . En effet, à tout classificateur $f \in \mathcal{F}$, on associe :

$$A_f := f^{-1}(\{1\}) = \{x \in \mathcal{X} / f(x) = 1\}.$$

Réiproquement, à toute partie mesurable A de \mathcal{X} , on fait correspondre le classificateur $\mathbb{1}_A$. On désigne ainsi souvent un classificateur f via la partie A_f de \mathcal{X} auquel il correspond.

Comme en section 2.1, on définit la *fonction de régression* $\eta : \mathcal{X} \rightarrow \mathbb{R}$ par

$$\eta(X) := \mathbb{E}[Y | X] = \mathbb{P}(Y = 1 | X) \tag{4}$$

presque sûrement. Celle-ci est toujours définie (puisque Y est bornée), à un ensemble de mesure nulle pour P_X près (voir la parenthèse 20).

Règles par partition en classification

De même qu'en régression, le premier exemple de règle de classification que nous définissons est celui des *règles par partition*.

Exemple 3 (Règle de classification par partition) Comme à l'exemple 1 on considère \mathcal{A} une partition mesurable de \mathcal{X} , finie ou dénombrable. Pour tout $x \in \mathcal{X}$, on note $\mathcal{A}(x)$ l'unique élément de la partition \mathcal{A} qui contient x . On pose alors, pour tout entier $n \geq 1$, tout $x \in \mathcal{X}$ et $(x_i, y_i)_{1 \leq i \leq n} \in (\mathcal{X} \times \mathcal{Y})^n$:

$$\hat{f}_{\mathcal{A}}^{p-c}((x_i, y_i)_{1 \leq i \leq n}; x) := \begin{cases} 1 & \text{si } \text{Card}\{i \in \{1, \dots, n\} / y_i = 1 \text{ et } x_i \in \mathcal{A}(x)\} \\ & > \text{Card}\{i \in \{1, \dots, n\} / y_i = 0 \text{ et } x_i \in \mathcal{A}(x)\} \\ 0 & \text{sinon.} \end{cases}$$

Ceci définit la *règle de classification par partition* associée à la partition \mathcal{A} , notée $\hat{f}_{\mathcal{A}}^{p-c}$.

Remarque 31 Les remarques faites en régression, à la suite de l'exemple 1, s'appliquent ici. Dans l'exemple 3, la partition \mathcal{A} peut varier avec la taille de l'échantillon. Lorsque $\mathcal{X} \subset \mathbb{R}^p$ et $\mathcal{A} = \mathcal{A}_{\mathcal{X}}^{\text{cub}}(h)$ comme à l'exemple 2, on obtient la règle de classification par partition cubique de pas h . Les arbres de décision sont un autre exemple classique de règle de classification par partition. Enfin, quand $\mathcal{A}(x)$ ne contient aucun point X_i de l'échantillon, on peut noter que la convention ici est de poser $\hat{f}_{\mathcal{A}}^{p-c}(D_n; x) = 0$ par défaut. D'autres choix sont possibles, mais celui-ci est naturel, puisque la classe 0 est très souvent la classe « par défaut » (par exemple, lorsqu'il s'agit de déterminer si un objet est présent dans une image, on choisit la classe 0 pour encoder l'absence de l'objet).

Parenthèse 32 (Lien régression/classification) Les règles d'apprentissage des exemples 1 et 3 sont liées. Une partition \mathcal{A} étant donnée, pour tout échantillon D_n et tout $x \in \mathcal{X}$:

$$\hat{f}_{\mathcal{A}}^{p-c}(D_n; x) = \mathbf{1}_{\hat{f}_{\mathcal{A}}^{p-r}(D_n; x) > 1/2}.$$

On dit que $\hat{f}_{\mathcal{A}}^{p-c}$ est la règle par plug-in associée à $\hat{f}_{\mathcal{A}}^{p-r}$, une notion étudiée en détail en section 2.3.

Parenthèse 33 (Multiclasse) L'exemple 3 s'étend directement au cas d'un ensemble \mathcal{Y} fini (classification multiclasse). Dans chaque case $A \in \mathcal{A}$, on procède à un vote majoritaire parmi les étiquettes y_i des observations $x_i \in A$.

Coût 0–1

Une fonction de coût naturelle en classification est le coût 0–1, défini par :

$$\forall y, y' \in \mathcal{Y}, \quad c(y, y') = \mathbb{1}_{y \neq y'}.$$

Puisque l'on a choisi $\mathcal{Y} = \{0, 1\}$, on peut aussi écrire le coût 0–1 comme le coût quadratique : $\mathbb{1}_{y \neq y'} = (y - y')^2$. Le risque associé, appelé *risque 0–1* et défini par

$$\forall f \in \mathcal{F}, \quad \mathcal{R}_P(f) = \mathbb{E}[\mathbb{1}_{f(X) \neq Y}] = \mathbb{P}(f(X) \neq Y),$$

mesure la probabilité que f commette une erreur de classification sur de nouvelles observations. La proposition suivante précise ce qu'est un classifieur idéal avec le coût 0–1 et donne une formule simple pour l'excès de risque.

Proposition 2 *On suppose $\mathcal{Y} = \{0, 1\}$ et l'on considère le coût 0–1 défini par c : $(y, y') \mapsto \mathbb{1}_{y \neq y'}$.*

- (i) *Le classifieur défini par $f^*(x) = \mathbb{1}_{\eta(x) > 1/2}$ pour tout $x \in \mathcal{X}$ est un classifieur de Bayes.*
- (ii) *Un classifieur f est un classifieur de Bayes si et seulement si*

$$f(X) = \mathbb{1}_{\eta(X) > 1/2} \quad \text{presque sûrement,}$$

sauf éventuellement sur l'événement $\{\eta(X) = 1/2\}$.

- (iii) *Le risque de Bayes vaut :*

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

- (iv) *L'excès de risque de tout classifieur $f \in \mathcal{F}$ s'écrit :*

$$\ell(f^*, f) = \mathbb{E}\left[|2\eta(X) - 1|\mathbb{1}_{f^*(X) \neq f(X)}\right]. \quad (5)$$

Démonstration Pour tout $f \in \mathcal{F}$, on définit le risque 0–1 conditionnellement à X :

$$\begin{aligned} \mathbb{P}(f(X) \neq Y | X) &= \mathbb{1}_{f(X)=0}\mathbb{P}(Y=1 | X) + \mathbb{1}_{f(X)=1}\mathbb{P}(Y=0 | X) \\ &= \mathbb{1}_{f(X)=0}\eta(X) + \mathbb{1}_{f(X)=1}(1 - \eta(X)) \\ &\geq \min\{\eta(X), 1 - \eta(X)\}. \end{aligned} \quad (6)$$

Remarquons que cette dernière inégalité est une égalité quand $f = f^*$. En intégrant, on obtient que pour tout classifieur f ,

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

Ceci démontre (i) et (iii). De plus, (6) entraîne que :

$$\begin{aligned} &\mathbb{P}(f(X) \neq Y | X) - \mathbb{P}(f^*(X) \neq Y | X) \\ &= \mathbb{1}_{f(X) \neq f^*(X)}\left[\max\{\eta(X), 1 - \eta(X)\} - \min\{\eta(X), 1 - \eta(X)\}\right] \\ &= \mathbb{1}_{f(X) \neq f^*(X)}|2\eta(X) - 1|, \end{aligned}$$

d'où (iv) en passant à l'espérance. On en déduit que $\ell(f^*, f) = 0$ si et seulement si

$$\mathbb{1}_{f(X) \neq f^*(X)} |2\eta(X) - 1| = 0$$

presque sûrement, ce qui démontre (ii). \square

Remarque 34 (Classification zéro-erreur) Le risque de Bayes est nul si et seulement si $\eta(X) \in \{0, 1\}$ presque sûrement, c'est-à-dire, lorsque l'étiquette Y est une fonction déterministe de X . On parle alors de cas « zéro-erreur » (car il est idéalement possible de réaliser une classification sans erreur), situation où l'on peut obtenir de meilleures vitesses d'apprentissage que dans le cas général (voir la section 3.8).

Parenthèse 35 (Unicité de η) La proposition 2 est cohérente avec le fait que η est définie à un ensemble de mesure nulle près pour P_X (voir la parenthèse 20 en section 2.1) : son énoncé reste valable quel que soit le représentant choisi pour η .

La formule (5) montre que l'excès de risque de $f \in \mathcal{F}$ s'interprète comme une « distance » entre f et f^* (la mesure pour P_X de la différence symétrique entre $A_f = \{x \in \mathcal{X} / f(x) = 1\}$ et $A_{f^*} = \{x \in \mathcal{X} / f^*(x) = 1\}$), pondérée par $|2\eta(X) - 1|$ qui mesure la difficulté locale du problème de classification en X .

Dans le cas zéro-erreur, on a exactement :

$$\forall f \in \mathcal{F}, \quad \ell(f^*, f) = \mathbb{P}(f(X) \neq f^*(X)) = P_X(A_f \Delta A_{f^*}).$$

Dans le cas général, on ajoute une pondération $|2\eta(X) - 1|$ qui est d'autant plus grande que $\eta(X)$ est loin de $1/2$ (et donc que le problème de classification est « facile »). Schématiquement, il y a trois situations :

1. Lorsque $\eta(X)$ est proche de $1/2$, le problème de classification est très difficile (X n'apporte quasiment aucune information sur Y), mais il l'est pour tous les classifieurs (même pour f^* qui « connaît » P). Dès lors, avoir un excès de risque petit est « facile ».
2. Lorsque $\eta(X)$ est éloigné de $1/2$, le problème de classification est facile (on est proche du cas zéro-erreur), et toute erreur d'« estimation » de f^* se paye cher dans l'excès de risque. Mais ce type d'erreur se produit rarement si l'on a suffisamment de données, puisque le niveau de bruit est faible.
3. Lorsque $\eta(X)$ est à une distance intermédiaire de $1/2$, on est dans la situation la plus difficile vis-à-vis de l'excès de risque. Le problème de classification est « faisable » sans être évident. C'est là qu'on voit si une règle de classification fonctionne bien. Les bornes inférieures en pire cas reposent souvent sur l'analyse de ce type de situation. L'ordre de grandeur typique d'une telle valeur « intermédiaire » de $|\eta(X) - 1/2|$ dépend du problème. Par exemple, si \mathcal{X} est fini, avec un échantillon de taille $n \gg \text{Card}(\mathcal{X})$, cet ordre de grandeur est $n^{-1/2}$ (voir les propositions 17 et 18 en section 7).

Coût asymétrique

On peut considérer d'autres mesures d'erreur en classification. On définit ainsi, pour tout $w = (w_0, w_1) \in [0, +\infty[^2$ tel que $w_0 + w_1 > 0$, le *coût asymétrique* :

$$c_w : (y, y') \in \{0, 1\}^2 \mapsto w_{y'} \mathbb{1}_{y \neq y'}.$$

Le risque associé s'écrit, pour tout $f \in \mathcal{F}$:

$$\begin{aligned} \mathcal{R}_P^w(f) &= \mathbb{E}[w_Y \mathbb{1}_{f(X) \neq Y}] \\ &= w_1 \mathbb{P}(f(X) = 0 \text{ et } Y = 1) + w_0 \mathbb{P}(f(X) = 1 \text{ et } Y = 0). \end{aligned}$$

Une telle mesure d'erreur se justifie pleinement lorsque les deux types d'erreurs de classification (étiquetter 0 au lieu de 1, étiquetter 1 au lieu de 0) ont un impact différent en pratique. Par exemple, pour la détection de spams, mieux vaut laisser passer un spam à tort que d'étiquetter comme spam un mail important.

Remarque 36 (Classes déséquilibrées) Lorsque les classes sont déséquilibrées (par exemple, si $\mathbb{P}(Y = 1)$ est proche de 0), il est également utile de choisir un coût asymétrique, pour « forcer » la prise en compte de la classe sous-représentée (ici, la classe 1). En effet, avec le coût 0–1, le classifieur constant égal à 0 a un risque très faible, égal à $\mathbb{P}(Y = 1)$, et il peut être extrêmement difficile de faire mieux avec une règle de classification non constante ! Prendre un coût asymétrique avec $w_1 \gg w_0$ rééquilibre la situation et permet de réellement « apprendre » quelque chose, quitte à faire un peu moins bien au sens du coût 0–1.

La proposition 2 se généralise au coût asymétrique c_w comme suit.

Proposition 3 *On suppose $\mathcal{Y} = \{0, 1\}$ et l'on considère le coût asymétrique $c_w : (y, y') \mapsto w_{y'} \mathbb{1}_{y \neq y'}$ avec $w_0, w_1 \geq 0$ tels que $w_0 + w_1 > 0$.*

- (i) *Le classifieur défini par $f_w^*(x) = \mathbb{1}_{\eta(x) > w_0/(w_0 + w_1)}$ pour tout $x \in \mathcal{X}$ est un classifieur de Bayes.*
- (ii) *Un classifieur f est un classifieur de Bayes si et seulement si*

$$f(X) = f_w^*(X) \quad \text{presque sûrement,}$$

sauf éventuellement sur l'événement $\{\eta(X) = w_0/(w_0 + w_1)\}$.

- (iii) *Le risque de Bayes vaut :*

$$\mathcal{R}_P^{*,w} = \mathbb{E}\left[\min\left\{w_1 \eta(X), w_0(1 - \eta(X))\right\}\right].$$

- (iv) *L'excès de risque de tout classifieur $f \in \mathcal{F}$ s'écrit :*

$$\ell^w(f_w^*, f) = (w_0 + w_1) \mathbb{E}\left[\left|\eta(X) - \frac{w_0}{w_0 + w_1}\right| \mathbb{1}_{f_w^*(X) \neq f(X)}\right].$$

Remarque 37 (Coût 0–1) Lorsque $w_0 = w_1 = 1$, le coût asymétrique c_w est le coût 0–1 et l'on retrouve exactement la proposition 2.

Remarque 38 (Fonction de coût générale en classification) En classification binaire, avec $\mathcal{Y} = \{0, 1\}$, toute fonction de coût c sur \mathcal{Y} (non identiquement nulle) s'écrit comme un coût asymétrique avec $w_0 = c(1, 0)$ et $w_1 = c(0, 1)$, puisque $c(y, y) = 0$ pour tout $y \in \mathcal{Y}$ par hypothèse. La proposition 3 couvre donc l'ensemble des fonctions de coût possibles, à l'exception du coût identiquement nul pour lequel tout classifieur est un classifieur de Bayes. D'autres mesures d'erreur existent toutefois en classification si l'on considère des « pseudo-classifieurs » $f : \mathcal{X} \rightarrow \mathbb{R}$ (malgré le fait que les étiquettes ne prennent que deux valeurs), voir la section 4.

Remarque 39 (Classification binaire et test d'hypothèse) Supposons que $p = \mathbb{P}(Y = 1) \in]0, 1[$. On peut alors définir $P_{X,j}$ la loi de X sachant $Y = j$ pour $j \in \{0, 1\}$. Soit μ une mesure qui domine $P_{X,0}$ et $P_{X,1}$ (on peut toujours prendre $\mu = \frac{1}{2}(P_{X,0} + P_{X,1})$, mais souvent d'autres choix sont judicieux). On note g_j la densité de $P_{X,j}$ par rapport à μ pour $j \in \{0, 1\}$. Alors, on peut envisager le problème de classification (attribuer une étiquette 0 ou 1 à un point X_{n+1} tiré suivant la loi P_X) comme un problème de test de $H_0 : \langle X_{n+1} \sim P_{X,0} \rangle$ contre $H_1 : \langle X_{n+1} \sim P_{X,1} \rangle$ (ce que l'on peut rapprocher du point de vue proposé à la parenthèse 10 en section 1.2). Le test de rapport de vraisemblance rejette H_0 si et seulement si

$$\frac{g_1(X_{n+1})}{g_0(X_{n+1})} > t$$

où $t \geq 0$ est un seuil à choisir en fonction du niveau voulu. Or, pour P_X presque tout x , on a :

$$\eta(x) = \frac{pg_1(x)}{(1-p)g_0(x) + pg_1(x)}. \quad (7)$$

Par conséquent, pour P_X presque tout x , le prédicteur de Bayes $f_w^*(x)$ vaut 1 si et seulement si :

$$\eta(x) = \frac{pg_1(x)}{(1-p)g_0(x) + pg_1(x)} > \frac{w_0}{w_0 + w_1}.$$

Ceci équivaut à :

$$\frac{g_1(x)}{g_0(x)} > \frac{\mathbb{P}(Y = 0)}{\mathbb{P}(Y = 1)} \frac{w_0}{w_1} =: t(w).$$

Le classifieur de Bayes pour un coût asymétrique quelconque est donc équivalent au test de rapport de vraisemblance, avec une correspondance explicite entre les poids (w_0, w_1) et le seuil t . On remarque notamment que l'asymétrie entre H_0 et H_1 dans un test d'hypothèses se traduit, au niveau de la classification, par l'usage de poids asymétriques dans la fonction de coût c_w . Les poids des deux classes jouent aussi un rôle via le rapport $\mathbb{P}(Y = 0)/\mathbb{P}(Y = 1)$, voir la remarque 36 à ce sujet. Il y a toutefois une grosse différence entre classification et test : ce qui précède nécessite la connaissance de $P_{X,0}$ et $P_{X,1}$ (au moins du rapport de vraisemblance g_1/g_0). S'il faut les estimer à l'aide d'un échantillon, cela devient un problème aussi difficile (voire plus) que le problème de classification initial : il faut estimer (le quotient de) deux densités. L'idée de l'apprentissage statistique est précisément de proposer des méthodes se focalisant directement sur le problème de classification.

2.3 Liens entre régression et classification

On se limite désormais au coût de classification 0–1. L’expression $\mathbb{1}_{\hat{\eta}(x)>1/2}$ donnée par la proposition 2 pour le classifieur de Bayes suggère que l’on peut aborder le problème de classification en deux temps. D’abord, on estime la fonction de régression par $\hat{\eta}$. Ensuite, on utilise le classifieur $x \mapsto \mathbb{1}_{\hat{\eta}(x)>1/2}$. On parle alors de classification par « plug-in »¹⁰.

Définition 2 Soit $\hat{\eta}$ une règle de régression. La règle de classification qui à tout échantillon $D_n \in (\mathcal{X} \times \{0, 1\})^n$ et à tout $x \in \mathcal{X}$ associe

$$\hat{f}_{\hat{\eta}}(D_n; x) := \mathbb{1}_{\hat{\eta}(D_n; x)>1/2}$$

est appelée *règle de classification par plug-in associée à $\hat{\eta}$* .

Les règles de classification par partition sont des exemples de règle par plug-in (associées à la règle de régression par partition correspondante), voir la parenthèse 32 en section 2.2. La section 5 donne d’autres exemples. Enfin, certains réseaux de neurones sont souvent, pour des raisons algorithmiques, approchés en classification par des règles par plug-in.

Parenthèse 40 (Coût asymétrique et plug-in) Quand on utilise un coût asymétrique c_w , on peut également définir une règle par plug-in « asymétrique » par :

$$\hat{f}_{\hat{\eta}, w}(D_n; x) := \mathbb{1}_{\hat{\eta}(D_n; x)>w_0/(w_0+w_1)}.$$

Au vu de la parenthèse 32, on peut donc définir des règles par partition asymétriques en classification.

Parenthèse 41 (Étiquettes ± 1 et plug-in) La définition 2 repose fortement sur le fait que les étiquettes Y_i valent 0 ou 1. Si ce n’est pas le cas (par exemple, avec l’autre convention classique $Y_i \in \{-1, 1\}$), il faut modifier en conséquence la définition des règles par plug-in.

Parenthèse 42 (Plug-in et classification multiclasse) L’idée de plug-in se généralise au cas multiclasse. Pour cela, on suppose donnée une règle de régression (multivariée) $\hat{\eta}$ qui estime

$$\eta(X) = (\mathbb{P}(Y = y | X))_{y \in \mathcal{Y}},$$

et la règle de classification par plug-in associée est définie par :

$$\hat{f}_{\hat{\eta}}(D_n; x) \in \operatorname{argmax}_{y \in \mathcal{Y}} (\hat{\eta}(D_n; x))_y.$$

10. On peut traduire le terme anglais « plug-in » par substitution empirique, ou simplement substitution.

Quand on utilise une règle par plug-in, on peut majorer son excès de risque de classification 0–1 à l'aide de l'excès de risque quadratique de la règle de régression correspondante.

Proposition 4 Soit P une loi sur $\mathcal{X} \times \{0, 1\}$, η la fonction de régression associée, $\widehat{\eta}$ une règle de régression et $\widehat{f}_{\widehat{\eta}}$ la règle de classification par plug-in associée (définition 2). On a alors, pour le coût 0–1 en classification :

$$\begin{aligned}\ell(f^*, \widehat{f}_{\widehat{\eta}}(D_n)) &\leq 2 \mathbb{E} \left[|\widehat{\eta}(D_n; X) - \eta(X)| \mid D_n \right] \\ &\leq 2 \sqrt{\mathbb{E} \left[(\widehat{\eta}(D_n; X) - \eta(X))^2 \mid D_n \right]}.\end{aligned}$$

En particulier, si $\widehat{\eta}$ est faiblement consistante pour P en régression avec le coût quadratique, alors $\widehat{f}_{\widehat{\eta}}$ est faiblement consistante pour P en classification avec le coût 0–1.

Démonstration On utilise la formule (5) pour l'excès de risque 0–1 :

$$\ell(f^*, \widehat{f}_{\widehat{\eta}}(D_n)) = \mathbb{E} \left[|2\eta(X) - 1| \mathbf{1}_{f^*(X) \neq \widehat{f}_{\widehat{\eta}}(D_n; X)} \mid D_n \right].$$

Or, pour tout $x \in \mathcal{X}$, $f^*(x) \neq \widehat{f}_{\widehat{\eta}}(D_n; x)$ signifie que $\eta(x)$ et $\widehat{\eta}(D_n; x)$ sont de part et d'autre de 1/2, si bien que :

$$\left| \eta(X) - \frac{1}{2} \right| \mathbf{1}_{f^*(X) \neq \widehat{f}_{\widehat{\eta}}(D_n; X)} \leq |\eta(X) - \widehat{\eta}(D_n; X)|.$$

En intégrant (conditionnellement à D_n), on obtient la première inégalité. La deuxième s'en déduit par l'inégalité de Jensen. \square

On rappelle que, par la proposition 1,

$$\mathbb{E} \left[(\widehat{\eta}(D_n; X) - \eta(X))^2 \mid D_n \right]$$

est égal à l'excès de risque de $\widehat{\eta}(D_n)$ en régression avec le coût quadratique. En ce sens, la proposition 4 établit un lien entre les excès de risque en classification et régression : si $\widehat{\eta}$ estime bien η (en régression), alors la règle par plug-in associée est une bonne règle de classification.

Le majorant intermédiaire

$$\mathbb{E} \left[|\widehat{\eta}(D_n; X) - \eta(X)| \mid D_n \right]$$

ne s'interprète pas en général comme un excès de risque en régression, même s'il s'agit bien d'une erreur d'estimation L^1 de η par $\widehat{\eta}$.

Remarque 43 (Variantes de la proposition 4) La proposition 4 se généralise à un coût asymétrique quelconque (exercice 4). On peut également obtenir une borne

plus fine dans le cas zéro-erreur (exercice 5). Mentionnons enfin que le théorème 1 en section 4 peut être vu comme une généralisation de la proposition 4 à un « coût convexe » quelconque (au lieu du coût quadratique).

Remarque 44 (La classification est-elle plus facile ?) La proposition 4 est souvent interprétée comme « la classification est plus facile que la régression ». Du point de vue de la consistance faible, c'est exact : si l'on sait construire une règle de régression consistante pour toute loi P telle que Y est bornée¹¹, alors on peut en déduire une règle de classification universellement consistante. Cette interprétation est en revanche fausse vis-à-vis du problème général de prévision : en classification comme en régression, on cherche toujours à faire aussi bien que possible (par exemple, avoir un excès de risque qui tend vers 0 avec la *meilleure vitesse* possible, un objectif plus ambitieux que la consistance). Or, les vitesses optimales de classification et de régression ne correspondent pas toujours comme ce que suggère la proposition 4 : il ne suffit pas de savoir atteindre la vitesse optimale en régression (avec le risque quadratique) pour atteindre la vitesse optimale en classification (avec le risque 0–1). De plus, les règles de classification les plus naturelles¹² conduisent souvent à des problèmes d'optimisation difficiles, nécessitant souvent de recourir à d'autres méthodes¹³. La parenthèse 45 ci-dessous détaille ces deux points sur un exemple.

Parenthèse 45 (Discrimination linéaire et plug-in) On peut illustrer la remarque 44 ci-dessus en considérant l'exemple de la discrimination linéaire (exemple 5 en section 3.3) et en s'appuyant sur les résultats de la suite de ce texte. Soit $\mathcal{X} = \mathbb{R}^p$ et $\mathcal{P}(S_{\text{class}}^{\text{lin}})$ l'ensemble des lois P sur $\mathbb{R}^p \times \{0, 1\}$ telles que le classifieur de Bayes pour le risque 0–1 est de la forme $f_P^* : \mathbf{x} \mapsto \mathbf{1}_{\langle \mathbf{w}^*, \mathbf{x} \rangle \geq 0}$, avec $\mathbf{w}^* \in \mathbb{R}^p$. Alors, d'après les résultats de la section 3.7 et la proposition 17 en section 7, la vitesse (minimax) optimale d'apprentissage en classification 0–1 est de l'ordre de $\sqrt{p/n}$. Cette vitesse est atteinte pour un minimiseur du risque empirique sur $S_{\text{class}}^{\text{lin}}$, qui nécessite de résoudre un problème algorithmique difficile (voir la remarque 52 en section 3.3). En régression avec le risque quadratique, savoir que $P \in \mathcal{P}(S_{\text{class}}^{\text{lin}})$ autorise la fonction de régression η_P à être très irrégulière. Par exemple, dans le demi-espace d'équation $\langle \mathbf{w}^*, \mathbf{x} \rangle \geq 0$, la seule contrainte est que η_P doit prendre des valeurs dans $[1/2, 1]$. Le risque minimax sur $\mathcal{P}(S_{\text{class}}^{\text{lin}})$ est donc minoré par une constante absolue $\kappa > 0$ (c'est un corollaire du résultat de l'exercice 37). Par conséquent, il est très facile de construire une règle optimale (au sens du minimax) en régression — par exemple, en posant $\hat{f} \equiv 1/2$ —, mais la règle par plug-in associée n'est pas optimale en classification 0–1.

Audibert et Tsybakov (2007, section 1) identifient les causes générales de ce problème : tout dépend du type d'hypothèse que l'on fait sur P . Les règles par plug-in fonctionnent très bien si l'on suppose η_P régulière, mais elles ne sont pas appropriées (en tout cas, théoriquement) sous la seule hypothèse que la *frontière de décision* optimale (en classification 0–1) est régulière.

Signalons que pour des raisons algorithmiques, on utilise souvent malgré tout

11. Il suffit d'avoir la consistance en régression lorsque $Y \in \{0, 1\}$ presque sûrement.

12. Par exemple, les règles minimisant le risque empirique, avec le coût 0–1.

13. Par exemple, minimiser le risque empirique avec l'un des coûts convexes définis en section 4.

des méthodes de type plug-in à la place de la minimisation du risque empirique 0–1 sur $S_{\text{class}}^{\text{lin}}$, par exemple via la minimisation d'un Φ -risque empirique sur S^{lin} ; on y perd au niveau des garanties théoriques, mais on y gagne énormément en temps de calcul.

À notre connaissance, la question de savoir si une règle de classification optimale sur $\mathcal{P}(S_{\text{class}}^{\text{lin}})$ peut avoir une complexité polynomiale en pire cas reste un problème théorique ouvert.

3 Minimisation du risque empirique

De nombreuses règles d'apprentissage reposent sur le principe de minimisation du risque empirique. Cette section présente les grandes lignes de leur analyse théorique commune. Le lecteur intéressé peut consulter l'article de survol de Boucheron *et al.* (2005) pour plus de détails et une bibliographie complète.

Dans un premier temps, nous considérons un problème de prévision général, avec les notations de la section 1.

3.1 Principe

Définition 3 (Risque empirique) Le *risque empirique* d'un prédicteur f , sur un échantillon $D_n \in (\mathcal{X} \times \mathcal{Y})^n$, avec la fonction de coût $c : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}^+$, est défini par :

$$\widehat{\mathcal{R}}_n(f) = \widehat{\mathcal{R}}_n^c(f; D_n) = \frac{1}{n} \sum_{i=1}^n c(f(X_i), Y_i).$$

Le risque empirique de f est ainsi l'erreur moyenne de prévision de f sur l'échantillon D_n , mesurée avec le coût c .

Parenthèse 46 (Mesure empirique) Définissons la mesure empirique de l'échantillon D_n par :

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

Alors, on peut écrire le risque empirique ainsi :

$$\widehat{\mathcal{R}}_n^c(f) = \mathcal{R}_{P_n}^c(f).$$

Désormais, on suppose toujours que le risque et le risque empirique sont associés à une même fonction de coût c , sauf mention explicite du contraire. Nous ne précisons donc plus nécessairement la dépendance en c dans les notations. Alors, pour tout $f \in \mathcal{F}$ fixé, $\widehat{\mathcal{R}}_n(f)$ estime sans biais le risque de f :

$$\mathbb{E}[\widehat{\mathcal{R}}_n(f)] = \mathbb{E}[c(f(X), Y)] = \mathcal{R}_P(f).$$

Puisque l'on veut trouver un prédicteur dont le risque est petit, à partir de la seule connaissance de D_n , il paraît naturel de choisir un prédicteur dont le risque empirique est petit. Ceci conduit à la définition suivante.

Définition 4 Soit $S \subset \mathcal{F}$, $D_n \in (\mathcal{X} \times \mathcal{Y})^n$ et $f \in \mathcal{F}$. On dit que f est un minimiseur du risque empirique¹⁴ sur S si :

$$f \in S \quad \text{et} \quad \widehat{\mathcal{R}}_n(f; D_n) = \inf_{g \in S} \widehat{\mathcal{R}}_n(g; D_n).$$

De même, on dit qu'une règle d'apprentissage \widehat{f} est une règle par minimisation du risque empirique sur S si pour tout $n \geq 1$ et presque tout échantillon D_n de taille n , $\widehat{f}(D_n)$ est un minimiseur du risque empirique sur S . Le sous-ensemble S de l'ensemble \mathcal{F} des prédicteurs est appelé *modèle*.

Dans la suite, on note \widehat{f}_S une règle d'apprentissage par minimisation du risque empirique sur S même s'il y en a en général plusieurs (voire une infinité). Ceci signifie que l'on étudie les propriétés de n'importe laquelle de ces règles, en s'appuyant uniquement sur le fait qu'elle vérifie la définition 4. Le problème du choix de S est discuté en section 3.9.

Le principe de minimisation du risque empirique est une généralisation de la méthode des moindres carrés, qui correspond à utiliser le coût quadratique en régression ; on parle alors d'*estimateur des moindres carrés*.

Les règles minimisant le risque empirique appartiennent à la famille des *M-estimateurs*, ou estimateurs par minimum de contraste (Bickel et Doksum, 2001, chapitre 2), qui sont des estimateurs de la forme :

$$\widehat{f} \in \operatorname{argmin}_{f \in S} \left\{ \sum_{i=1}^n \gamma(f; (X_i, Y_i)) \right\}.$$

Par exemple, les estimateurs du maximum de vraisemblance sont des M-estimateurs. Dans ce texte, on se limite à une fonction de contraste γ de la forme :

$$(f; (x_i, y_i)) \mapsto c(f(x_i), y_i).$$

Parenthèse 47 (Existence, unicité et calculabilité de \widehat{f}_S) La définition 4 précise *un* minimiseur du risque empirique sur S car son unicité n'est en rien garantie (il peut y en avoir une infinité!). Il n'en existe pas non plus toujours au moins un. Et même lorsqu'il en existe, le calcul d'un minimiseur du risque empirique peut nécessiter un coût de calcul prohibitif. Pour ces raisons, on introduit la notion de minimiseur approché du risque empirique sur S . Pour tout $\rho \geq 0$, $S \subset \mathcal{F}$, $D_n \in (\mathcal{X} \times \mathcal{Y})^n$ et $f \in \mathcal{F}$, on dit que f est un ρ -minimiseur du risque empirique sur S si :

$$f \in S \quad \text{et} \quad \widehat{\mathcal{R}}_n(f; D_n) \leq \inf_{g \in S} \widehat{\mathcal{R}}_n(g; D_n) + \rho.$$

14. Le terme anglais est « empirical risk minimizer », souvent abrégé ERM.

De même, on dit qu'une règle d'apprentissage \hat{f} est une règle par ρ -minimisation du risque empirique sur S si pour tout $n \geq 1$ et presque tout échantillon D_n de taille n , $\hat{f}(D_n)$ est un ρ -minimiseur du risque empirique sur S .

3.2 Exemples en régression

On se place en régression ($\mathcal{Y} = \mathbb{R}$), comme à la section 2.1. On a alors de nombreux exemples classiques de règles par minimisation du risque empirique. Tout d'abord, les règles par partition définies précédemment sont des minimiseurs du risque empirique.

Proposition 5 Soit \mathcal{A} une partition mesurable de \mathcal{X} , finie ou dénombrable. En régression, on définit le modèle par partition associé :

$$\begin{aligned} S_{\text{reg}}^{\text{part}}(\mathcal{A}) &:= \{f \in \mathcal{F} / f \text{ est constante sur } A \text{ pour tout } A \in \mathcal{A}\} \\ &= \overline{\text{vect}\{\mathbf{1}_A / A \in \mathcal{A}\}}. \end{aligned}$$

Alors, si c est le coût quadratique, la règle de régression par partition $\hat{f}_{\mathcal{A}}^{\text{P-r}}$ associée à \mathcal{A} (définie à l'exemple 1 en section 2.1) est une règle par minimisation du risque empirique sur $S_{\text{reg}}^{\text{part}}(\mathcal{A})$.

Démonstration Soit $n \geq 1$ un entier et $D_n = (x_i, y_i)_{1 \leq i \leq n}$ un échantillon. Pour tout $A \in \mathcal{A}$ et tout $x \in A$, $\hat{f}_{\mathcal{A}}^{\text{P-r}}(D_n; x)$ est égal à la moyenne des y_i tels que $x_i \in A$ (qui vaut par convention 0 si aucun des x_i n'appartient à A). Par conséquent, $\hat{f}_{\mathcal{A}}^{\text{P-r}}(D_n) \in S_{\text{reg}}^{\text{part}}(\mathcal{A})$. De plus, tout $f \in S_{\text{reg}}^{\text{part}}(\mathcal{A})$ peut s'écrire $f = \sum_{A \in \mathcal{A}} f_A \mathbf{1}_A$, de telle sorte que son risque empirique est égal à :

$$\sum_{A \in \mathcal{A}} \sum_{i / x_i \in A} (y_i - f_A)^2.$$

Or, pour tout $A \in \mathcal{A}$,

$$u \mapsto \sum_{i / x_i \in A} (y_i - u)^2$$

est minimale lorsque u est égal à la moyenne des y_i tels que $x_i \in A$. Donc $\hat{f}_{\mathcal{A}}^{\text{P-r}}(D_n)$ minimise le risque empirique sur $S_{\text{reg}}^{\text{part}}(\mathcal{A})$. \square

Exemple 4 (Régression linéaire) Si $\mathcal{X} \subset \mathbb{R}^p$, pour tout $\mathbf{w} \in \mathbb{R}^p$, on définit le prédicteur

$$f_{\mathbf{w}} : \mathbf{x} \in \mathcal{X} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle = \sum_{j=1}^p w_j x_j.$$

Alors, minimiser le risque empirique sur

$$S^{\text{lin}} := \{f_{\mathbf{w}} / \mathbf{w} \in \mathbb{R}^p\}$$

correspond à la régression linéaire : on cherche un prédicteur qui dépend linéairement des covariables.

Lorsque c est le coût quadratique, on obtient le très classique estimateur des moindres carrés du modèle linéaire (Cornillon et Matzner-Løber, 2011), illustré par la figure 3. Si l'on suppose que la matrice $\mathbf{X} = (x_{i,j})_{1 \leq i \leq n, 1 \leq j \leq p}$ est de rang p , on dispose alors d'une formule close :

$$\hat{f}_{S^{\text{lin}}} = \hat{f}_{\hat{\mathbf{w}}} \quad \text{avec} \quad \hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \quad \text{et} \quad \mathbf{y} = (Y_i)_{1 \leq i \leq n}.$$

Il est également classique de considérer

$$S_J^{\text{lin}} := \{f_{\mathbf{w}} / \mathbf{w} \in \mathbb{R}^p \text{ avec } w_j = 0 \text{ pour tout } j \notin J\}$$

pour un $J \subset \{1, \dots, p\}$ bien choisi. L'ensemble J correspond aux indices des covariables retenues pour prévoir Y au sein du modèle S_J^{lin} . Le problème de choisir J , appelé *sélection de variables*, est un domaine de recherche en tant que tel. C'est un cas particulier du problème de sélection de modèles, qui est discuté en section 3.9.

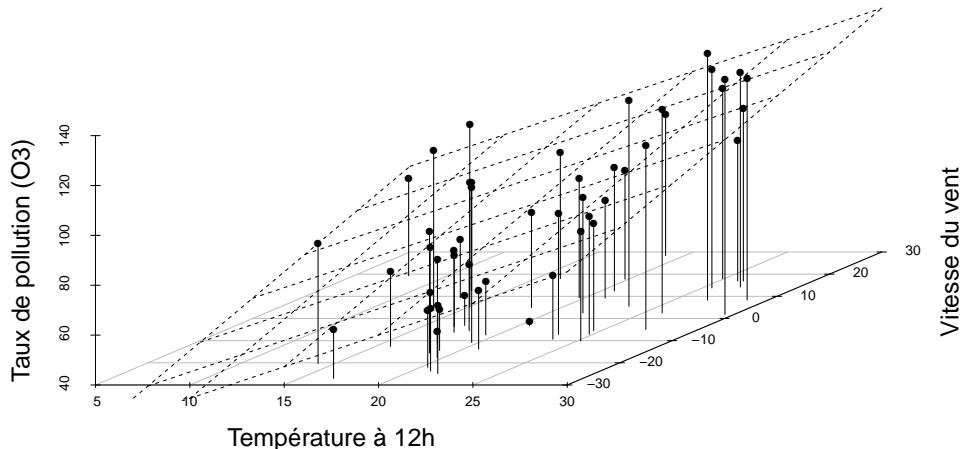


FIGURE 3 – *Taux de pollution à l'ozone en fonction de données météorologiques : ajustement d'un modèle de régression linéaire (en ajoutant la variable constante, comme suggéré à la remarque 48). Données issues de Cornillon et Matzner-Løber (2011).*

Remarque 48 (Régression affine) On s'intéresse souvent à des prédicteurs s'écrivant comme une fonction *affine* des covariables, c'est-à-dire, de la forme

$$f_{\mathbf{w}, b} : \mathbf{x} \in \mathcal{X} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle + b$$

avec $\mathbf{w} \in \mathbb{R}^p$ et $b \in \mathbb{R}$. Ce cas est en fait déjà couvert par l'exemple 4, en ajoutant une covariable constante (si elle n'y est pas déjà), c'est-à-dire, en remplaçant chaque $\mathbf{x} = (x_1, \dots, x_p) \in \mathbb{R}^p$ par $(1, x_1, \dots, x_p) \in \mathbb{R}^{p+1}$.

Remarque 49 (Transformations des \mathbf{x}_i) Il est classique d'appliquer la régression linéaire après avoir transformé la variable explicative \mathbf{x} , afin de modéliser une relation non-linéaire entre \mathbf{x} et la variable d'intérêt y . Formellement, on se donne une famille $(\varphi_k)_{1 \leq k \leq m}$ de fonctions mesurables $\mathcal{X} \rightarrow \mathbb{R}$ et l'on remplace chaque $\mathbf{x} \in \mathbb{R}^p$ par $(\varphi_k(\mathbf{x}))_{1 \leq k \leq m} \in \mathbb{R}^m$.

Par exemple, si $p = 1$, la régression polynomiale s'obtient en remplaçant $\mathbf{x} \in \mathbb{R}$ par le vecteur $(1, x_1, x_1^2, \dots, x_1^d) \in \mathbb{R}^{d+1}$. On ajuste un polynôme de degré d pour modéliser la relation entre \mathbf{x} et y . On peut faire de même en dimension $p > 1$, en introduisant si besoin des covariables de la forme $x_{i_1}^\alpha x_{i_2}^\beta$ pour tenir compte d'interactions entre variables.

Il est aussi habituel d'utiliser des transformations non polynomiales, par exemple en choisissant pour $(\varphi_k)_{1 \leq k \leq m}$ les m premières fonctions d'une famille libre de \mathcal{F} : base de Fourier, base d'ondelettes, etc. (la famille de fonctions choisie dépendant des informations dont on dispose sur la régularité de η). Györfi *et al.* (2002, section 10.3) démontrent alors un résultat de consistance sous des hypothèses assez faibles sur $(\varphi_k)_{1 \leq k \leq m}$.

On peut ainsi écrire la règle par partition associée à la partition \mathcal{A} comme une règle par régression linéaire obtenue après transformation de chaque $\mathbf{x} \in \mathbb{R}^p$ en $(\mathbf{1}_{\mathbf{x} \in A})_{A \in \mathcal{A}}$.

Parenthèse 50 (Régression linéaire avec \mathcal{X} général) Il est possible d'étendre l'exemple 4 au cas où \mathcal{X} est un espace préhilbertien : il suffit de disposer d'un produit scalaire sur \mathcal{X} pour pouvoir parler de régression linéaire. Ceci est notamment utile lorsque \mathcal{X} est un espace fonctionnel, cas auquel on peut se ramener en transformant les covariables à l'aide d'un noyau semi-défini positif.

Il existe encore d'autres règles de régression qui minimisent un risque empirique, par exemple certains réseaux de neurones (voir le livre de Györfi *et al.* (2002, chapitre 16), par exemple) et les réseaux à fonction de base radiale¹⁵ (voir le livre de Györfi *et al.* (2002, chapitre 17) et le texte de Viennet (2006), par exemple).

3.3 Exemples en classification

Considérons le problème de classification binaire supervisée ($\mathcal{Y} = \{0, 1\}$), comme en section 2.2. Un premier exemple classique de règle de classification par minimisation du risque empirique est celui des règles par partition.

15. Le terme anglais est « radial basis function (RBF) network ».

Proposition 6 Soit \mathcal{A} une partition mesurable de \mathcal{X} , finie ou dénombrable. En classification binaire, on définit le modèle par partition associé :

$$\begin{aligned} S_{\text{class}}^{\text{part}}(\mathcal{A}) &:= \{f \in \mathcal{F} / f \text{ est constante sur } A \text{ pour tout } A \in \mathcal{A}\} \\ &= \left\{ f : x \in \mathcal{X} \mapsto \sum_{A \in \mathcal{A}} \alpha_A \mathbb{1}_A / \alpha \in \{0, 1\}^{\mathcal{A}} \right\} \\ &= \{\mathbb{1}_{\bigcup_{B \in \mathcal{B}} B} / \mathcal{B} \subset \mathcal{A}\} \\ &= S_{\text{reg}}^{\text{part}}(\mathcal{A}) \cap \{f : \mathcal{X} \rightarrow \{0, 1\} \text{ mesurable}\}. \end{aligned}$$

Alors, si c est le coût 0–1, la règle de classification par partition $\hat{f}_{\mathcal{A}}^{\text{p-c}}$ associée à \mathcal{A} (définie à l'exemple 3 en section 2.2) est une règle par minimisation du risque empirique sur $S_{\text{class}}^{\text{part}}(\mathcal{A})$.

Démonstration Soit $n \geq 1$ un entier et $D_n = (x_i, y_i)_{1 \leq i \leq n}$ un échantillon. Pour tout $A \in \mathcal{A}$ et tout $x \in A$, $\hat{f}_{\mathcal{A}}^{\text{p-c}}(D_n; x)$ est égal à :

$$\mathbb{1}_{\text{Card}\{i / y_i=1 \text{ et } x_i \in A\} > \text{Card}\{i / y_i=0 \text{ et } x_i \in A\}} \in \underset{u \in \{0, 1\}}{\operatorname{argmin}} \left\{ \sum_{i / x_i \in A} \mathbb{1}_{y_i \neq u} \right\}. \quad (8)$$

Par conséquent, $\hat{f}_{\mathcal{A}}^{\text{p-c}}(D_n) \in S_{\text{class}}^{\text{part}}(\mathcal{A})$. De plus, tout $f \in S_{\text{class}}^{\text{part}}(\mathcal{A})$ pouvant s'écrire $f = \sum_{A \in \mathcal{A}} f_A \mathbb{1}_A$ avec $f_A \in \{0, 1\}$, son risque empirique vaut :

$$\sum_{A \in \mathcal{A}} \sum_{i / x_i \in A} \mathbb{1}_{y_i \neq f_A}.$$

Au vu de (8), sur $S_{\text{class}}^{\text{part}}(\mathcal{A})$, le risque empirique est donc minimal en $\hat{f}_{\mathcal{A}}^{\text{p-c}}(D_n)$. \square

Exemple 5 (Discrimination linéaire) Si $\mathcal{X} \subset \mathbb{R}^p$, pour tout $\mathbf{w} \in \mathbb{R}^p$, on définit le classifieur

$$f_{\mathbf{w}}^{\text{class}} : \mathbf{x} \in \mathcal{X} \mapsto \mathbb{1}_{\langle \mathbf{w}, \mathbf{x} \rangle \geq 0}.$$

Alors, minimiser le risque empirique sur

$$S_{\text{class}}^{\text{lin}} := \{f_{\mathbf{w}}^{\text{class}} / \mathbf{w} \in \mathbb{R}^p\}$$

correspond à une règle de discrimination linéaire : on cherche un classifieur qui sépare \mathcal{X} en deux demi-espaces.

Le lecteur intéressé peut consulter le livre de Devroye *et al.* (1996, chapitre 4) au sujet de la discrimination linéaire. Signalons que ce type de règle est à la base de plusieurs autres règles de classification importantes, telles que les réseaux de neurones (qui sont un autre exemple de règle minimisant le risque empirique), les SVM (voir la parenthèse 54 ci-dessous) et dans une moindre mesure les arbres de décision.

Remarque 51 (Discrimination affine) Comme en régression, le plus souvent on s'intéresse plutôt à une séparation « affine », c'est-à-dire, à des classificateurs de la forme :

$$f_{(\mathbf{w}, b)}^{\text{class}} : \mathbf{x} \in \mathcal{X} \mapsto \mathbb{1}_{\langle \mathbf{w}, \mathbf{x} \rangle + b \geq 0}$$

avec $b \in \mathbb{R}$. Pour cette raison, de même qu'en régression, on ajoute en général une covariable constante (voir la remarque 48). On peut aussi procéder à des transformations générales des \mathbf{x}_i , comme à la remarque 49.

Remarque 52 (Difficulté algorithmique) En dimension $p = 1$, la minimisation exacte du risque empirique sur $S_{\text{class}}^{\text{lin}}$ est faisable, avec une complexité de l'ordre de $n \ln(n)$, d'après Kearns *et al.* (1997). En revanche, lorsque $p \geq 2$, c'est un problème NP-dur. Une minimisation approchée (à p/n près) est possible en énumérant $2 \binom{n}{p}$ classificateurs, lorsque les X_i sont en position générale (Devroye *et al.*, 1996, chapitre 4).

Remarque 53 (Sélection de variables) De même qu'en régression, il est souvent judicieux de considérer seulement un sous-ensemble des covariables disponibles, c'est-à-dire, remplacer le modèle $S_{\text{class}}^{\text{lin}}$ par :

$$S_{\text{class}, J}^{\text{lin}} := \{f_{\mathbf{w}}^{\text{class}} / \mathbf{w} \in \mathbb{R}^p \text{ avec } w_j = 0 \forall j \notin J\}$$

où $J \subset \{1, \dots, p\}$ est bien choisi.

Parenthèse 54 (Discrimination linéaire avec \mathcal{X} général) Comme pour la régression linéaire, on peut définir une règle de classification par discrimination linéaire dès que \mathcal{X} est un espace vectoriel muni d'un produit scalaire. Et lorsque ce n'est pas le cas, on peut s'y ramener en transformant les variables explicatives $x \in \mathcal{X}$ à l'aide d'un noyau semi-défini positif (c'est le principe des SVM non-linéaires).

3.4 Décomposition du risque : erreur d'approximation, erreur d'estimation

Désormais, jusqu'à la fin de la section 3, on considère $S \subset \mathcal{F}$ un modèle et \hat{f}_S un minimiseur du risque empirique sur S . Avec probabilité 1, puisque $\hat{f}_S \in S$, l'excès de risque $\ell(f^*, \hat{f}_S)$ est minoré par :

$$\ell(f^*, S) := \inf_{f \in S} \ell(f^*, f). \quad (9)$$

La quantité (9) est appelée *erreur d'approximation du modèle* S . Elle représente l'erreur minimale (mesurée par l'excès de risque) commise par un prédicteur dans S . On l'interprète souvent comme une « distance » entre le prédicteur de Bayes f^* et le modèle S (même lorsque l'excès de risque ne définit pas une distance sur \mathcal{F}).

Lorsque l'infimum dans (9) est atteint, on note f_S^* un des meilleurs prédicteurs de S :

$$f_S^* \in \operatorname{argmin}_{f \in S} \ell(f^*, f).$$

On a alors :

$$\ell(f^*, S) = \ell(f^*, f_S^*).$$

On définit l'*erreur d'estimation* de \hat{f}_S comme la différence entre l'excès de risque de \hat{f}_S et l'erreur d'approximation :

$$\mathcal{R}_P(\hat{f}_S) - \inf_{f \in S} \mathcal{R}_P(f) = \ell(f^*, \hat{f}_S) - \ell(f^*, S) \geq 0. \quad (10)$$

Ainsi, l'excès de risque de \hat{f}_S peut se décomposer en la somme de l'erreur d'approximation et de l'erreur d'estimation :

$$\ell(f^*, \hat{f}_S) = \underbrace{\ell(f^*, S)}_{\text{erreur d'approximation}} + \underbrace{\ell(f^*, \hat{f}_S) - \ell(f^*, S)}_{\text{erreur d'estimation}}. \quad (11)$$

Par analogie avec la décomposition biais-variance du risque quadratique — voir par exemple l'équation (44) en section 5.2 —, l'erreur d'approximation est souvent appelée « biais (du modèle S) », et l'erreur d'estimation est souvent appelée « variance ».

L'erreur d'approximation peut être calculée ou majorée explicitement dans de nombreux exemples ; les exercices 6 et 7 traitent le cas des règles par partition, en régression et en classification.

L'analyse de l'erreur d'estimation fait l'objet des sous-sections suivantes. En régression avec le coût quadratique, l'ordre de grandeur typique de l'erreur d'estimation est le nombre de « paramètres » du modèle divisé par le nombre d'observations. L'exercice 8 et la remarque 105 qui suit son énoncé justifient cette affirmation dans le cas de règles par partition en régression. L'exemple 6 en section 3.6 et l'exercice 18 traitent le cas des règles par partition en classification 0–1.

Parenthèse 55 (Intérêt de (11)) La décomposition (11) peut sembler vide de sens au premier abord. Son intérêt réside dans le fait que l'erreur d'estimation est toujours positive ou nulle, et s'interprète comme l'augmentation du risque induite par la difficulté d'apprendre au sein du modèle S . Ainsi, lorsque S est décrit par un nombre fini de paramètres, l'erreur d'estimation est la conséquence des erreurs commises en estimant les paramètres du modèle S . Par exemple, en régression linéaire (exemple 4), l'erreur d'estimation est (approximativement) proportionnelle au nombre p de covariables considérées dans le modèle. Signalons également que la décomposition (11) s'avère particulièrement éclairante sur le problème de sélection de modèles, voir la section 3.9.

Parenthèse 56 (Autres définitions de l'erreur d'estimation) On trouve parfois une autre définition de l'erreur d'estimation dans la littérature

statistique, en prenant l'espérance sur D_n :

$$\mathbb{E}[\mathcal{R}_P(\hat{f}_S) - \inf_{f \in S} \mathcal{R}_P(f)].$$

On la nomme également parfois « erreur d'estimation du modèle S », ce qui suppose implicitement que tous les minimiseurs du risque empirique ont le même risque, ou bien que l'on a redéfini l'erreur d'estimation comme :

$$\sup_{\hat{f}_S \in \operatorname{argmin}_{f \in S} \widehat{\mathcal{R}}_n(f)} \{\mathcal{R}_P(\hat{f}_S) - \inf_{f \in S} \mathcal{R}_P(f)\}.$$

Enfin, il arrive aussi que l'on nomme erreur d'estimation un *majorant* de $\mathcal{R}_P(\hat{f}_S) - \inf_{f \in S} \mathcal{R}_P(f)$ ou de son espérance.

3.5 Majoration générale de l'erreur d'estimation

La proposition suivante fournit un majorant simple de l'erreur d'estimation de tout minimiseur du risque empirique sur S .

Proposition 7 Soit $S \subset \mathcal{F}$ un modèle et \hat{f}_S un minimiseur du risque empirique sur S . On a alors :

$$\ell(f^*, \hat{f}_S) - \ell(f^*, S) \leq 2 \sup_{f \in S} |\mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f)|. \quad (12)$$

Démonstration Pour tout $g \in S$, on a :

$$\begin{aligned} & \mathcal{R}_P(\hat{f}_S) - \mathcal{R}_P(g) \\ &= \underbrace{\mathcal{R}_P(\hat{f}_S) - \widehat{\mathcal{R}}_n(\hat{f}_S)}_{\leq \sup_{f \in S} |\mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f)|} + \underbrace{\widehat{\mathcal{R}}_n(\hat{f}_S) - \widehat{\mathcal{R}}_n(g)}_{\leq 0} + \underbrace{\widehat{\mathcal{R}}_n(g) - \mathcal{R}_P(g)}_{\leq \sup_{f \in S} |\mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f)|} \\ &\leq 2 \sup_{f \in S} |\mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f)|. \end{aligned}$$

En prenant le sup sur $g \in S$ dans le membre de gauche, on obtient le résultat annoncé. \square

Parenthèse 57 (Autre démonstration de la proposition 7) On peut aussi voir cette proposition 7 comme un corollaire du lemme 2 énoncé en section 3.9, avec $\mathcal{E} = S$, $\mathcal{C} = \widehat{\mathcal{R}}_n$, $\mathcal{R} = \mathcal{R}_P$ et

$$A = B : f \in \mathcal{E} \mapsto \sup_{f \in S} |\mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f)|.$$

Le majorant de l'erreur d'estimation donné par (12) s'interprète comme une mesure de complexité du modèle S , puisque c'est une fonction croissante de S . On démontre en section 3.7 que cette complexité peut dans certains cas se majorer par une fonction de la « dimension » de S .

Parenthèse 58 (Minimiseur approché du risque empirique) La proposition 7 se généralise à $\hat{f}_{S,\rho}$, n'importe quel ρ -minimiseur du risque empirique sur S . On a alors :

$$\ell(f^*, \hat{f}_{S,\rho}) \leq \underbrace{\ell(f^*, S)}_{\text{approximation}} + \underbrace{2 \sup_{f \in S} |\mathcal{R}_P(f) - \hat{\mathcal{R}}_n(f)|}_{\text{estimation}} + \underbrace{\rho}_{\text{optimisation}}.$$

L'excès de risque de $\hat{f}_{S,\rho}$ est ainsi majoré par la somme de trois termes correspondant aux trois sources d'erreur : approximation (par S), estimation (au sein de S) et optimisation (du risque empirique sur S). Précisons tout de même qu'il ne s'agit que d'une majoration : par exemple, ceci ne prouve pas que commettre une erreur d'optimisation ρ augmente nécessairement le risque de ρ (on laisse au lecteur le soin de trouver des contre-exemples).

Parenthèse 59 (Analyse globale et analyse locale du risque) On dit parfois que

$$2 \sup_{f \in S} |\mathcal{R}_P(f) - \hat{\mathcal{R}}_n(f)|$$

et les quantités qui l'estiment sont des mesures « globales » de la complexité de S . Ceci fait référence au fait que ces quantités mesurent l'écart maximum entre risque et risque empirique sur le modèle S , qui est ici considéré *globalement*. Une telle approche est parfois judicieuse (voir la section 3.7), parfois sous-optimale (par exemple, quand on peut obtenir des « vitesses rapides » en classification, voir les sections 3.8 et 7), voire totalement inopérante (en régression avec le coût quadratique, lorsque S est un espace vectoriel, ce supremum est infini). Dans ces derniers cas, une analyse plus fine (dite « locale ») du risque de \hat{f}_S est possible, au prix d'efforts théoriques plus importants. En une phrase, elles reviennent à démontrer que l'on peut appliquer le lemme 2 (en section 3.9) avec

$$\mathcal{E} = S, \quad \mathcal{C} = \hat{\mathcal{R}}_n, \quad \mathcal{R} = \mathcal{R}_P \quad \text{et} \quad A(f) \approx B(f) \leq \epsilon_n \ell(f^*, f) + \epsilon'_n,$$

où ϵ_n et ϵ'_n sont suffisamment petits. Boucheron *et al.* (2005) donnent une bonne vue d'ensemble sur ces méthodes, des résultats avancés pouvant être trouvés dans l'article de Massart et Nédélec (2006).

Si l'on s'intéresse à l'espérance de l'erreur d'estimation, une majoration plus fine que celle de la proposition 7 est possible (et elle se généralise à un minimiseur approché du risque empirique, voir l'exercice 9).

Proposition 8 Soit $S \subset \mathcal{F}$ un modèle et \hat{f}_S une règle par minimisation du risque empirique sur S . Si le risque \mathcal{R}_P et le risque empirique $\hat{\mathcal{R}}_n$ sont définis avec le même coût c , on a :

$$\mathbb{E}[\ell(f^*, \hat{f}_S(D_n)) - \ell(f^*, S)] \leq \mathbb{E}\left[\sup_{f \in S} \{\mathcal{R}_P(f) - \hat{\mathcal{R}}_n(f)\}\right]. \quad (13)$$

Remarque 60 (Processus empirique) Les propositions 7 et 8 mettent en évidence l'importance de l'écart maximal entre risque et risque empirique sur le modèle. De manière générale, la théorie des processus empiriques étudie ce type de quantité (van der Vaart et Wellner, 1996).

Parenthèse 61 (Mesurabilité du supremum) On a supposé ici implicitement que

$$\sup_{f \in S} \{ \mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f) \}$$

est mesurable et que son espérance est bien définie. Sous les hypothèses générales de ce texte, ceci est toujours vrai lorsque S est fini ou dénombrable. Dans le cas général, il faudrait ajouter une condition de séparabilité. Dans la suite, nous supposons toujours (implicitement) qu'une telle condition est vérifiée.

Parenthèse 62 (Valeur absolue dans le supremum) Les propositions 7 et 8 font apparaître deux quantités très similaires :

$$\sup_{f \in S} |\mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f)| \quad \text{et} \quad \sup_{f \in S} \{ \mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f) \}.$$

Le deuxième supremum (sans valeur absolue) et bien sûr inférieur ou égal au premier, mais de combien ? Dans bien des cas, ces deux quantités sont très proches. Par exemple, en classification 0–1, si S est « symétrique » — c'est-à-dire, pour tout $f \in S$, $1 - f \in S$ —, alors ces deux quantités sont égales, puisque

$$\mathcal{R}_P(1 - f) - \widehat{\mathcal{R}}_n(1 - f) = -[\mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f)].$$

Lorsque S n'est pas symétrique, ajouter la valeur absolue dans le supremum revient à remplacer S par $S \cup \{1 - f / f \in S\}$, ce qui n'augmente que très peu les bornes démontrées dans les sections 3.6 et 3.7 ci-après (voir aussi les parenthèses 64 et 65).

3.6 Cas d'un modèle fini

À titre d'illustration, nous démontrons dans cette section une majoration générale du risque de \widehat{f}_S lorsque S est fini et la fonction de coût c bornée.

Proposition 9 Soit $S \subset \mathcal{F}$ un modèle et \widehat{f}_S une règle par minimisation du risque empirique sur S . On suppose que le risque \mathcal{R}_P et le risque empirique $\widehat{\mathcal{R}}_n$ sont définis tous les deux avec le même coût c , et que des constantes $a < b$ existent telles que, pour tout $f \in S$:

$$c(f(X), Y) \in [a, b] \quad \text{presque sûrement.}$$

Alors, pour tout $x \geq 0$ et $n \geq 1$:

$$\mathbb{P}\left(\ell(f^*, \widehat{f}_S(D_n)) < \ell(f^*, S) + (b - a)\sqrt{\frac{2[x + \ln(2 \operatorname{Card} S)]}{n}}\right) \geq 1 - e^{-x}. \quad (14)$$

Démonstration Étant donné la proposition 7, il suffit de majorer :

$$\sup_{f \in S} |\mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f)|.$$

Or, pour $f \in S$ fixé,

$$\widehat{\mathcal{R}}_n(f) - \mathcal{R}_P(f) = \frac{1}{n} \sum_{i=1}^n \left(c(f(X_i), Y_i) - \mathbb{E}[c(f(X_i), Y_i)] \right)$$

est la somme de n variables aléatoires indépendantes, centrées, à valeurs dans un intervalle d'amplitude $(b-a)/n$. D'après l'inégalité de Hoeffding (théorème 5 en section 8), on a donc, pour tout $f \in S$ et tout $z \geq 0$:

$$\mathbb{P}(|\widehat{\mathcal{R}}_n(f) - \mathcal{R}_P(f)| \geq z) \leq 2 \exp\left(\frac{-2z^2 n}{(b-a)^2}\right).$$

Par une borne d'union sur $f \in S$, on en déduit que pour tout $z \geq 0$:

$$\mathbb{P}(|\widehat{\mathcal{R}}_n(f) - \mathcal{R}_P(f)| \geq z) \leq 2 \operatorname{Card}(S) \exp\left(\frac{-2z^2 n}{(b-a)^2}\right)$$

d'où le résultat. \square

En faisant un peu plus attention dans la démonstration, on peut gagner un facteur 2 à l'intérieur du logarithme (exercice 10). On peut également démontrer un résultat similaire lorsque c n'est pas bornée (exercice 11). En particulier, si pour tout $f \in S$,

$$\operatorname{var}(c(f(X), Y)) \leq v,$$

alors pour tout $\delta \in (0, 1]$:

$$\mathbb{P}\left(\ell(f^*, \widehat{f}_S(D_n)) < \ell(f^*, S) + \sqrt{\frac{v \operatorname{Card} S}{n\delta}}\right) \geq 1 - \delta. \quad (15)$$

En comparant (14) avec $x = \ln(1/\delta)$ et (15), on observe que l'on perd beaucoup en ne supposant plus que c est bornée : non seulement la dépendance en δ (la probabilité de l'événement défavorable) passe de $\sqrt{\ln(1/\delta)}$ à $\delta^{-1/2}$, mais surtout la dépendance en $\operatorname{Card}(S)$ (issue d'une borne d'union) passe de $\sqrt{\ln(\operatorname{Card} S)}$ à $\sqrt{\operatorname{Card} S}$.

Si l'on s'intéresse à l'espérance de l'excès de risque, une technique de preuve similaire mène au résultat suivant.

Proposition 10 *Sous les hypothèses de la proposition 9, on a pour tout $n \geq 1$:*

$$\mathbb{E}\left[\ell(f^*, \widehat{f}_S(D_n))\right] \leq \ell(f^*, S) + (b-a)\sqrt{\frac{\ln(\operatorname{Card} S)}{2n}}. \quad (16)$$

Si $f^* \in S$, la proposition 10 montre que l'espérance de l'excès de risque de \hat{f}_S est majorée par $\kappa(S)/\sqrt{n}$, où $\kappa(S)$ ne dépend pas de n . En particulier, si S est fixé (fini) et si n tend vers l'infini, \hat{f}_S est consistant. De plus, cette borne n'est pas améliorable en général (à une constante numérique près), comme le démontre la proposition 17 en section 7.

Il faut cependant interpréter ce résultat avec précaution, car la constante $\kappa(S)$ est de l'ordre de $\sqrt{\ln(\text{Card } S)}$. Ainsi, lorsque S est fini mais « très grand » (par exemple, de cardinal supérieur à e^n), la proposition 10 n'apporte aucune information. C'est pire encore lorsque c n'est pas borné et que l'on suppose uniquement sa variance majorée : la borne (15) sur l'erreur d'estimation n'est « petite » que lorsque $\text{Card}(S) \ll n$.

Comme les résultats de la section 3.5, les propositions 9 et 10 majorent l'excès de risque de \hat{f}_S par une somme de deux termes, qui correspondent chacun à une source d'erreur (approximation, complexité de S). La nouveauté ici est que le terme de complexité a une forme explicite, fonction du cardinal de S uniquement, de l'ordre de :

$$\sqrt{\frac{\ln(\text{Card } S)}{n}}.$$

Parenthèse 63 (Minimiseur approché) Au vu de leur méthode de démonstration (qui s'appuie sur les propositions 7 et 8 ou leur démonstration), les propositions 9 et 10 et l'équation (15) se généralisent au cas de $\hat{f}_{S,\rho}$, un ρ -minimiseur du risque empirique sur S , au prix d'ajouter l'erreur d'optimisation ρ à la borne sur l'excès de risque. Voir les exercices 10, 11 et 12.

Partition finie

On peut appliquer les propositions 9 et 10 au cas des règles par partition finie en classification.

Exemple 6 (Règle de classification par partition finie) On considère une règle par partition $\hat{f}_{\mathcal{A}}^{p-c}$ en classification binaire supervisée (comme à l'exemple 3 en section 2.2), avec une partition \mathcal{A} finie. D'après la proposition 6, $\hat{f}_{\mathcal{A}}^{p-c}$ minimise le risque empirique sur le modèle $S_{\text{class}}^{\text{part}}(\mathcal{A})$, qui est fini de cardinal $2^{\text{Card}(\mathcal{A})}$. Le coût de classification 0–1 étant à valeurs dans $[0, 1]$, les propositions 9 et 10 s'appliquent avec $b - a = 1$. On obtient que pour tout $x \geq 0$:

$$\mathbb{P}\left(\ell(f^*, \hat{f}_{\mathcal{A}}^{p-c}(D_n)) \leq \ell(f^*, S_{\text{class}}^{\text{part}}(\mathcal{A})) + \sqrt{\frac{2[x + \ln(2) \text{Card}(\mathcal{A})]}{n}}\right) \geq 1 - e^{-x}$$

et $\mathbb{E}\left[\ell(f^*, \hat{f}_{\mathcal{A}}^{p-c}(D_n))\right] \leq \ell(f^*, S_{\text{class}}^{\text{part}}(\mathcal{A})) + \sqrt{\frac{\ln(2) \text{Card}(\mathcal{A})}{2n}}.$

Le terme correspondant à la complexité de $S_{\text{class}}^{\text{part}}(\mathcal{A})$ (erreur d'estimation) est ici proportionnel à $\sqrt{\text{Card}(\mathcal{A})/n}$. Il est à noter que $\text{Card}(\mathcal{A})$ est le nombre de « paramètres »

du modèle $S_{\text{class}}^{\text{part}}(\mathcal{A})$: choisir parmi $S_{\text{class}}^{\text{part}}(\mathcal{A})$ revient à choisir une étiquette pour chaque élément de \mathcal{A} . La section 3.7 indique une manière de définir une notion plus générale de « dimension » d'un ensemble de classificateurs, fini ou infini. Notons que l'on peut utiliser ici le résultat de l'exercice 7 pour préciser (ou majorer) ce que vaut l'erreur d'approximation dans les bornes ci-dessus.

3.7 Cas d'un modèle quelconque

Comme en section 3.6, on suppose ici que c est bornée (à valeurs dans $[a, b]$). En revanche, on considère un modèle S qui n'est pas nécessairement fini. Compte-tenu de la proposition 8, il reste à majorer :

$$\mathbb{E} \left[\sup_{f \in S} \{ \mathcal{R}_P(f) - \hat{\mathcal{R}}_n(f) \} \right].$$

Nous présentons ci-dessous une version modernisée de l'approche proposée par Vapnik et Chervonenkis pour traiter le cas de la classification avec le coût 0–1.

Pour simplifier l'exposition, nous considérons ici l'espérance de l'excès de risque (voir l'exercice 21 pour une majoration avec grande probabilité), et nous ne donnons que les grandes lignes de la démonstration. Des explications détaillées sont données par Boucheron *et al.* (2005, section 3).

3.7.1 Symétrisation

La première étape, valable pour un problème de prévision général et un coût quelconque, repose sur l'idée de symétrisation (proposition 20 en section 8.4). Soit $\varepsilon_1, \dots, \varepsilon_n$ des variables de Rademacher¹⁶ indépendantes entre elles et indépendantes de D_n . On a alors :

$$\mathbb{E} \left[\sup_{f \in S} \{ \mathcal{R}_P(f) - \hat{\mathcal{R}}_n(f) \} \right] \leq 2 \mathbb{E} \left[\sup_{f \in S} \left\{ \frac{1}{n} \sum_{i=1}^n \varepsilon_i c(f(X_i), Y_i) \right\} \right]. \quad (17)$$

Définissons la *moyenne de Rademacher* de $\mathcal{B} \subset \mathbb{R}^n$ par :

$$\text{Rad}_n(\mathcal{B}) = \mathbb{E} \left[\sup_{\beta \in \mathcal{B}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \beta_i \mid \mathcal{B} \right], \quad (18)$$

où les variables de Rademacher indépendantes $\varepsilon_1, \dots, \varepsilon_n$ sont indépendantes de \mathcal{B} (si jamais \mathcal{B} est aléatoire). Alors, on peut réécrire (17) comme :

$$\begin{aligned} \mathbb{E} \left[\sup_{f \in S} \{ \mathcal{R}_P(f) - \hat{\mathcal{R}}_n(f) \} \right] &\leq 2 \mathbb{E} [\text{Rad}_n(\mathcal{B}_S(D_n))] \\ \text{avec } \mathcal{B}_S(D_n) &:= \left\{ (c(f(X_i), Y_i))_{1 \leq i \leq n} \mid f \in S \right\}. \end{aligned} \quad (19)$$

¹⁶ Une variable de Rademacher est une variable aléatoire ε de loi uniforme sur $\{-1, 1\}$.

On appelle *complexité de Rademacher* (globale) de S la quantité

$$\text{Rad}_n(\mathcal{B}_S(D_n)) = \mathbb{E} \left[\sup_{f \in S} \left\{ \frac{1}{n} \sum_{i=1}^n \varepsilon_i c(f(X_i), Y_i) \right\} \middle| D_n \right]$$

qui apparaît dans le membre de droite de (19).

Parenthèse 64 (Sur les moyennes de Rademacher) La moyenne de Rademacher $\text{Rad}_n(\mathcal{B})$ est un objet classique qui possède des propriétés intéressantes (Boucheron *et al.*, 2005, théorème 3.3) ; voir aussi l'exercice 24. Notons qu'elle est souvent définie avec une valeur absolue dans le supremum, ce qui correspond à « symétriser » \mathcal{B} et utiliser la définition (18) :

$$\mathbb{E} \left[\sup_{\beta \in \mathcal{B}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \beta_i \right| \middle| \mathcal{B} \right] = \text{Rad}_n(\mathcal{B} \cup (-\mathcal{B})).$$

De même, la complexité de Rademacher est souvent définie avec une valeur absolue dans le supremum.

Parenthèse 65 (Que perd-on avec (17) ou (19)?) En classification 0–1, si S est « symétrique » (pour tout $f \in S$, $1 - f \in S$), alors les inégalités (17) et (19) ne peuvent pas faire perdre plus qu'un facteur 4 par rapport à la borne sur l'erreur d'estimation fournie par la proposition 8 (voir l'exercice 13).

Parenthèse 66 (Lien avec le rééchantillonnage) La complexité de Rademacher (globale) de S peut s'interpréter comme un estimateur par rééchantillonnage de

$$\sup_{f \in S} \{ \mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f) \}$$

(Arlot, 2007, section 1.3.2), qui est précisément la quantité que l'on cherche à majorer ici.

3.7.2 Classification 0–1 : entropie combinatoire empirique

On se place désormais dans le cadre de la classification binaire avec le coût 0–1. Alors,

$$\mathcal{B}_S(D_n) = \{ (\mathbf{1}_{f(X_i) \neq Y_i})_{1 \leq i \leq n} / f \in S \} \subset \{0, 1\}^n$$

est nécessairement fini ! On peut donc appliquer le lemme 4 (énoncé en section 8.2) qui montre que lorsque \mathcal{B} est fini,

$$\text{Rad}_n(\mathcal{B}) \leq \frac{1}{n} \sqrt{2 \ln(\text{Card } \mathcal{B}) \sup_{\beta \in \mathcal{B}} \left\{ \sum_{i=1}^n \beta_i^2 \right\}}. \quad (20)$$

Ici, puisque $\sum_{i=1}^n \beta_i^2 \leq n$ pour tout $\beta \in \mathcal{B}_S(D_n)$, en combinant (19) et (20), on obtient que :

$$\mathbb{E} \left[\sup_{f \in S} \{ \mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f) \} \right] \leq 2\mathbb{E} \left[\sqrt{\frac{2 \ln(\text{Card } \mathcal{B}_S(D_n))}{n}} \right]. \quad (21)$$

En comparant (21) avec la majoration (16) obtenue lorsque S est fini, on observe qu'à un facteur 4 près, tout se passe comme si S était fini de cardinal $\text{Card } \mathcal{B}_S(D_n)$.

Focalisons-nous maintenant sur le cardinal de $\mathcal{B}_S(D_n)$. Première remarque : il ne dépend pas de Y_1, \dots, Y_n . En effet, pour tout $f, g \in S$, le fait que

$$(\mathbf{1}_{f(X_i) \neq Y_i})_{1 \leq i \leq n} = (\mathbf{1}_{g(X_i) \neq Y_i})_{1 \leq i \leq n}$$

est réalisé ou non ne dépend pas des valeurs des Y_i . En prenant $Y_i = 0$, on obtient que :

$$\text{Card}(\mathcal{B}_S(D_n)) = \text{Card} \left\{ (f(X_i))_{1 \leq i \leq n} / f \in S \right\} =: T_S(X_1, \dots, X_n)$$

le cardinal de la « trace » de X_1, \dots, X_n sur S . Le logarithme de $T_S(X_1, \dots, X_n)$ est appelé *entropie combinatoire empirique*¹⁷ de S pour l'échantillon X_1, \dots, X_n :

$$\begin{aligned} H_S(X_1, \dots, X_n) &:= \ln T_S(X_1, \dots, X_n) \\ &= \ln \text{Card} \left\{ (f(X_i))_{1 \leq i \leq n} / f \in S \right\}. \end{aligned} \quad (22)$$

On peut donc réécrire (21) sous la forme :

$$\mathbb{E} \left[\sup_{f \in S} \{ \mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f) \} \right] \leq \frac{2\sqrt{2}}{\sqrt{n}} \mathbb{E} \left[\sqrt{H_S(X_1, \dots, X_n)} \right]. \quad (23)$$

Par conséquent, on a :

$$\mathbb{E} \left[\sup_{f \in S} \{ \mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f) \} \right] \leq \frac{2\sqrt{2}}{\sqrt{n}} \sqrt{\sup_{x_1, \dots, x_n \in \mathcal{X}} \{ H_S(x_1, \dots, x_n) \}}. \quad (24)$$

3.7.3 Classe de Vapnik-Chervonenkis

Avec une hypothèse supplémentaire sur S , on peut obtenir une majoration simple de l'entropie combinatoire, et donc de l'excès de risque moyen en combinant (24) et (13).

Définition 5 (Classe de Vapnik-Chervonenkis) Soit S un ensemble de classifieurs $\mathcal{X} \rightarrow \{0, 1\}$. Pour tout entier $k \geq 1$, on pose :

$$\mathcal{C}(S, k) := \sup_{x_1, \dots, x_k \in \mathcal{X}} T_S(x_1, \dots, x_k) = \sup_{x_1, \dots, x_k \in \mathcal{X}} \text{Card} \left\{ (f(x_i))_{1 \leq i \leq k} / f \in S \right\}.$$

17. Vapnik (2000, section 2.3) utilise le terme « random entropy » pour H_S .

On dit que S est une *classe de Vapnik-Chervonenkis* si :

$$V(S) := \sup \{ k \geq 1 / \mathcal{C}(S, k) = 2^k \} < +\infty.$$

On appelle alors $V(S)$ la *dimension de Vapnik-Chervonenkis* de S .

Pour tout $k \geq 1$ et tout $x_1, \dots, x_k \in \mathcal{X}$, on a $T_S(x_1, \dots, x_k) \leq 2^k$ et donc $\mathcal{C}(S, k) \leq 2^k$. Le cas d'égalité se produit lorsque S est capable d'expliquer parfaitement n'importe quel ensemble d'étiquettes $y_1, \dots, y_k \in \{0, 1\}$ associées à ces k points. On dit alors que S « pulvérise » ou « hâche » $\{x_1, \dots, x_k\}$. Un tel modèle n'a aucune chance de généraliser correctement (on parle de surapprentissage, voir la section 3.9), sauf peut-être dans le cas zéro-erreur. Avoir $T_S(x_1, \dots, x_k) < 2^k$ est donc une bonne chose, et une classe de Vapnik-Chervonenkis est un modèle S qui vérifie ceci pour n'importe quel échantillon de taille $k > V(S)$. La dimension de Vapnik-Chervonenkis est ainsi une mesure du pouvoir séparateur de S (sa capacité à « séparer » deux sous-ensembles quelconques de $x_1, \dots, x_k \in \mathcal{X}$).

Parenthèse 67 (Autre définition) On trouve parfois une autre définition de la dimension de Vapnik-Chervonenkis :

$$\inf \{ k \geq 1 / \mathcal{C}(S, k) < 2^k \}$$

qui est égale à $V(S) + 1$. Nous avons choisi ici la définition la plus habituelle.

Un modèle S fini est toujours une classe de Vapnik-Chervonenkis, de dimension $V(S) \leq \ln_2(\text{Card } S)$. En revanche, on peut trouver des modèles relativement simples qui ne sont pas des classes de Vapnik-Chervonenkis, par exemple

$$S = \{\mathbf{1}_A / A \subset \mathbb{R}^2 \text{ convexe}\}$$

si $\mathcal{X} = \mathbb{R}^2$.

Le lemme de Sauer ci-dessous montre que $T_S(x_1, \dots, x_k)$ est en fait beaucoup plus petit que 2^k pour toute classe de Vapnik-Chervonenkis de dimension $V(S) < k/2$.

Lemme 1 (Lemme de Sauer) Soit S une classe de Vapnik-Chervonenkis de dimension $V(S)$. Alors, pour tout $n \geq 1$, on a :

$$\forall x_1, \dots, x_n \in \mathcal{X}, \quad T_S(x_1, \dots, x_n) \leq \sum_{i=0}^{V(S)} \binom{n}{i}.$$

En particulier, pour tout $n > 2V(S)$, on a :

$$\forall x_1, \dots, x_n \in \mathcal{X}, \quad H_S(x_1, \dots, x_n) \leq V(S) \ln \left(\frac{en}{V(S)} \right).$$

Devroye *et al.* (1996, théorèmes 13.2 et 13.3) démontrent le lemme 1. En combinant (24) avec le lemme de Sauer, on obtient que si S est une classe de Vapnik-Chervonenkis de dimension $V(S)$, alors pour tout $n > 2V(S)$:

$$\mathbb{E} \left[\sup_{f \in S} \{ \mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f) \} \right] \leq 2\sqrt{2} \sqrt{\frac{V(S)}{n} \ln \left(\frac{en}{V(S)} \right)}. \quad (25)$$

Lorsque S est fini, $V(S) \leq \ln_2(\text{Card } S)$ et (25) redonne la borne (16), à un facteur $\ln(n)$ près. En utilisant directement (24) avec la majoration $H_S(x_1, \dots, x_n) \leq \ln \text{Card}(S)$, ce terme logarithmique disparaît et l'on retrouve la borne (16) obtenue en section 3.6, à un facteur 4 près.

L'intérêt de la borne (25) est qu'elle s'applique à bon nombre de modèles infinis classiques, pour lesquels on peut démontrer que S est une classe de Vapnik-Chervonenkis (voir les exercices 15 et 16). Par exemple, si $\mathcal{X} = \mathbb{R}^p$ avec $p \geq 1$,

$$S = \{\mathbf{1}_A / A \text{ demi-espace de } \mathbb{R}^p\}$$

est une classe de Vapnik-Chervonenkis de dimension $V(S) = p + 1$. Cette borne s'applique également à l'analyse de certains réseaux de neurones.

La borne (25) possède cependant plusieurs défauts. D'une part, elle est pessimiste : au vu de l'inégalité (24), il s'agit d'un pire cas sur x_1, \dots, x_n , donc sur la loi P ; si $\mathcal{X} = \mathbb{R}^p$ mais que X_i appartient à un sous-espace vectoriel de \mathbb{R}^p de petite dimension, la borne (25) ne peut pas en tenir compte (voir l'exercice 17). D'autre part, le facteur $\sqrt{\ln(n)}$ de la borne (25) est sous-optimal : sous les mêmes hypothèses, on peut établir par un raisonnement différent que

$$\mathbb{E} \left[\sup_{f \in S} \{ \mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f) \} \right] \leq \kappa \sqrt{\frac{V(S)}{n}}$$

où κ est une constante universelle (Boucheron *et al.*, 2005, section 3). Enfin, la dimension de Vapnik-Chervonenkis $V(S)$ n'est pas toujours facile à calculer.

3.7.4 Récapitulatif

L'analyse de Vapnik des règles de minimisation du risque empirique en classification 0–1 peut se résumer à la chaîne d'inégalités suivante :

$$\begin{aligned}
& \mathbb{E}[\ell(f^*, \hat{f}_S(D_n))] - \ell(f^*, S) \\
& \leq \mathbb{E}\left[\sup_{f \in S}\{\mathcal{R}_P(f) - \hat{\mathcal{R}}_n(f)\}\right] && \text{(proposition 8)} \\
& \leq 2\mathbb{E}[\text{Rad}_n(\mathcal{B}_S(D_n))] && \text{(équation (19))} \\
& \leq \frac{2\sqrt{2}}{\sqrt{n}} \mathbb{E}\left[\sqrt{H_S(X_1, \dots, X_n)}\right] && \text{(équation (23))} \\
& \leq \frac{2\sqrt{2}}{\sqrt{n}} \sqrt{\sup_{x_1, \dots, x_n \in \mathcal{X}} H_S(x_1, \dots, x_n)} && \text{(équation (24))} \\
& \leq 2\sqrt{2} \sqrt{\frac{V(S)}{n} \ln\left(\frac{en}{V(S)}\right)}. && \text{(équation (25))}
\end{aligned}$$

Lorsque le résultat obtenu à la dernière équation n'est pas assez précis (par exemple, si S n'est pas une classe de Vapnik-Chervonenkis), il est souvent intéressant d'utiliser l'un des résultats intermédiaires ci-dessus (par exemple pour les règles par partition, comme cela est détaillé ci-après). Rappelons toutefois que la première inégalité (issue de la proposition 8) est parfois déjà trop large, voire non informative, comme le signale la parenthèse 59 en section 3.5.

Parenthèse 68 (Liens avec la consistance) Vapnik (2000, chapitre 2) justifie les différentes étapes de l'analyse ci-dessus du risque de \hat{f}_S en reliant la « consistance » de \hat{f}_S (en un sens différent de la définition 1) à ces différentes bornes sur l'erreur d'estimation. Vapnik propose deux définitions de la consistance (pour une loi P) d'une règle d'apprentissage \hat{f}_S minimisant le risque empirique sur S . La définition « traditionnelle » est :

$$\mathcal{R}_P(\hat{f}_S(D_n)) \xrightarrow[n \rightarrow +\infty]{(p)} \inf_{f \in S} \mathcal{R}_P(f) \quad \text{et} \quad \hat{\mathcal{R}}_n(\hat{f}_S(D_n)) \xrightarrow[n \rightarrow +\infty]{(p)} \inf_{f \in S} \mathcal{R}_P(f);$$

si $f^* \in S$, ceci implique bien la consistance faible de la définition 1. Pour régler des problèmes apparaissant dans certaines situations particulières, Vapnik définit une notion de « consistance non-triviale », qui implique la première condition de la consistance traditionnelle. Alors, la consistance non-triviale (pour une loi P fixée) est *équivalente* (Vapnik, 2000, théorème 2.1) au fait que

$$\sup_{f \in S}\{\mathcal{R}_P(f) - \hat{\mathcal{R}}_n(f)\} \xrightarrow[n \rightarrow +\infty]{(p)} 0,$$

ce qui montre la finesse de la proposition 8. De plus,

$$\frac{\mathbb{E}[H_S(X_1, \dots, X_n)]}{n} \xrightarrow[n \rightarrow +\infty]{} 0$$

est une condition suffisante de consistance non-triviale pour P , et l'on a une condition nécessaire assez semblable (Vapnik, 2000, théorèmes 2.3 et 2.4) ; on n'a donc pas perdu grand chose avec l'équation (23). Enfin, la consistance non-triviale *universelle* de \hat{f}_S est équivalente à

$$\frac{\sup_{x_1, \dots, x_n \in \mathcal{X}} H_S(x_1, \dots, x_n)}{n} \xrightarrow[n \rightarrow +\infty]{} 0$$

et au fait que S est une classe de Vapnik-Chervonenkis (Vapnik, 2000, sections 2.7 et 3.5) ; les inégalités (24) et (25) ne font donc pas trop perdre par rapport aux résultats précédents quand on s'intéresse à des bornes *universelles*, valables pour toutes les lois P simultanément. Nous renvoyons le lecteur intéressé aux énoncés précis de ces résultats, qui nécessitent quelques hypothèses non mentionnées ici. Notons aussi que ces résultats sont formulés ici pour la classification 0–1 mais sont valables dans un cadre bien plus général.

Malgré les arguments de Vapnik, une analyse « globale » du risque de \hat{f}_S (au sens précisé à la parenthèse 59) n'est pas toujours la plus fine qui soit. Comme on l'a déjà mentionné, pour certaines lois P (en classification, on peut penser au cas zéro-erreur et à la condition de marge, voir les sections 3.8 et 7), la vitesse d'apprentissage réelle de \hat{f}_S est meilleure que celle que l'on peut déduire de la suite de majorations proposée par Vapnik.

On peut aussi déduire de la chaîne d'inégalités ci-dessus des majorations *observables* de l'erreur d'estimation. Si $V(S)$ et

$$\sup_{x_1, \dots, x_n \in \mathcal{X}} H_S(x_1, \dots, x_n)$$

sont calculables en un temps raisonnable, (25) et (24) fournissent des majorations déterministes (disponibles *avant* d'avoir vu les observations D_n). En utilisant (23) et le fait que l'entropie combinatoire empirique se concentre (Boucheron *et al.*, 2013, théorème 6.14), on obtient une majoration fondée sur $H_S(X_1, \dots, X_n)$, qui prend en compte la loi des X_i . En utilisant (19) et le résultat de l'exercice 19, on obtient une majoration fondée sur la complexité de Rademacher globale $\text{Rad}_n(\mathcal{B}_S(D_n))$, qui prend en compte la loi P des observations.

Parenthèse 69 (Lien avec la pénalisation) Disposer de majorations observables de

$$\sup_{f \in S} \{ \mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f) \}$$

ou son espérance est également utile pour construire des pénalités pour le problème de sélection de modèles, comme indiqué en section 3.9.

3.7.5 Cas des règles par partition

Afin d'illustrer les différentes étapes intermédiaires de l'analyse de Vapnik récapitulée ci-dessus, il est intéressant de considérer le cas particulier des règles par partition. On fixe \mathcal{A} une partition mesurable de \mathcal{X} et l'on note $S = S_{\text{class}}^{\text{part}}(\mathcal{A})$ le modèle par partition associé.

La dernière borne (25) n'apporte pas vraiment d'information nouvelle. Si \mathcal{A} est finie et alors $S_{\text{class}}^{\text{part}}(\mathcal{A})$ est fini et l'on retrouve le résultat déjà obtenu en section 3.6 (en un peu moins précis). Si \mathcal{A} est infinie (dénombrable), $S_{\text{class}}^{\text{part}}(\mathcal{A})$ n'est pas une classe de Vapnik-Chervonenkis et (25) ne s'applique donc pas.

Supposons désormais que \mathcal{A} est infinie : en remontant la chaîne d'inégalités établie par l'analyse de Vapnik, peut-on obtenir un résultat sur la règle par partition $\hat{f}_{\mathcal{A}}^{\text{p-c}}$?

La borne (24) reposant sur une majoration uniforme de l'entropie combinatoire empirique n'est pas plus informative. En effet, dès que $\text{Card}(\mathcal{A}) \geq n$, on a :

$$\sup_{x_1, \dots, x_n \in \mathcal{X}} H_{S_{\text{class}}^{\text{part}}(\mathcal{A})}(x_1, \dots, x_n) = n \ln(2).$$

En revanche, l'inégalité (23) permet d'établir que l'erreur d'estimation tend vers zéro lorsque n tend vers l'infini (voir l'exercice 18). En ajoutant des conditions sur la distribution de X , on peut même obtenir des bornes sur la vitesse de décroissance vers zéro de l'erreur d'estimation pour une règle par partition cubique.

Signalons enfin que l'inégalité (19) ne peut pas faire perdre plus qu'un facteur 4 par rapport à la borne sur l'erreur d'estimation fournie par la proposition 8, puisque le modèle $S_{\text{class}}^{\text{part}}(\mathcal{A})$ est « symétrique » (voir l'exercice 13) : pour tout $f \in S_{\text{class}}^{\text{part}}(\mathcal{A})$, on a également $1 - f \in S_{\text{class}}^{\text{part}}(\mathcal{A})$.

3.7.6 Extensions

L'analyse de Vapnik-Chervonenkis permet d'obtenir une majoration avec grande probabilité de l'erreur d'estimation, en partant de la borne fournie par la proposition 7 (voir les exercices 20 et 21).

On peut également l'étendre au cadre de la classification avec un coût asymétrique (exercice 22) ou avec certains coûts convexes (Giraud, 2014, section 9.3.2).

Plus généralement, l'analyse de Vapnik-Chervonenkis s'étend à un cadre très général, incluant le cadre de la prévision lorsque

$$\left\{ c(y, f(x)) / x \in \mathcal{X}, y \in \mathcal{Y}, f \in S \right\}$$

est borné (Vapnik, 2000).

Parenthèse 70 (Pour un minimiseur approché) Puisque l'analyse de Vapnik s'appuie sur la proposition 8 (ou éventuellement la proposition 7), elle se généralise au cas de $\hat{f}_{S,\rho}$, un ρ -minimiseur du risque empirique sur S , au prix d'ajouter l'erreur d'optimisation ρ à chacune des bornes sur l'excès de risque.

3.8 Classification zéro-erreur

En classification, lorsque $\eta(X) \in \{0, 1\}$ presque sûrement, on parle de cas « zéro-erreur ». L'étiquette Y est une fonction déterministe de X et le risque de Bayes est nul. On peut alors obtenir de meilleures bornes sur le risque de \hat{f}_S que dans le cas général.

Lorsque S est fini et $f^* \in S$, on obtient une borne de risque en $1/n$, à comparer à la vitesse en $1/\sqrt{n}$ fournie par les résultats de la section 3.6 (sans hypothèse sur P).

Proposition 11 *On se place en classification binaire avec le coût 0–1. On suppose que P est une loi zéro-erreur, c'est-à-dire que $Y = \eta(X)$ presque sûrement. Soit $S \subset \mathcal{F}$ un modèle fini et \hat{f}_S une règle d'apprentissage minimisant le risque empirique sur S . Alors, si $f^* \in S$, on a pour tout entier $n \geq 1$ et tout $x \geq 0$:*

$$\begin{aligned} \mathbb{P}\left(\ell(f^*, \hat{f}_S(D_n)) < \frac{x + \ln(\text{Card}(S))}{n}\right) &\geq 1 - e^{-x} \\ \text{et} \quad \mathbb{E}\left[\ell(f^*, \hat{f}_S(D_n))\right] &\leq \frac{1 + \ln(\text{Card}(S))}{n}. \end{aligned}$$

Démonstration Puisque $f^* \in S$ et que l'on est dans le cas zéro-erreur :

$$\min_{f \in S} \mathcal{R}_P(f) = \mathcal{R}_P(f^*) = 0.$$

Avec probabilité 1, on a donc $\widehat{\mathcal{R}}_n(\hat{f}_S) = \widehat{\mathcal{R}}_n(f^*) = 0$. Pour tout $\epsilon \geq 0$, on a :

$$\begin{aligned} \mathbb{P}(\mathcal{R}_P(\hat{f}_S) \geq \epsilon) &\leq \mathbb{P}(\exists f \in S / \widehat{\mathcal{R}}_n(f) = 0 \text{ et } \mathcal{R}_P(f) \geq \epsilon) \\ &\leq \sum_{f \in S / \mathcal{R}_P(f) \geq \epsilon} \underbrace{\mathbb{P}(\widehat{\mathcal{R}}_n(f) = 0)}_{=(1-\mathcal{R}_P(f))^n \leq (1-\epsilon)^n} \\ &\leq \text{Card}(S)(1-\epsilon)^n \leq \text{Card}(S)e^{-\epsilon n}. \end{aligned}$$

On en déduit la première inégalité en prenant $\epsilon = [x + \ln \text{Card}(S)]/n$. La deuxième inégalité s'obtient par intégration :

$$\begin{aligned} \mathbb{E}\left[\ell(f^*, \hat{f}_S(D_n))\right] &= \int_0^{+\infty} \mathbb{P}(\ell(f^*, \hat{f}_S(D_n)) \geq t) dt \\ &\leq \int_0^{+\infty} \min\{1, (\text{Card } S)e^{-tn}\} dt \\ &= \frac{\ln \text{Card}(S)}{n} + \text{Card}(S) \int_{(\ln \text{Card}(S))/n}^{+\infty} e^{-tn} dt \\ &= \frac{1 + \ln \text{Card}(S)}{n}. \end{aligned}$$

□

De même qu'en section 3.7, on peut généraliser le résultat de la proposition 11 au cas d'un modèle S infini dès que l'on dispose d'un contrôle uniforme sur son entropie combinatoire empirique.

Proposition 12 *On se place en classification binaire avec le coût 0–1. On suppose que P est une loi zéro-erreur, c'est-à-dire que $Y = \eta(X)$ presque sûrement. Soit $S \subset \mathcal{F}$ un modèle et \hat{f}_S une règle d'apprentissage minimisant le risque empirique sur S . Alors, si $f^* \in S$, on a pour tout entier $n \geq 1$ et tout $x \geq 0$:*

$$\mathbb{P}\left(\ell(f^*, \hat{f}_S(D_n)) < \frac{2}{\ln(2)} \frac{1}{n} \left[x + \ln(2) + \sup_{x_1, \dots, x_{2n} \in \mathcal{X}} H_S(x_1, \dots, x_{2n}) \right]\right) \geq 1 - e^{-x}.$$

Devroye *et al.* (1996, théorème 12.7) démontrent la proposition 12 et donnent des références bibliographiques. Des résultats similaires sous des hypothèses plus générales (incluant notamment le cas $f^* \notin S$) peuvent être obtenus avec des techniques plus avancées (Massart et Nédélec, 2006), voir par exemple la proposition 18 en section 7.

En particulier, si S est une classe de Vapnik-Chervonenkis, on obtient une borne de risque de l'ordre de $V(S) \ln(n)/n$, qui est bien meilleure que la borne en $\sqrt{V(S) \ln(n)/n}$ obtenue en section 3.7. Le fait d'avoir une loi P « zéro-erreur » induit ici un gain important dans la borne de risque. Comme indiqué en section 7, ces bornes sont améliorables en pire cas, et donc cette différence correspond à une certaine réalité : on peut obtenir des vitesses d'apprentissage significativement meilleures dans le cas zéro-erreur.

Signalons pour finir qu'il existe toute une famille d'hypothèses intermédiaires entre le cas zéro-erreur et le cas général (connues sous le nom de « condition de marge »), qui est évoquée en section 7.

3.9 Choix d'un modèle

Les sections précédentes étudient les propriétés de \hat{f}_S en supposant le modèle S donné *a priori*. En pratique, choisir le modèle S le plus approprié est un problème important.

Le problème de choix d'un modèle (ou « sélection de modèles ») est un domaine de recherche en tant que tel, que nous ne pouvons évoquer ici que très brièvement. Un bon survol du domaine est offert par les livres de Bertrand *et al.* (2016), Giraud (2014, chapitre 2) et Massart (2007). On peut également consulter Arlot (2017) à propos de la sélection d'estimateurs, dont la sélection de modèles est un cas particulier.

Problème

Soit $(S_m)_{m \in \mathcal{M}_n}$ une collection (finie ou dénombrable) de modèles. Pour chaque modèle S_m , on suppose donnée $\hat{f}_m = \hat{f}_{S_m}$ une règle d'apprentissage par minimisation du risque empirique. On cherche alors à choisir $\hat{m} = \hat{m}(D_n) \in \mathcal{M}_n$, à l'aide des données uniquement.

Un exemple est donné par le problème de sélection de variables (évoqué à l'exemple 4 en section 3.2) : pour tout $m \subset \{1, \dots, p\}$, \hat{f}_m est l'estimateur des moindres carrés du

modèle linéaire n'utilisant que les variables explicatives X^j dont l'indice j fait partie de m . Choisir un modèle revient alors à choisir un ensemble de variables explicatives.

Objectif

De manière générale, on peut avoir deux sortes d'objectifs en choisissant un modèle. Soit l'on veut que le prédicteur correspondant $\hat{f}_{\hat{m}(D_n)}(D_n)$ ait un risque

$$\mathcal{R}_P(\hat{f}_{\hat{m}(D_n)}(D_n))$$

aussi petit que possible. On parle d'objectif de prévision, ou d'estimation. Sur le plan théorique, deux types de résultats garantissent que \hat{m} est une bonne procédure de sélection de modèles pour la prévision. On dit que \hat{m} est *asymptotiquement optimale* ou *efficace* (« efficient » en anglais) si :

$$\frac{\ell(f^*, \hat{f}_{\hat{m}(D_n)}(D_n))}{\inf_{m \in \mathcal{M}_n} \left\{ \ell(f^*, \hat{f}_m(D_n)) \right\}} \xrightarrow[n \rightarrow +\infty]{\text{p.s.}} 1.$$

À taille d'échantillon n fixée, on dit que \hat{m} vérifie une *inégalité oracle* lorsque :

$$\ell(f^*, \hat{f}_{\hat{m}(D_n)}(D_n)) \leq C_n \inf_{m \in \mathcal{M}_n} \left\{ \ell(f^*, \hat{f}_m(D_n)) \right\} + R_n$$

a lieu en espérance ou avec grande probabilité¹⁸, avec $C_n \geq 1$ et $R_n = o(1)$ « petits ». Un élément m^* de \mathcal{M}_n qui réalise l'infimum de l'excès de risque ci-dessus (s'il existe) est appelé « oracle », d'où le nom « inégalité oracle »¹⁹.

Soit l'on veut identifier « le vrai modèle », c'est-à-dire qu'on suppose que f^* appartient à S_{m^*} un « petit » modèle appartenant à la collection $(S_m)_{m \in \mathcal{M}_n}$, et l'on cherche à retrouver S_{m^*} . On souhaite alors comprendre les mécanismes du phénomène étudié. On parle d'objectif d'identification. Sur le plan théorique, on cherche à établir que

$$\mathbb{P}(\hat{m}(D_n) = m^*)$$

est proche de 1 (à n fixé), ou tend vers 1 lorsque n tend vers l'infini (consistance en sélection).

En apprentissage statistique, le plus souvent, les modèles considérés ne prétendent pas décrire exactement le mécanisme générant les données. Les modèles ne sont alors qu'un moyen pratique pour construire des prédicteurs. C'est pourquoi nous nous focalisons désormais sur l'objectif de prévision.

18. On dit qu'un événement est de grande probabilité lorsque sa probabilité est proche de 1. Dans ce cas précis, on cherche par exemple à minorer la probabilité d'avoir une inégalité oracle par $1 - Ln^{-\alpha}$, où $L, \alpha > 0$ sont des constantes absolues, ou bien par $1 - e^{-x}$ où x est un réel pouvant être choisi arbitrairement grand (le terme R_n dépendant alors de x).

19. La notion d'inégalité oracle a été proposée par Donoho et Johnstone (1994) comme une manière d'évaluer la performance d'une procédure de sélection de variables, différente de l'approche minimax (qui se focalise sur le risque en pire cas, voir la section 7).

Remarque 71 (Prévision et interprétation de \hat{m}) Si l'on choisit \hat{m} avec un objectif de prévision, il faut faire attention à ne pas surinterpréter le modèle sélectionné $S_{\hat{m}}$. Ce n'est pas forcément « le vrai modèle », ne serait-ce que parce qu'il se peut qu'aucun modèle parmi $(S_m)_{m \in \mathcal{M}_n}$ ne soit correct. Et même s'il existe un vrai modèle S_{m^*} (contenant f^*), il est tout-à-fait possible que $S_{\hat{m}}$ soit faux !

Considérons le cas de la sélection de variables, en régression avec le coût quadratique, dans un cadre simple :

$$\forall \mathbf{x} \in \mathcal{X} = \mathbb{R}^3, \quad \eta(\mathbf{x}) = w_1 x_1 + w_2 x_2 + w_3 x_3$$

avec $w_1 = 10$ $w_2 = \frac{1}{10}$ $w_3 = 10^{-10}$,

un niveau de bruit $\text{var}(\varepsilon | X) = 1$ et des variables explicatives X^1 , X^2 et X^3 indépendantes, centrées et de même variance égale à 1. Le « vrai modèle » est alors $S_{m^*} = S^{\text{lin}}$, celui qui prend en compte les trois variables. Mais si n n'est pas gigantesque, il est impossible d'évaluer correctement l'influence de la troisième variable X^3 ; alors, le modèle $S_{\{1,2\}}^{\text{lin}}$ qui ne prend en compte que les deux premières variables fournit de meilleures prévisions que le vrai modèle S^{lin} . En effet, tant que l'on ne dispose pas de suffisamment d'observations, il vaut mieux ignorer l'influence (minime) de la variable X^3 pour se concentrer sur l'estimation précise des paramètres w_1 et w_2 . Et si n est petit, pour une raison similaire, le modèle $S_{\{1\}}^{\text{lin}}$ qui ne prend en compte que la première variable X^1 donne souvent les meilleures prévisions. En résumé, on a un exemple où le modèle oracle S_{m^*} est égal (avec grande probabilité) à $S_{\{1\}}^{\text{lin}}$ pour de petites valeurs de n , puis à $S_{\{1,2\}}^{\text{lin}}$ pour des valeurs de n moyennes à grandes, et n'est égal à $S_{m^*} = S^{\text{lin}}$ que pour de très grandes valeurs de n . La figure 4 ci-après donne un autre exemple (moins caricatural) où le meilleur modèle (pour le risque) ne contient pas f^* .

Il y a encore une autre difficulté d'interprétation pour un modèle $S_{\hat{m}}$ sélectionné avec un objectif de prévision. Celui-ci peut s'avérer (légèrement) plus complexe que nécessaire. Par exemple, en sélection de variables, incorporer quelques variables « inutiles » (dont f^* ne dépend pas) n'augmente pas beaucoup le risque de $\hat{f}_{\hat{m}}$. Dès lors, une procédure \hat{m} peut très bien être asymptotiquement optimale pour la prévision tout en incluant quelques variables inutiles avec une probabilité non-nulle (asymptotiquement).

Parenthèse 72 (Collection de modèles variant avec n) La collection de modèles $(S_m)_{m \in \mathcal{M}_n}$ est autorisée à varier avec la taille de l'échantillon. Ceci correspond au fait que plus on a d'observations disponibles, plus on peut s'autoriser à considérer des modèles complexes. Par exemple, en régression linéaire (exemple 4), avec peu d'observations, il est naturel de ne considérer que quelques variables explicatives potentielles (les plus vraisemblables) ; en revanche, si l'on dispose d'un grand échantillon, on peut légitimement chercher à prendre en compte des variables peu influentes, en augmentant le nombre de variables explicatives potentielles.

Obtenir une règle asymptotiquement optimale lorsque \mathcal{M}_n varie avec n est beaucoup plus difficile que si l'on a une collection $(S_m)_{m \in \mathcal{M}}$ finie et fixe quand

n tend vers l'infini. En effet, avec une collection de modèles fixe, le surapprentissage est (asymptotiquement) impossible ! Ainsi, minimiser le risque empirique — c'est-à-dire, la procédure (26) avec $\text{crit}(m; D_n) = \widehat{\mathcal{R}}_n(\widehat{f}_m(D_n))$ — est asymptotiquement optimal dès que

$$\forall m \in \mathcal{M}, \quad \sup_{f \in S_m} |\widehat{\mathcal{R}}_n(f) - \mathcal{R}_P(f)| \xrightarrow[n \rightarrow +\infty]{} 0.$$

C'est en particulier le cas en classification 0–1 si S_m est une classe de Vapnik-Chervonenkis, d'après les résultats de la section 3.7. Il n'y a alors même pas besoin de pénaliser le risque empirique ! En revanche, si la collection de modèles grandit avec n , le surapprentissage devient possible et il est nécessaire de pénaliser le risque empirique (voir la parenthèse 74 à la fin de cette section).

Enjeux du problème : éviter surapprentissage et sous-apprentissage

Il y a deux écueils principaux à éviter quand on choisit un modèle. Le *surapprentissage* (ou sur-ajustement ; « overfitting » en anglais) se produit lorsque l'on choisit un modèle S_m beaucoup trop complexe (trop grand) compte-tenu du nombre d'observations disponibles. Alors, la richesse de ce modèle fait qu'il est possible d'y trouver un prédicteur \widehat{f}_m qui commet très peu d'erreurs sur les observations, sans aucune garantie sur ses capacités de *généralisation* (c'est-à-dire sur son risque $\mathcal{R}(\widehat{f}_m)$) : son erreur d'estimation peut être très grande. Le prédicteur \widehat{f}_m surapprend car il « apprend par cœur » le jeu de données (erreurs de mesure comprises), sans essayer de généraliser.

L'archétype du surapprentissage est lorsque $S_m = \mathcal{F}$ l'ensemble de tous les prédicteurs. Pour simplifier, on suppose que les X_i sont distincts et \mathcal{X} est infini. Alors, il existe une infinité de prédicteurs $f \in \mathcal{F}$ tels que $f(X_i) = Y_i$ pour tout $i \in \{1, \dots, n\}$, et ces prédicteurs ont tous un risque empirique nul. Tout prédicteur $\widehat{f}_m = \widehat{f}_{\mathcal{F}}$ minimisant le risque empirique sur \mathcal{F} est donc parfait en apparence. Or, certains de ces prédicteurs sont très mauvais : par exemple, si l'on est en classification 0–1, le prédicteur défini par $f(x) = 1 - f^*(x)$ pour tout $x \notin \{X_1, \dots, X_n\}$ et $f(X_i) = Y_i$ pour $i \in \{1, \dots, n\}$ est quasiment le pire prédicteur possible²⁰, alors que c'est un minimiseur du risque empirique sur \mathcal{F} . Sans information supplémentaire, \widehat{f}_m peut donc être très mauvais et il faut éviter de choisir un tel modèle.

L'écueil opposé, que l'on appelle *sous-apprentissage* par analogie (« underfitting » en anglais), se produit quand on choisit un modèle S_m beaucoup trop simple au regard de la complexité du phénomène étudié. Alors, l'erreur d'approximation de S_m est grande, et cela suffit à rendre \widehat{f}_m totalement inopérant. L'archétype du sous-apprentissage est lorsque S_m est l'ensemble des prédicteurs constants sur \mathcal{X} (un cas particulier de modèle par partition). Hormis le cas où f^* est constant ou presque, le risque de \widehat{f}_m est alors largement plus élevé que le risque de Bayes, simplement car \widehat{f}_m doit réaliser une prévision de Y sans utiliser la valeur de X .

Une fois ces deux écueils évités, la sélection de modèles vise à choisir un bon modèle de complexité « intermédiaire » — idéalement, le meilleur. Au vu de la décomposi-

20. Il atteint la pire valeur possible pour le risque lorsque la loi de X est sans atome.

tion (11) de l'excès de risque de \hat{f}_m , il s'agit de réaliser un *compromis entre erreur d'approximation et erreur d'estimation*, aussi appelé *compromis biais-variance*²¹.

Une bonne manière de visualiser ce compromis est de tracer l'erreur d'approximation, l'espérance de l'erreur d'estimation et leur somme (l'excès de risque moyen) en fonction de la complexité de S . Un exemple d'un tel graphe est proposé à la figure 4.

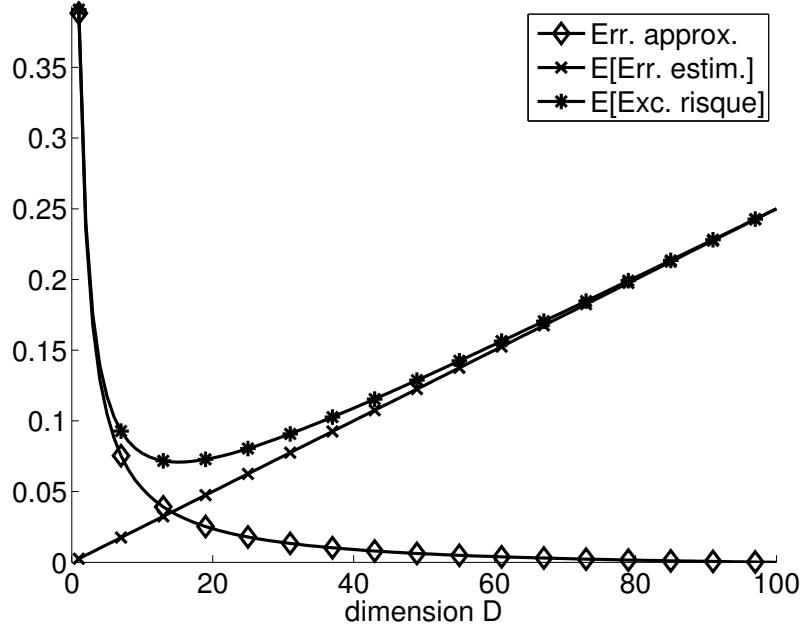


FIGURE 4 – *Erreur d'approximation, espérance de l'erreur d'estimation et espérance de l'excès de risque en fonction de la dimension D des modèles. Données simulées en régression sur un plan d'expérience fixe, avec un coût quadratique : on observe $\mathbf{y} = \mathbf{f}^* + \boldsymbol{\varepsilon} \in \mathbb{R}^n$ avec $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$, on propose $\hat{\mathbf{f}} = \hat{\mathbf{f}}(\mathbf{y}) \in \mathbb{R}^n$ pour « reconstruire » le signal \mathbf{f}^* , et l'on mesure l'excès de risque par $n^{-1} \|\mathbf{f}^* - \hat{\mathbf{f}}\|_2^2$. Pour chaque $D \in \{1, \dots, n\}$, on considère l'estimateur des moindres carrés sur le modèle engendré par les D premiers vecteurs de la base canonique de \mathbb{R}^n . Ici, on a pris $n = 100$, $\sigma^2 = 1/4$ et $f_i^* \propto 1/i^2$ (avec une renormalisation pour avoir $n^{-1} \|\mathbf{f}^*\|_2^2 = 1$).*

Une autre option est d'étudier théoriquement ce que vaut l'excès de risque de \hat{f}_S (ou des majorations de celui-ci, par exemple celles obtenues en section 3.7). Prenons l'exemple des règles par partition cubique de $\mathcal{X} = [0, 1]$ en régression avec le coût quadratique (exemple 2), en supposant que X est de loi uniforme sur $[0, 1]$, η est de classe C^1 et la variance résiduelle $\text{var}(Y | X) = \sigma^2$ ne dépend pas de X (homoscédasticité).

21. L'origine de la terminologie « compromis biais-variance » est expliquée par la parenthèse 86 en section 5.2.

L'exercice 6 donne alors l'ordre de grandeur de l'erreur d'approximation, tandis que l'exercice 8 et la remarque 105 qui suit son énoncé donnent l'ordre de grandeur de l'erreur d'estimation. On en déduit que si $h_n \rightarrow 0$ et $nh_n \rightarrow +\infty$,

$$\mathbb{E}[\ell(f^*, \widehat{f}_h^{\text{cub-r}}(D_n))] \sim \frac{h_n^2}{12} \int_0^1 (\eta'(x))^2 dx + \frac{\sigma^2}{nh_n}.$$

Minimiser en h_n cet équivalent nécessite donc de trouver un compromis entre l'erreur d'approximation (proportionnelle à h_n^2) et l'erreur d'estimation (proportionnelle à $1/(nh_n)$). Ce compromis est réalisé pour $h_n \propto n^{-1/3}$, la constante multiplicative dépendant du niveau de bruit σ^2 et de la régularité de η , via $\|\eta'\|_2^2$. Notons que l'ordre de grandeur optimal de h_n serait différent si η était moins régulière, ce qui rend le problème de sélection de modèles difficile : le modèle optimal dépend en général fortement de la loi P (inconnue) des observations.

Méthodes

La plupart des méthodes de sélection de modèles sont de la forme :

$$\widehat{m} \in \operatorname{argmin}_{m \in \mathcal{M}_n} \{\operatorname{crit}(m; D_n)\} \quad (26)$$

où $\operatorname{crit} : \mathcal{M}_n \rightarrow \mathbb{R}$ (le critère de sélection de modèles, qui est en général fonction des données D_n) estime ou majore le risque $\mathcal{R}_P(\widehat{f}_m(D_n))$ pour tout $m \in \mathcal{M}_n$. Le résultat suivant éclaire sur les bonnes manières de construire ce critère.

Lemme 2 (Lemme fondamental de l'apprentissage) *Soit \mathcal{E} un ensemble et $\mathcal{C}, \mathcal{R}, A, B$ des applications $\mathcal{E} \rightarrow \mathbb{R}$ telles que :*

$$\forall x \in \mathcal{E}, \quad -A(x) \leq \mathcal{C}(x) - \mathcal{R}(x) \leq B(x). \quad (27)$$

Alors, pour tout $\widehat{x} \in \operatorname{argmin}_{x \in \mathcal{E}} \mathcal{C}(x)$, on a :

$$\mathcal{R}(\widehat{x}) - A(\widehat{x}) \leq \inf_{x \in \mathcal{E}} \{\mathcal{R}(x) + B(x)\}.$$

Démonstration Pour tout $x \in \mathcal{E}$, on a :

$$\begin{aligned} \mathcal{R}(\widehat{x}) &= \underbrace{\mathcal{C}(\widehat{x})}_{\leq \mathcal{C}(x)} + \underbrace{\mathcal{R}(\widehat{x}) - \mathcal{C}(\widehat{x})}_{\leq A(\widehat{x})} \\ &\leq \mathcal{C}(x) + A(\widehat{x}) \\ &= \mathcal{R}(x) + \underbrace{\mathcal{C}(x) - \mathcal{R}(x)}_{\leq B(x)} + A(\widehat{x}) \\ &\leq \mathcal{R}(x) + B(x) + A(\widehat{x}). \end{aligned}$$

Le résultat s'en déduit en prenant l'infimum sur $x \in \mathcal{E}$. \square

On peut aisément démontrer une version plus générale de ce résultat (voir l'exercice 23).

Parenthèse 73 (Pourquoi le lemme 2 est-il fondamental?) Nous appelons « lemme fondamental de l'apprentissage » le lemme 2 car celui-ci est la pierre angulaire de quasiment toutes les analyses théoriques de procédures de sélection de modèles définies par (26), ainsi que de l'analyse des règles par minimisation du risque empirique, que ce soit par la méthode « globale » de Vapnik ou par la méthode « locale » plus fine (voir les parenthèses 57 et 59 en section 3.5). Plus généralement, ce lemme peut servir à étudier n'importe quelle méthode d'apprentissage définie comme solution d'un problème d'optimisation.

Pour appliquer le lemme 2 à l'analyse d'une procédure de sélection de modèles de la forme (26), on pose :

$$\forall m \in \mathcal{E} = \mathcal{M}_n, \quad \mathcal{R}(m) = \mathcal{R}_P(\hat{f}_m(D_n)) \quad \text{et} \quad \mathcal{C}(m) = \text{crit}(m; D_n).$$

Ensuite, on distingue deux grandes familles de procédures, chacune s'analysant théoriquement via un choix particulier pour les fonctions A et B .

Une première famille de procédures est de la forme (26) avec

$$\text{crit}(m; D_n) \approx \mathcal{R}_P(\hat{f}_m(D_n)) \quad \text{pour tout} \quad m \in \mathcal{M}_n.$$

Ceci correspond notamment au *principe d'estimation sans biais du risque*²², selon lequel on doit prendre un critère tel que :

$$\mathbb{E}[\text{crit}(m; D_n)] = \mathbb{E}[\mathcal{R}_P(\hat{f}_m(D_n))]$$

(exactement ou approximativement) pour tout $m \in \mathcal{M}_n$. Alors, si le critère et le risque se concentrent suffisamment autour de leur espérance, on obtient que, pour chacun des $m \in \mathcal{M}_n$, le critère $\text{crit}(m; D_n)$ et le risque $\mathcal{R}_P(\hat{f}_m(D_n))$ sont « proches » avec grande probabilité. Si de plus \mathcal{M}_n n'est pas « trop grande », une borne d'union²³ permet d'en déduire qu'avec grande probabilité, $\text{crit}(m; D_n)$ et $\mathcal{R}_P(\hat{f}_m(D_n))$ sont « proches » *simultanément* pour tous les $m \in \mathcal{M}_n$. Par exemple, AIC (critère d'information d'Akaike) et C_p (dû à Mallows) reposent sur ce principe, de même que la validation croisée (Arlot et Celisse, 2010; Arlot, 2017). Précisons ce que signifie que le critère et le risque sont « proches » en regardant ce que le lemme 2 nous dit d'une procédure de cette famille. Si la condition (27) est vérifiée²⁴ avec

$$A(m) = B(m) \leq \epsilon_1 \ell(f^*, \hat{f}_m(D_n)) + \epsilon_2 \tag{28}$$

22. Le principe d'estimation sans biais du risque est aussi appelé « heuristique de Mallows » ou « heuristique d'Akaike », en référence aux travaux qui ont mené aux critères C_p et AIC, respectivement. En tout rigueur, il faudrait parler ici d'estimation du risque *moyen*, le risque étant une quantité aléatoire. On conserve cette terminologie dans la suite par abus de langage.

23. On a utilisé une borne d'union pour démontrer la proposition 9 en section 3.6.

24. En l'occurrence, c'est plutôt la condition plus générale (63) de l'exercice 23 qu'il est possible d'établir ici, la conclusion restant la même.

pour des constantes $\epsilon_1 \in]0, 1]$ et $\epsilon_2 \geq 0$, alors le lemme 2 fournit l'inégalité oracle :

$$\ell(f^*, \hat{f}_m(D_n)) \leq \frac{1 + \epsilon_1}{1 - \epsilon_1} \inf_{m \in \mathcal{M}_n} \left\{ \ell(f^*, \hat{f}_m(D_n)) \right\} + \frac{2\epsilon_2}{1 - \epsilon_1}. \quad (29)$$

De plus, si (28) a lieu avec

$$\epsilon_1 = o(1) \quad \text{et} \quad \epsilon_2 \ll \inf_{m \in \mathcal{M}_n} \left\{ \ell(f^*, \hat{f}_m(D_n)) \right\} \quad (30)$$

lorsque n tend vers l'infini, l'inégalité oracle (29) est optimale au premier ordre, c'est-à-dire qu'elle entraîne que \hat{m} est asymptotiquement optimale. Ces conditions (27), (28) et (30) sont non seulement suffisantes pour obtenir une inégalité oracle optimale au premier ordre, mais surtout elles expliquent comment un grand nombre de procédures de sélection de modèles sont construites. En particulier, elles soulignent une limitation majeure du principe d'estimation sans biais du risque : il s'applique en général uniquement à des collections $(S_m)_{m \in \mathcal{M}_n}$ « petites » (de cardinal majoré par un polynôme en n), pour lesquelles il est possible d'avoir (28) *simultanément* pour tous les $m \in \mathcal{M}_n$, avec ϵ_1 et ϵ_2 assez petits. À l'inverse, les collections « exponentielles » (de cardinal de l'ordre de $e^{\beta n}$ avec $\beta > 0$) nécessitent l'utilisation d'autres méthodes (Birgé et Massart, 2007).

Une deuxième famille de procédures est donnée par (26) avec

$$\text{crit}(m; D_n) \geq \mathcal{R}_P(\hat{f}_m(D_n)) \quad (31)$$

pour tout $m \in \mathcal{M}_n$ (à une constante additive près). Alors, la condition (27) est vérifiée avec

$$A(m) = 0 \quad \text{et} \quad B(m) = \text{crit}(m; D_n) - \mathcal{R}_P(\hat{f}_m(D_n))$$

et l'on obtient :

$$\ell(f^*, \hat{f}_m(D_n)) \leq \inf_{m \in \mathcal{M}_n} \left\{ \ell(f^*, \hat{f}_m(D_n)) + \text{crit}(m; D_n) - \mathcal{R}_P(\hat{f}_m(D_n)) \right\}.$$

Ceci constitue une inégalité-oracle si la majoration (31) n'est pas trop large²⁵.

Par exemple, les procédures construites pour de grandes familles de modèles $(S_m)_{m \in \mathcal{M}_n}$ vérifient généralement (31). C'est notamment le cas des procédures de sélection de variables en grande dimension par « pénalisation L^0 » (Giraud, 2014, chapitre 2), qui s'apparentent alors au critère BIC (critère d'information bayésien). La méthode de « minimisation structurelle du risque », issue de l'analyse de Vapnik présentée en section 3.7, passe également par l'obtention de (31), comme on le détaille à la fin du paragraphe suivant.

Le lecteur intéressé par les procédures vérifiant (31) peut consulter l'article de survol de Arlot et Celisse (2010, section 3.2) pour un exposé bref (mais plus détaillé qu'ici) à leur sujet.

25. Il suffit en fait que cette majoration soit fine pour les « bons » modèles.

Pénalisation

Une partie importante des procédures de sélection de modèles de la forme (26) utilisent un critère empirique pénalisé :

$$\text{crit}(m; D_n) = \widehat{\mathcal{R}}_n(\widehat{f}_m(D_n)) + \text{pen}(m; D_n) \quad (32)$$

où $\text{pen} : \mathcal{M}_n \rightarrow \mathbb{R}$ (souvent fonction des données D_n) est une « pénalité » qui vise à compenser l'optimisme du risque empirique de $\widehat{f}_m(D_n)$ comme estimateur du risque $\mathcal{R}_P(\widehat{f}_m(D_n))$. Cet optimisme est en général plus grand pour les modèles plus complexes : la pénalité est alors une fonction croissante de la « complexité » de S_m .

L'objectif étant de minimiser le risque, une pénalité idéale serait :

$$\text{pen}_{\text{id}}(m; D_n) := \mathcal{R}_P(\widehat{f}_m(D_n)) - \widehat{\mathcal{R}}_n(\widehat{f}_m(D_n)).$$

Celle-ci est bien sûr inconnue (car fonction de P), mais elle éclaire les deux stratégies classiques pour construire une bonne pénalité.

En effet, si l'on suit le principe d'estimation sans biais du risque, alors on fait en sorte d'avoir $\text{pen}(m; D_n)$ proche de $\text{pen}_{\text{id}}(m; D_n)$ ou de son espérance pour tout $m \in \mathcal{M}_n$. Ceci conduit aux critères AIC et C_p , ainsi qu'aux pénalités covariance²⁶ (Efron, 2004).

La deuxième approche, où l'on utilise un critère qui majore le risque, revient ici à chercher une pénalité telle que, avec grande probabilité :

$$\forall m \in \mathcal{M}_n, \quad \text{pen}(m; D_n) \geq \text{pen}_{\text{id}}(m; D_n).$$

Or, puisque $\widehat{f}_m(D_n) \in S_m$, on a toujours :

$$\text{pen}_{\text{id}}(m) \leq \text{pen}_{\text{id},g}(m) := \sup_{f \in S_m} \{\mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f)\}. \quad (33)$$

On nomme ce majorant la « pénalité idéale globale »²⁷. Or, l'analyse de Vapnik-Chervonenkis présentée en section 3.7 établit des majorations observables de $\text{pen}_{\text{id},g}(m)$. On peut donc utiliser la complexité de Rademacher (globale), l'entropie combinatoire empirique ou la dimension de Vapnik-Chervonenkis pour construire des pénalités. Cette approche, proposée par Vapnik et Chervonenkis, est appelée minimisation structurelle du risque (« structural risk minimization » en anglais).

Parenthèse 74 (Pourquoi pénaliser le risque empirique ?) Que se passe-t-il si l'on minimise le risque empirique de $\widehat{f}_m(D_n)$ sans le pénaliser ? Prenons l'exemple d'une collection de modèles S_m emboîtés. Alors,

$$\widehat{\mathcal{R}}_n(\widehat{f}_m(D_n)) = \inf_{f \in S_m} \{\widehat{\mathcal{R}}_n(f)\}$$

26. Les pénalités covariance généralisent C_p , la dimension de S_m comme espace vectoriel étant remplacée par la notion plus générale de nombre de degrés de liberté (« degrees of freedom » en anglais).

27. Le nom « pénalité idéale globale » correspond au fait que $\text{pen}_{\text{id},g}(m)$ majore la pénalité idéale à l'aide d'une quantité qui considère S_m « globalement » (au sens de la parenthèse 59 en section 3.5).

est une fonction décroissante de S_m , si bien que minimiser le risque empirique conduit à toujours sélectionner le plus grand modèle. Ceci conduit en général au surapprentissage (sauf dans certains cas particuliers où il est impossible de surapprendre, voir la parenthèse 72), d'où la nécessité de pénaliser.

Un autre exemple instructif est donné par certains estimateurs des moindres carrés (en régression ou en estimation de densité, voir Arlot (2017, parenthèse 8 en section 3.2)) pour lesquels on a :

$$\begin{aligned}\mathbb{E}[\mathcal{R}_P(\hat{f}_m(D_n)) - \mathcal{R}_P^*] &\approx \text{Erreur d'approximation} + \mathbb{E}[\text{Erreur d'estimation}] \\ \mathbb{E}[\widehat{\mathcal{R}}_n(\hat{f}_m(D_n)) - \mathcal{R}_P^*] &\approx \text{Erreur d'approximation} - \mathbb{E}[\text{Erreur d'estimation}].\end{aligned}$$

L'optimisme du risque empirique (comme estimateur du risque) se traduit sur l'erreur d'estimation, qui est prise en compte à l'opposé de sa contribution à l'excès de risque. Par conséquent, minimiser le risque empirique favorise les modèles les plus complexes (ceux dont l'erreur d'estimation est la plus grande), ce qui conduit au surapprentissage lorsque l'erreur d'estimation n'est pas négligeable devant l'erreur d'approximation. La pénalisation compense cet optimisme du risque empirique et permet de prendre en compte comme il faut l'erreur d'estimation.

4 Coûts convexes en classification

La minimisation du risque empirique en classification avec le coût 0–1 se heurte à une difficulté majeure, d'ordre algorithmique : sauf exception²⁸, cette approche nécessite de résoudre des problèmes d'optimisation combinatoire difficiles (voir par exemple la remarque 52 en section 3.3).

Pour contourner cette difficulté, une idée naturelle est la suivante. On commence par estimer la fonction de régression par minimisation du risque empirique²⁹ avec le coût quadratique, sur un modèle S convexe. Comme il s'agit d'un problème d'optimisation convexe, on dispose d'algorithmes efficaces pour le résoudre. Enfin, on en déduit un classifieur par plug-in, comme défini en section 2.3. La méthode d'analyse présentée en section 3 fournit alors une borne sur le risque quadratique d'estimation de la fonction de régression (voir l'exercice 24), et la proposition 4 permet d'en déduire une borne sur le risque 0–1 en classification.

Toutefois, une telle approche est discutable car le coût quadratique n'est pas forcément adapté à l'objectif final de classification. Par exemple, en un point $x \in \mathcal{X}$ tel que $\eta(x) > 1/2$, surestimer $\eta(x)$ ne devrait pas coûter autant que sous-estimer $\eta(x)$. On aimerait donc pouvoir remplacer le coût quadratique par une autre fonction de coût (convexe). Sous quelles conditions une telle approche est-elle sensée, du point de vue de l'objectif de classification ? Et peut-on généraliser la proposition 4 à de telles fonctions de coût ?

Cette section répond à ces deux questions. Pour aller plus loin, le lecteur intéressé

28. Les règles par partition se calculent très facilement en classification 0–1, par exemple.

29. On peut aussi utiliser un risque empirique régularisé, comme en régression ridge ou lasso.

est invité à consulter les articles de Bartlett *et al.* (2006) et Boucheron *et al.* (2005, section 4.2).

4.1 Pseudo-classificateurs et Φ -risque

On suppose, jusqu'à la fin de la section 4, que $\mathcal{Y} = \{-1, 1\}$. À une bijection près, cette convention est équivalente à la convention $\mathcal{Y} = \{0, 1\}$ prise dans le reste de ce texte (voir aussi la remarque 29 en section 2.2). On choisit ici la convention la plus classique et la plus naturelle pour définir le Φ -risque. Signalons que la fonction de régression η s'écrit alors :

$$\eta(X) = \mathbb{E}[Y | X] = 2\zeta(X) - 1 \quad \text{où} \quad \zeta(X) := \mathbb{P}(Y = 1 | X).$$

La probabilité conditionnelle ζ d'avoir l'étiquette 1 ne coïncide donc plus avec la fonction de régression.

On appelle « pseudo-classifieur »³⁰ une fonction mesurable $g : \mathcal{X} \rightarrow \overline{\mathbb{R}}$, où $\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, +\infty\}$. On note $\overline{\mathcal{F}}$ l'ensemble des pseudo-classificateurs. On peut remarquer que si $f \in \mathcal{F}$, c'est-à-dire si f est une application mesurable $\mathcal{X} \rightarrow \{-1, 1\}$, alors f est également un pseudo-classifieur : $\mathcal{F} \subset \overline{\mathcal{F}}$. Dans l'autre sens, si l'on dispose d'un pseudo-classifieur $g \in \overline{\mathcal{F}}$, on lui associe de manière canonique³¹ le classifieur :

$$\text{signe}(g) : x \in \mathcal{X} \mapsto \begin{cases} 1 & \text{si } g(x) > 0 \\ -1 & \text{sinon.} \end{cases}$$

Soit $\Phi : \overline{\mathbb{R}} \rightarrow [0, +\infty]$ une fonction mesurable, ne pouvant prendre la valeur $+\infty$ qu'à l'infini (hypothèse que l'on suppose désormais toujours vérifiée). La fonction de coût associée à Φ est définie par :

$$c_\Phi : (y, y') \in \mathbb{R}^2 \mapsto \Phi(yy').$$

Le risque associé à Φ (appelé Φ -risque), pour un pseudo-classifieur $g \in \overline{\mathcal{F}}$ et une distribution P sur $\mathcal{X} \times \{-1, 1\}$, est alors défini par :

$$\mathcal{R}_P^\Phi(g) := \mathbb{E}[\Phi(Yg(X))]$$

où l'on rappelle que (X, Y) est de loi P .

Remarque 75 (Interprétation de $g(x)$) La valeur $g(x)$ d'un pseudo-classifieur g en un point $x \in \mathcal{X}$ n'est pas nécessairement une estimation de la valeur $\eta(x)$ prise par la fonction de régression. En général, on interprète seulement $g(x)$ comme un « niveau de confiance » que l'on a en la prévision faite par le classifieur $\text{signe}(g)$ en x . Plus $|g(x)|$ est grand, plus ce niveau de confiance est élevé. Pour que cette interprétation soit cohérente avec le fait que l'on cherche à minimiser le Φ -risque de g , il faut prendre

30. La valeur d'un pseudo-classifieur est souvent appelée « score de classification ». En anglais, on parle de « margin classifier » ; voir aussi la remarque 75.

31. Cette association canonique est l'équivalent de l'approche par plug-in définie en section 2.3.

Φ décroissante, ce qui est le cas de presque tous les exemples classiques. Disposer d'une telle information, et pas seulement du classifieur $\text{signe}(g)$, est souvent très utile, notamment dans des applications où la classification aide à prendre une décision lourde de conséquences (détection de piétons par une voiture autonome, aide au diagnostic médical, etc.) ; voir aussi à ce sujet la parenthèse 81 en section 4.4. C'est également une information précieuse en classification multiclasse « un contre tous » (voir la parenthèse 27 en section 2.2).

La quantité $yg(x)$ est souvent appelée « marge » du pseudo-classifieur g , pour une observation (x, y) . Lorsque Φ est décroissante, minimiser le Φ -risque revient donc à maximiser la marge, d'où le nom français « séparateurs à vaste marge » pour les SVM³².

Parenthèse 76 (Convention pour $\text{signe}(0)$) On a choisi ci-dessus de définir $\text{signe}(0) = -1$, par cohérence avec la définition du classifieur plug-in en section 2.3 et notre choix de prendre la classe -1 (ou la classe 0) comme classe « par défaut ». On aurait pu choisir la convention $\text{signe}(0) = 1$, sans changement majeur dans les résultats de la suite de cette section.

Parenthèse 77 (Valeurs infinies des pseudo-classificateurs) On autorise les pseudo-classificateurs à prendre des valeurs infinies pour avoir plus facilement l'existence d'un classifieur de Bayes pour le Φ -risque. Ceci nous amène également à autoriser $\pm\infty$ comme argument de la fonction Φ . On aurait pu se restreindre à des valeurs finies dans les deux cas, au prix de quelques détails techniques supplémentaires.

4.2 Exemples

Les exemples les plus classiques de Φ -risques sont les suivants. Les fonctions Φ correspondantes sont représentées à la figure 5.

— *Risque 0–1*. On pose :

$$\forall u \in \mathbb{R}, \quad \Phi_{0-1}(u) = \mathbf{1}_{u \leq 0}.$$

Le Φ_{0-1} -risque coïncide alors presque avec le risque 0–1 défini en section 2.2. En effet, pour tout $g \in \overline{\mathcal{F}}$, on a :

$$\mathcal{R}_P^{\Phi_{0-1}}(g) \geq \mathcal{R}_P^{0-1}(\text{signe}(g))$$

avec égalité si $\mathbb{P}(g(X) = 0) = 0$.

³² En anglais, les SVM (« support vector machines ») sont également connus sous le nom « maximum-margin classifiers ».

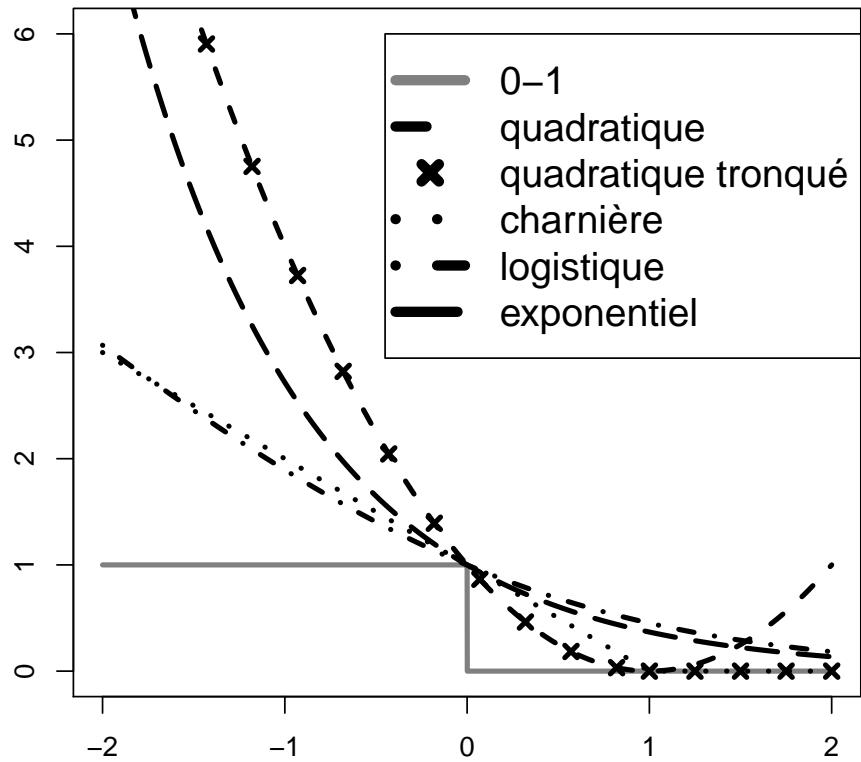


FIGURE 5 – Fonctions Φ définissant les exemples de Φ -risques mentionnés dans le texte.

Démonstration Par définition, on a pour tout $g \in \bar{\mathcal{F}}$:

$$\begin{aligned}
 \mathcal{R}_P^{\Phi_{0-1}}(g) &= \mathbb{E}[\mathbf{1}_{Yg(X) \leq 0}] = \mathbb{P}(Yg(X) \leq 0) \\
 &= \mathbb{P}(Y \neq \text{signe}(g(X)) \text{ ou } g(X) = 0) \\
 &\geq \mathcal{R}_P^{0-1}(\text{signe}(g)),
 \end{aligned}$$

avec égalité si $\mathbb{P}(g(X) = 0) = 0$. □

- *Risque quadratique.* Le Φ -risque est égal au risque quadratique, défini dans le cadre de la régression en section 2.1, lorsque Φ est définie par :

$$\forall u \in \mathbb{R}, \quad \Phi(u) = (1 - u)^2.$$

En effet, on a alors, pour tout $g \in \bar{\mathcal{F}}$:

$$\mathcal{R}_P^\Phi(g) = \mathbb{E}\left[\left(Yg(X) - 1\right)^2\right] = \mathbb{E}\left[\left(g(X) - Y\right)^2\right]$$

puisque $Y \in \{-1, 1\}$.

- *Risque quadratique tronqué*³³. Il correspond à la fonction définie par :

$$\forall u \in \mathbb{R}, \quad \Phi(u) = (1 - u)_+^2 = (\max\{1 - u, 0\})^2.$$

La différence par rapport au risque quadratique est que l'on ne comptabilise pas d'erreur pour g lorsque $g(x)$ est du signe de y et de valeur absolue supérieure à 1.

- *Risque charnière*³⁴. Soit Φ la fonction définie par :

$$\forall u \in \mathbb{R}, \quad \Phi(u) = (1 - u)_+ = \max\{1 - u, 0\}.$$

La fonction de coût associée est à la base des SVM. Un fait remarquable est que f^* , le classifieur de Bayes pour le risque 0–1, est également un pseudo-classifieur de Bayes pour le risque charnière, ce qui n'est pas le cas pour les autres Φ -risques classiques (voir la table 1 en section 4.5).

- *Risque logistique*. Il correspond à la fonction définie par :

$$\forall u \in \mathbb{R}, \quad \Phi(u) = \ln_2(1 + e^{-u}).$$

La fonction de coût associée est à la base de la régression logistique (voir la parenthèse 79).

- *Risque exponentiel*. On l'obtient avec la fonction :

$$\forall u \in \mathbb{R}, \quad \Phi(u) = \exp(-u).$$

La fonction de coût associée est à la base de adaboost, qui est une méthode de type boosting.

Bartlett *et al.* (2006) donnent d'autres exemples de fonctions Φ classiques. D'autres manières de mesurer le risque d'un pseudo-classifieur sont possibles, dont le critère AUC.

Parenthèse 78 (Valeurs de Φ à l'infini) Dans les exemples ci-dessus, on n'a pas précisé les valeurs de Φ en $\pm\infty$. À chaque fois, il faut comprendre que l'on utilise les valeurs limites de Φ .

33. En anglais, la fonction de coût associée au risque quadratique tronqué est appelée « truncated quadratic loss » ou « squared hinge loss » (puisque c'est le carré du coût charnière).

34. En anglais, la fonction de coût associée au risque charnière est appelée « hinge loss » ou « soft loss ».

Parenthèse 79 (Régression logistique) On suppose que $\mathcal{X} = \mathbb{R}^p$. On définit la fonction logistique standard (ou sigmoïde) τ par :

$$\forall u \in \mathbb{R}, \quad \tau(u) = \frac{1}{1 + e^{-u}}.$$

Cette fonction vérifie $\tau' = \tau(1 - \tau)$ et $\tau(-u) = 1 - \tau(u)$. On dit alors que (X, Y) suit le modèle logistique si

$$\ln \frac{\mathbb{P}(Y = 1 | X = \mathbf{x})}{\mathbb{P}(Y = -1 | X = \mathbf{x})} = a + \mathbf{b}^\top \mathbf{x}$$

avec $a \in \mathbb{R}$ et $\mathbf{b} \in \mathbb{R}^p$. De manière équivalente, le modèle logistique est défini par :

$$\mathbb{P}(Y = 1 | X = \mathbf{x}) = \tau(f_\theta(\mathbf{x})) \quad \text{et} \quad \mathbb{P}(Y = -1 | X = \mathbf{x}) = \tau(-f_\theta(\mathbf{x}))$$

en posant

$$\boldsymbol{\theta} = (a, \mathbf{b}) \in \Theta = \mathbb{R}^{p+1} \quad \text{et} \quad f_\theta(\mathbf{x}) = a + \mathbf{b}^\top \mathbf{x}. \quad (34)$$

On note alors P_θ la loi de (X, Y) . Une telle relation est vérifiée par exemple lorsque la loi de X sachant Y est gaussienne et que sa matrice de covariance Σ ne dépend pas de Y , comme supposé dans le cadre de l'analyse discriminante linéaire paramétrique (en anglais, « linear discriminant analysis », abrégé « LDA », voir l'exercice 26). Plus généralement, (X, Y) suit le modèle logistique si $\mathcal{L}(X | Y = 1)$ et $\mathcal{L}(X | Y = -1)$ appartiennent à une même famille exponentielle (Bickel et Doksum, 2001, section 1.6), quitte à remplacer X par la statistique suffisante $T(X)$.

Sous le modèle P_θ , la vraisemblance d'une observation $(\mathbf{x}, y) \in \mathbb{R}^p \times \{-1, 1\}$ s'écrit :

$$\tau(y f_\theta(\mathbf{x})).$$

En particulier, si $\boldsymbol{\theta}$ et \mathbf{x} sont connus (mais pas y), la vraisemblance est maximale pour $y = \text{signe}(f_\theta(\mathbf{x}))$. Et si l'on observe un échantillon $(X_i, Y_i)_{1 \leq i \leq n}$ de variables indépendantes et de loi P_θ , l'estimateur du maximum de vraisemblance de $\boldsymbol{\theta}$ s'écrit :

$$\begin{aligned} \hat{\boldsymbol{\theta}} &\in \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}} \left\{ \prod_{i=1}^n \tau(Y_i f_\theta(X_i)) \right\} \\ &= \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}} \left\{ \frac{1}{n} \sum_{i=1}^n \ln(\tau(Y_i f_\theta(X_i))) \right\} \\ &= \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^n \ln(1 + e^{-Y_i f_\theta(X_i)}) \right\}. \end{aligned} \quad (35)$$

Le classifieur associé est $\text{signe}(f_{\hat{\boldsymbol{\theta}}})$. On parle de *régression logistique*, et d'après (35), cette règle de classification est également obtenue en prenant le signe d'un minimiseur du Φ -risque empirique logistique sur $\{f_\theta / \boldsymbol{\theta} \in \Theta\}$ — qui est un espace vectoriel de dimension finie. Quelques remarques pour conclure :

— Si le modèle logistique est exact, l'erreur d'approximation (pour le risque

- logistique) du modèle $\{f_{\theta} / \theta \in \Theta\}$ est nulle.
- Lorsque Θ est de grande dimension, il est naturel d'ajouter dans (35) un terme de régularisation au Φ -risque empirique.
 - Malgré le lien indiqué ci-dessus, la régression logistique avec le modèle (34) fournit un classifieur différent de l'analyse discriminante linéaire. Le classifieur LDA utilise la formule (64) du classifieur de Bayes donnée à l'exercice 26 en y remplaçant $\mathbb{P}(Y = 1)$, Σ , μ_{-1} et μ_1 par leurs estimateurs empiriques classiques (une forme de « plug-in »), tandis que la régression logistique utilise les valeurs de a et \mathbf{b} qui minimisent le Φ -risque empirique.

4.3 Classification idéale pour le Φ -risque

À la suite de la section 1.2, on peut s'interroger sur l'ensemble des pseudo-classificateurs qui minimisent le Φ -risque sur $\bar{\mathcal{F}}$ (les pseudo-classificateurs de Bayes) et sur la valeur du Φ -risque optimal (le risque de Bayes). On suppose désormais que Φ est convexe³⁵, ce qui est le cas de tous les exemples ci-dessus sauf $\Phi = \Phi_{0-1}$.

Pour tout $\zeta \in [0, 1]$ et $\alpha \in \bar{\mathbb{R}}$, on pose :

$$C_{\zeta}^{\Phi}(\alpha) = \zeta\Phi(\alpha) + (1 - \zeta)\Phi(-\alpha).$$

Pour tout $g \in \bar{\mathcal{F}}$, on a presque sûrement :

$$\mathbb{E}\left[\Phi(Yg(X)) \mid X\right] = C_{\zeta(X)}^{\Phi}(g(X)).$$

Ainsi, en un point $x \in \mathcal{X}$ où l'étiquette Y vaut 1 avec probabilité $\zeta(x) = \zeta$ et où $g(x) = \alpha$, le « Φ -risque conditionnel »³⁶ de g est égal à $C_{\zeta}^{\Phi}(\alpha)$. Pour un pseudo-classificateur optimal, en un tel x , le Φ -risque conditionnel vaut donc :

$$H^{\Phi}(\zeta) := \inf_{\alpha \in \bar{\mathbb{R}}} C_{\zeta}^{\Phi}(\alpha)$$

(sauf éventuellement pour un ensemble de valeurs de x de mesure nulle pour P_X). On en déduit le résultat suivant.

Proposition 13 *On se place en classification binaire, $\mathcal{Y} = \{-1, 1\}$, et l'on considère le coût $c_{\Phi} : (y, y') \mapsto \Phi(yy')$ où Φ est une fonction convexe $\bar{\mathbb{R}} \rightarrow [0, +\infty]$. On a alors les deux résultats suivants :*

(i) *Le risque de Bayes (sur l'ensemble $\bar{\mathcal{F}}$ des pseudo-classificateurs) vaut :*

$$\inf_{g \in \bar{\mathcal{F}}} \mathcal{R}_P^{\Phi}(g) = \mathcal{R}_P^{\Phi \star} = \mathbb{E}\left[H^{\Phi}(\zeta(X))\right].$$

(ii) *Si g_{Φ}^* désigne un pseudo-classificateur, les deux affirmations suivantes sont équivalentes :*

35. La fonction Φ est convexe lorsque, pour tous $u, u' \in \bar{\mathbb{R}}$ et $\delta \in [0, 1]$, on a l'inégalité $\Phi(\delta u + (1 - \delta)u') \leq \delta\Phi(u) + (1 - \delta)\Phi(u')$.

36. Il s'agit ici du Φ -risque calculé conditionnellement à X (et D_n) ; la parenthèse 3 en section 1.1 signale une autre signification possible pour ce terme.

(a) $\mathcal{R}_P^\Phi(g_\Phi^*) = \inf_{g \in \bar{\mathcal{F}}} \mathcal{R}_P^\Phi(g)$, c'est-à-dire, g_Φ^* est un pseudo-classifieur de Bayes pour le Φ -risque.

(b) Avec probabilité 1 :

$$g_\Phi^*(X) \in \operatorname{argmin}_{\alpha \in \bar{\mathbb{R}}} C_{\zeta(X)}^\Phi(\alpha).$$

Démonstration Commençons par remarquer qu'on a, pour tout $g \in \bar{\mathcal{F}}$:

$$C_{\zeta(X)}^\Phi(g(X)) \geq \inf_{\alpha \in \bar{\mathbb{R}}} C_{\zeta(X)}^\Phi(\alpha) = H^\Phi(\zeta(X)). \quad (36)$$

En intégrant, on obtient que pour tout $g \in \bar{\mathcal{F}}$:

$$\mathcal{R}_P^\Phi(g) \geq \mathbb{E}[H^\Phi(\zeta(X))].$$

En particulier, on a :

$$\mathcal{R}_P^\Phi \geq \mathbb{E}[H^\Phi(\zeta(X))].$$

Or, Φ est convexe, donc C_ζ^Φ est convexe pour tout $\zeta \in [0, 1]$. Ainsi, pour tout $\zeta \in [0, 1]$:

$$\operatorname{argmin}_{\alpha \in \bar{\mathbb{R}}} C_\zeta^\Phi(\alpha) \neq \emptyset.$$

Il existe donc $g \in \bar{\mathcal{F}}$ tel que l'inégalité (36) est une égalité presque sûrement. Ceci démontre le point (i).

Pour le point (ii), supposons d'abord que la condition (a) est vérifiée. Alors, au vu de (36) et (i), la variable aléatoire

$$C_{\zeta(X)}^\Phi(g_\Phi^*(X)) - H^\Phi(\zeta(X))$$

est positive ou nulle (presque sûrement) et d'espérance nulle, donc elle est nulle presque sûrement. La condition (b) est donc vérifiée. Réciproquement, si (b) est vérifiée, alors (36) est une égalité presque sûrement, et en intégrant, on obtient avec (i) que :

$$\mathcal{R}_P(g_\Phi^*) = \mathcal{R}_P^\Phi.$$

Ainsi, la condition (a) est vérifiée. \square

La table 1 donne des formules pour g_Φ^* et H^Φ pour les exemples classiques de fonctions Φ .

Remarque 80 (Stricte convexité) Si Φ est strictement convexe³⁷ (ce qui est vérifié par les coûts quadratique, logistique et exponentiel), la fonction C_ζ^Φ est strictement convexe pour tout $\zeta \in [0, 1]$ et donc

$$\operatorname{argmin}_{\alpha \in \bar{\mathbb{R}}} C_{\zeta(X)}^\Phi(\alpha)$$

³⁷. La fonction Φ est strictement convexe lorsqu'elle est convexe et que, pour tous $u \neq u'$ et $\delta \in]0, 1[$, $\Phi(\delta u + (1 - \delta)u') < \delta\Phi(u) + (1 - \delta)\Phi(u')$.

est réduit à un point (éventuellement à l'infini). Alors, il existe un unique³⁸ pseudo-classifieur de Bayes pour le Φ -risque, défini par la condition (b) de la proposition 13.

4.4 Calibration pour la classification 0–1

Le classifieur associé à un minimiseur du Φ -risque est-il toujours un classifieur de Bayes pour le risque 0–1 ? C'est une condition indispensable si le Φ -risque est uniquement une étape intermédiaire, l'objectif final étant de minimiser un coût de classification 0–1. Lorsque c'est bien le cas, on dit que Φ est « calibrée pour la classification 0–1 ». Bartlett *et al.* (2006, définition 1) proposent de le définir formellement comme ceci.

Définition 6 (Fonction calibrée pour la classification 0–1) Une fonction $\Phi : \overline{\mathbb{R}} \rightarrow [0, +\infty]$ est calibrée³⁹ pour la classification 0–1 lorsque :

$$\begin{cases} \forall \zeta > \frac{1}{2}, \quad \inf_{\alpha \leq 0} C_\zeta^\Phi(\alpha) > \inf_{\alpha \in \overline{\mathbb{R}}} C_\zeta^\Phi(\alpha) = H^\Phi(\zeta) \\ \forall \zeta < \frac{1}{2}, \quad \inf_{\alpha > 0} C_\zeta^\Phi(\alpha) > \inf_{\alpha \in \overline{\mathbb{R}}} C_\zeta^\Phi(\alpha) = H^\Phi(\zeta). \end{cases}$$

La proposition suivante justifie le fait que la définition 6 correspond bien à l'intuition annoncée.

Proposition 14 *On suppose que \mathcal{X} est non-vide et Φ est convexe. Alors, les deux affirmations suivantes sont équivalentes.*

- (a) Φ est calibrée pour la classification 0–1.
- (b) Pour toute loi P sur $\mathcal{X} \times \{-1, 1\}$, pour tout $g_\Phi^* \in \overline{\mathcal{F}}$, si g_Φ^* est un pseudo-classifieur de Bayes pour le Φ -risque, alors $\text{signe}(g_\Phi^*)$ est un classifieur de Bayes pour le risque 0–1.

Démonstration Commençons par supposer que (a) est vérifiée. Soit P une loi sur $\mathcal{X} \times \{-1, 1\}$ et g_Φ^* un pseudo-classifieur de Bayes pour le Φ -risque. Alors, d'après la proposition 13, l'événement

$$\Omega = \left\{ g_\Phi^*(X) \in \operatorname{argmin}_{\alpha \in \overline{\mathbb{R}}} C_{\zeta(X)}^\Phi(\alpha) \right\}$$

est de probabilité 1. Sur Ω , si $\zeta(X) > 1/2$, la première partie de la définition de la calibration de Φ pour la classification 0–1 montre que $g_\Phi^*(X) > 0$, et donc $\text{signe}(g_\Phi^*(X)) = 1$. De même, sur Ω , si $\zeta(X) < 1/2$, alors $g_\Phi^*(X) \leq 0$ puisque Φ

38. L'unicité du pseudo-classifieur de Bayes pour le Φ -risque est à comprendre « à modification près des valeurs prises par le pseudo-classifieur sur un ensemble de mesure nulle pour P_X » : si g_1 et g_2 sont deux pseudo-classificateurs de Bayes pour le Φ -risque, alors $g_1(X) = g_2(X)$ presque sûrement.

39. En anglais, une telle fonction Φ est dite « classification-calibrated ». On parle aussi parfois de « Fisher consistency », même si cette notion a souvent une autre signification, proche de la consistance au sens de la définition 1 en section 1.4.

est calibrée, et donc $\text{signe}(g_\Phi^*(X)) = -1$. Au final, sur Ω , on a soit $\zeta(X) = 1/2$, soit

$$\text{signe}(g_\Phi^*)(X) = 2\mathbb{1}_{\zeta(X)>1/2} - 1.$$

Puisque $\mathbb{P}(\Omega) = 1$, la proposition 2 démontre que $\text{signe}(g_\Phi^*)$ est un classifieur de Bayes pour le risque 0–1. L'affirmation (b) est donc vérifiée.

Réciproquement, supposons que (b) est vérifiée. Fixons $\zeta > 1/2$. Puisque \mathcal{X} est non-vide, il existe une loi P sur $\mathcal{X} \times \{-1, 1\}$ telle que $\zeta(X) = \zeta$ presque sûrement. Soit

$$u \in \underset{\alpha \in \bar{\mathbb{R}}}{\operatorname{argmin}} C_\zeta^\Phi(\alpha)$$

quelconque. Alors, le pseudo-classifieur constant égal à u est un pseudo-classifieur de Bayes pour le Φ -risque d'après la proposition 13. Donc, d'après (b), le classifieur constant égal à $\text{signe}(u)$ est un classifieur de Bayes pour le risque 0–1, c'est-à-dire $\text{signe}(u) = 1$ d'après la proposition 2, et donc $u > 0$. Puisque C_ζ^Φ est une fonction convexe (car Φ est convexe), son infimum sur $[-\infty, 0]$ est atteint, de même que son infimum sur $\bar{\mathbb{R}}$. On a donc :

$$\forall \zeta > 1/2, \quad \inf_{\alpha \leq 0} C_\zeta^\Phi(\alpha) > \inf_{\alpha \in \bar{\mathbb{R}}} C_\zeta^\Phi(\alpha).$$

On traite de même le cas $\zeta < 1/2$, ce qui achève de démontrer que (b) implique (a). \square

Parenthèse 81 (Classification calibrée) Une autre notion de calibration, plus ancienne, existe en classification. On dit que g est calibrée (Gneiting *et al.*, 2007; Gneiting et Raftery, 2007) lorsque

$$\forall s \in [0, 1], \quad \mathbb{P}(Y = 1 | g(X) = s) = s,$$

de telle sorte que $g(X)$ fournit une information précieuse lorsque le problème de classification est associé à des décisions lourdes de conséquences. Cette définition n'est pas équivalente à la définition 6 et il ne faut pas les confondre, malgré la proximité terminologique. Dans ce texte, on considère uniquement la calibration au sens de la définition 6.

Toujours en supposant Φ convexe, on a une condition nécessaire et suffisante simple de calibration pour la classification 0–1.

Proposition 15 Soit $\Phi : \bar{\mathbb{R}} \rightarrow [0, +\infty]$ convexe. Alors, Φ est calibrée pour la classification 0–1 si et seulement si Φ est dérivable en 0 et $\Phi'(0) < 0$.

Démonstration Supposons Φ calibrée pour la classification 0–1. Soit $\zeta \in]1/2, 1]$ quelconque. La fonction C_ζ^Φ étant convexe (car Φ est convexe), elle admet une dérivée à droite en zéro, qui vaut :

$$C_{\zeta,d}^\Phi(0) = \zeta \Phi'_d(0) - (1 - \zeta) \Phi'_g(0).$$

Puisque Φ est calibrée, C_ζ^Φ atteint son minimum sur $]0, +\infty]$, et donc sa dérivée à droite en zéro est strictement négative :

$$\zeta \Phi'_d(0) < (1 - \zeta) \Phi'_g(0).$$

En faisant tendre ζ vers $1/2$ (par valeurs positives), on obtient l'inégalité $\Phi'_d(0) \leq \Phi'_g(0)$. Or, Φ étant convexe, on a toujours $\Phi'_d(0) \geq \Phi'_g(0)$. Par conséquent, Φ est dérivable en 0 et pour tout $\zeta > 1/2$ on a :

$$C_{\zeta,d}^{\Phi'}(0) = \zeta \Phi'(0) - (1 - \zeta) \Phi'(0) = (2\zeta - 1) \Phi'(0) < 0,$$

donc $\Phi'(0) < 0$.

Réciproquement, si Φ est dérivable en 0, pour tout $\zeta \in [0, 1]$, C_ζ^Φ est aussi dérivable en 0 et l'on a :

$$C_\zeta^{\Phi'}(0) = (2\zeta - 1) \Phi'(0).$$

Donc, si $\zeta > 1/2$, $C_\zeta^{\Phi'}(0) < 0$ et l'on a :

$$\inf_{\alpha \leq 0} C_\zeta^\Phi(\alpha) > \inf_{\alpha \in \overline{\mathbb{R}}} C_\zeta^\Phi(\alpha).$$

Si $\zeta < 1/2$, $C_\zeta^{\Phi'}(0) > 0$ et l'on a :

$$\inf_{\alpha > 0} C_\zeta^\Phi(\alpha) > \inf_{\alpha \in \overline{\mathbb{R}}} C_\zeta^\Phi(\alpha).$$

□

Pour les risques quadratique, quadratique tronqué, charnière, logistique et exponentiel, la fonction Φ est convexe et calibrée pour la classification 0–1.

Parenthèse 82 (Classification multiclassse) La notion de calibration pour la classification 0–1 peut s'étendre au cas multiclassse, et l'on dispose de généralisations de la proposition 15 à ce cadre (Tewari et Bartlett, 2007) ; voir aussi la parenthèse 42 en section 2.3.

4.5 Lien entre excès de Φ -risque et excès de risque 0–1

L'analyse théorique de règles minimisant le Φ -risque empirique (régiaralisé ou pas) mène naturellement à des garanties sur leur Φ -risque. On peut s'arrêter là si l'on souhaite effectivement disposer d'un pseudo-classifieur et que le Φ -risque mesure bien sa qualité (la remarque 75 en section 4.1 mentionne plusieurs raisons pour cela). Mais si l'on ne s'intéresse qu'à la qualité de prévision de l'étiquette Y , c'est sur un risque de classification — souvent le risque 0–1 — que l'on souhaite avoir des garanties. Il est alors utile de pouvoir relier Φ -risque et risque 0–1, ce que fait le théorème suivant.

Théorème 1 *On suppose Φ convexe et calibrée pour la classification 0–1. On pose, pour tout $\theta \in [-1, 1]$:*

$$\Psi(\theta) := \Phi(0) - H^\Phi\left(\frac{1+\theta}{2}\right).$$

Alors, Ψ est positive, paire et convexe. De plus, $\Psi(\theta) = 0$ si et seulement si $\theta = 0$. Enfin, pour tout pseudo-classifieur $g \in \bar{\mathcal{F}}$ et toute loi P sur $\mathcal{X} \times \{-1, 1\}$, on a :

$$\Psi(\mathcal{R}_P^{0-1}(\text{signe}(g)) - \mathcal{R}_P^{0-1*}) \leq \mathcal{R}_P^\Phi(g) - \mathcal{R}_P^{\Phi*}. \quad (37)$$

Démonstration Pour tout $\theta \in [-1, 1]$, on a :

$$\Psi(\theta) = \Phi(0) - H^\Phi\left(\frac{1+\theta}{2}\right) = C_{\frac{1+\theta}{2}}^\Phi(0) - H^\Phi\left(\frac{1+\theta}{2}\right) \geq 0, \quad (38)$$

ce qui montre la première affirmation.

Pour tout $\theta \in [-1, 1]$ et tout $\alpha \in \bar{\mathbb{R}}$, on a :

$$C_{\frac{1-\theta}{2}}^\Phi(\alpha) = \frac{1-\theta}{2}\Phi(\alpha) + \frac{1+\theta}{2}\Phi(-\alpha) = C_{\frac{1+\theta}{2}}^\Phi(-\alpha)$$

donc $H^\Phi((1-\theta)/2) = H^\Phi((1+\theta)/2)$, et Ψ est paire.

La fonction H^Φ est définie comme l'infimum d'une collection de fonctions linéaires (donc concaves), donc H^Φ est concave, et Ψ est convexe.

Puisque Φ est calibrée pour la classification 0–1, pour tout $\theta \in]0, 1]$, on a :

$$C_{\frac{1+\theta}{2}}^\Phi(0) \geq \inf_{\alpha \leq 0} C_{\frac{1+\theta}{2}}^\Phi(\alpha) > H^\Phi\left(\frac{1+\theta}{2}\right).$$

Avec (38), on obtient que $\Psi(\theta) > 0$. De même, pour tout $\theta \in [-1, 0[$, on a :

$$C_{\frac{1+\theta}{2}}^\Phi(0) \geq \inf_{\alpha \geq 0} C_{\frac{1+\theta}{2}}^\Phi(\alpha) = \inf_{\alpha > 0} C_{\frac{1+\theta}{2}}^\Phi(\alpha) > H^\Phi\left(\frac{1+\theta}{2}\right),$$

où l'égalité centrale provient de la continuité de $C_{\frac{1+\theta}{2}}^\Phi$ en 0 (puisque Φ est convexe à valeurs finies au voisinage de 0, elle est continue en 0). On a donc $\Psi(\theta) > 0$ pour tout $\theta \neq 0$. Lorsque $\theta = 0$, $H^\Phi(1/2) = \Phi(0)$ puisque Φ est convexe, et donc $\Psi(0) = 0$.

Enfin, pour tout $g \in \bar{\mathcal{F}}$ et toute loi P sur $\mathcal{X} \times \{-1, 1\}$, en posant $f^*(x) =$

$\mathbb{1}_{\zeta(x) > 1/2}$, on a :

$$\begin{aligned}
 & \Psi(\mathcal{R}_P^{0-1}(\text{signe}(g)) - \mathcal{R}_P^{0-1*}) \\
 &= \Psi\left(\mathbb{E}\left[\mathbb{1}_{f^*(X) \neq \mathbb{1}_{g(X)} > 0} |2\zeta(X) - 1|\right]\right) && (\text{proposition 2}) \\
 &\leq \mathbb{E}\left[\Psi\left(\mathbb{1}_{f^*(X) \neq \mathbb{1}_{g(X)} > 0} |2\zeta(X) - 1|\right)\right] && (\Psi \text{ convexe}) \\
 &= \mathbb{E}\left[\mathbb{1}_{f^*(X) \neq \mathbb{1}_{g(X)} > 0} \Psi(|2\zeta(X) - 1|)\right] && (\Psi(0) = 0) \\
 &= \mathbb{E}\left[\mathbb{1}_{f^*(X) \neq \mathbb{1}_{g(X)} > 0} \Psi(2\zeta(X) - 1)\right] && (\Psi \text{ paire}) \\
 &= \mathbb{E}\left[\mathbb{1}_{f^*(X) \neq \mathbb{1}_{g(X)} > 0} (\Phi(0) - H^\Phi(\zeta(X)))\right] && (\text{définition de } \Psi).
 \end{aligned}$$

Or, conditionnellement à X , si $f^*(X) \neq \mathbb{1}_{g(X) > 0}$, c'est-à-dire si $g(X)$ « se trompe de signe », alors son Φ -risque conditionnel $C_{\zeta(X)}^\Phi(g(X))$ est supérieur ou égal à

$$\inf_{\alpha \in \mathbb{R} / \alpha(2\zeta(X) - 1) < 0} C_{\zeta(X)}^\Phi(\alpha) = \Phi(0).$$

Autrement dit, on a :

$$\mathbb{1}_{f^*(X) \neq \mathbb{1}_{g(X)} > 0} \Phi(0) \leq \mathbb{1}_{f^*(X) \neq \mathbb{1}_{g(X)} > 0} C_{\zeta(X)}^\Phi(g(X)).$$

On obtient donc :

$$\begin{aligned}
 & \Psi(\mathcal{R}_P^{0-1}(\text{signe}(g)) - \mathcal{R}_P^{0-1*}) \\
 &\leq \mathbb{E}\left[\underbrace{\mathbb{1}_{f^*(X) \neq \mathbb{1}_{g(X)} > 0}}_{\leq 1} \underbrace{(C_{\zeta(X)}^\Phi(g(X)) - H^\Phi(\zeta(X)))}_{\geq 0}\right] \\
 &\leq \mathbb{E}\left[C_{\zeta(X)}^\Phi(g(X)) - H^\Phi(\zeta(X))\right] \\
 &= \mathcal{R}_P^\Phi(g) - \mathcal{R}_P^{\Phi*}.
 \end{aligned}$$

□

La fonction Ψ étant continue et ne s'annulant qu'en 0, le théorème 1 démontre notamment que si \hat{g} est une (pseudo-)règle de classification consistante pour le Φ -risque, alors $\text{signe}(\hat{g})$ est une règle de classification consistante pour le risque 0–1. Et la fonction Ψ permet de « transférer » une borne sur la vitesse d'apprentissage pour le Φ -risque en une borne sur la vitesse d'apprentissage pour le risque 0–1.

La table 1 donne des formules pour Ψ pour les exemples de fonctions Φ classiques. En particulier, pour le coût quadratique, $\Psi(\theta) = \theta^2$ et le théorème 1 donne un résultat strictement équivalent à la proposition 4 en section 2.3 (hormis le majorant intermédiaire).

Parenthèse 83 (À propos de la table 1) Les formules rassemblées dans

TABLE 1 – Fonctions H^Φ , g_Φ^* et Ψ pour les cinq exemples classiques de fonctions Φ .

$\Phi(u)$	$H^\Phi(\zeta)$	$g_\Phi^*(x)$	$\Psi(\theta)$
$(1-u)^2$	$1 - (2\zeta - 1)^2$	$\eta(x) = 2\zeta(x) - 1$	θ^2
$(1-u)_+^2$	$4\zeta(1-\zeta)$	$\eta(x) = 2\zeta(x) - 1$	θ^2
$(1-u)_+$	$2 \min\{\zeta, 1-\zeta\}$	$\text{signe}(2\zeta(x) - 1)$	$ \theta $
$\ln_2(1 + e^{-u})$	$-\zeta \ln_2 \zeta - (1-\zeta) \ln_2(1-\zeta)$	$\ln \frac{\zeta(x)}{1-\zeta(x)}$	$\frac{(1+\theta) \ln(1+\theta) + (1-\theta) \ln(1-\theta)}{2 \ln 2} \geq \frac{\theta^2}{2 \ln 2}$
e^{-u}	$2\sqrt{\zeta(1-\zeta)}$	$\frac{1}{2} \ln \frac{\zeta(x)}{1-\zeta(x)}$	$1 - \sqrt{1 - \theta^2} \geq \frac{\theta^2}{2}$

la table 1 cachent quelques petites ambiguïtés. Pour le risque quadratique tronqué ($\Phi : u \mapsto (1-u)_+^2$) et le risque charnière ($\Phi : u \mapsto (1-u)_+$) : lorsque $\zeta(x) = 0$, n’importe quelle valeur $g_\Phi^*(x) \in [-\infty, -1]$ convient, et lorsque $\zeta(x) = 1$, n’importe quelle valeur $g_\Phi^*(x) \in [1, +\infty]$ convient. Pour le risque logistique ($\Phi : u \mapsto \ln_2(1 + e^{-u})$) et le risque exponentiel ($\Phi : u \mapsto e^{-u}$) : lorsque $\zeta(x) = 0$, $g_\Phi^*(x) = -\infty$, et lorsque $\zeta(x) = 1$, $g_\Phi^*(x) = +\infty$.

Bartlett *et al.* (2006, théorème 1) montrent que la majoration (37) est inaméliorable lorsque $\text{Card}(\mathcal{X}) \geq 2$: pour tout $\theta \in [0, 1]$, on peut trouver une loi P et un pseudo-classifieur g d’excès de risque 0–1 égal à θ et qui réalise (presque) l’égalité dans (37). Cependant, le théorème 1 peut être amélioré sous certaines hypothèses sur la loi P . Ainsi, si P est une loi « zéro-erreur » (voir la remarque 34 en section 2.2), on a :

$$\Psi(1) \left[\mathcal{R}_P^{0-1}(\text{signe}(g)) - \mathcal{R}_P^{0-1*} \right] \leq \mathcal{R}_P^\Phi(g) - \mathcal{R}_P^{\Phi*}.$$

Bartlett *et al.* (2006, théorème 3) démontrent en fait un résultat encore plus général. L’exercice 29 donne son énoncé dans le cas particulier de la « condition de marge », évoquée en section 7.

Comme on l’a signalé en introduction de cette section, de nombreux classificateurs sont définis comme le signe d’un minimiseur du Φ -risque empirique (régularisé ou pas). Les SVM (avec le coût charnière) et la régression logistique (avec le coût logistique) en sont deux exemples emblématiques. Au vu de la table 1, il est tentant (mais périlleux) de comparer les mérites des différentes fonctions Φ envisageables en comparant uniquement les fonctions Ψ . On pourrait ainsi croire que le coût charnière, grâce à sa fonction de transfert $\Psi(\theta) = |\theta|$, permet d’obtenir de meilleures vitesses d’apprentissage que les autres coûts, dont les fonctions de transfert sont (approximativement) quadratiques. Faut-il donc préférer les SVM à la régression logistique ?

La réponse n’est pas toujours oui, et ceci pour plusieurs raisons. D’une part, on compare ici des *bornes supérieures* sur l’excès de risque, ce qui peut être trompeur : la borne peut être plus large en pratique pour certaines fonctions Φ que pour d’autres.

D'autre part, si l'on applique la borne à des fonctions g minimisant le Φ -risque empirique sur des modèles de complexité « raisonnable », il faut tenir compte du fait que pour un même problème de prévision, les vitesses d'apprentissage atteignables (pour l'excès de Φ -risque) ne sont pas forcément les mêmes pour deux fonctions Φ différentes ! En particulier, l'erreur d'approximation de g_Φ^* peut dépendre fortement de Φ . Avec le coût charnière, $g_\Phi^* = \text{signe}(2\zeta - 1)$ est une fonction irrégulière (elle « saute » brusquement lorsque $\zeta(x)$ franchit la valeur $1/2$), donc difficile à approcher avec la plupart des modèles « simples ». En revanche, avec le coût logistique (par exemple), si (X, Y) suit un modèle logistique (défini dans la parenthèse 79 en section 4.2), g_Φ^* est une fonction linéaire de X . Lorsque le modèle logistique est (à peu près) correct, on peut donc espérer obtenir de meilleurs résultats avec le coût logistique qu'avec le coût charnière, malgré une « moins bonne » fonction Ψ .

5 Moyenne locale

Une deuxième grande famille de règles d'apprentissage (régression ou classification) repose sur le principe de moyenne locale.

Ces règles peuvent être analysées simultanément, notamment via le théorème de Stone — démontré en section 5.2 — qui est historiquement le premier résultat démontrant qu'une règle d'apprentissage est universellement consistante (dans le cas des plus proches voisins, voir le corollaire 2 en section 5.4).

5.1 Définition

Une règle de régression par moyenne locale est une règle qui fait une prévision en $x \in \mathcal{X}$ à l'aide d'une moyenne (pondérée) des Y_i qui correspondent à un X_i « voisin » de x . Une règle de classification s'en déduit par plug-in.

Définition 7 (Règle d'apprentissage par moyenne locale) Soit, pour tout entier $n \geq 1$ et tout $x, x_1, \dots, x_n \in \mathcal{X}$, des réels positifs :

$$W_1(x_{1\dots n}; x), \dots, W_i(x_{1\dots n}; x), \dots, W_n(x_{1\dots n}; x).$$

On définit alors la règle de régression par moyenne locale⁴⁰ associée : pour tout $x \in \mathcal{X}$ et $(x_i, y_i)_{1 \leq i \leq n} \in (\mathcal{X} \times \mathbb{R})^n$,

$$\hat{\eta}((x_i, y_i)_{1 \leq i \leq n}; x) := \sum_{i=1}^n W_i(x_{1\dots n}; x) y_i. \quad (39)$$

La règle de classification par moyenne locale associée à $(W_i)_{1 \leq i \leq n}$ s'en déduit par plug-in :

$$\hat{f}((x_i, y_i)_{1 \leq i \leq n}; x) := \mathbf{1}_{\hat{\eta}((x_i, y_i)_{1 \leq i \leq n}; x) > 1/2}. \quad (40)$$

40. On parle aussi de règle par moyennage local ; les termes « local average estimator » et « local averaging estimate » sont utilisés en anglais.

Bien que cela ne soit pas dans la définition, les poids W_i sont en général choisis de somme 1 :

$$\forall x, x_1, \dots, x_n \in \mathcal{X}, \quad \sum_{i=1}^n W_i(x_{1\dots n}; x) = 1 \quad (41)$$

ou bien « approximativement » de somme 1 (voir la condition (a) du théorème 2 en section 5.2). Dès lors, la définition (39) indique bien que $\hat{\eta}(D_n; x)$ est une moyenne des Y_i .

Le caractère local de la moyenne n'est pas présent dans la définition ci-dessus, mais il apparaît clairement dans les exemples qui suivent, ainsi que dans la condition (c) du théorème 2. L'intuition derrière la définition 7 est que faire la moyenne d'un nombre suffisant de Y_i pour calculer $\hat{\eta}$ permet d'éviter le surapprentissage, tandis que le caractère local de la moyenne (39) permet de rendre compte des variations de η (et ainsi éviter le sous-apprentissage).

Parenthèse 84 (Sur les poids W_i) Implicitement, dans la définition 7, les poids W_i dépendent aussi de n (ne serait-ce que parce que ce sont des fonctions de $n + 1$ variables). On note donc ces poids $(W_{i,n})_{1 \leq i \leq n}$ lorsque le contexte le nécessite. Pour que $\hat{\eta}$ soit bien une application mesurable, en toute rigueur, il faut supposer que chaque W_i est une fonction mesurable de ses entrées $(x_{1\dots n}; x)$. On suppose toujours cette hypothèse vérifiée dans la suite.

Le premier exemple de règle par moyenne locale est celui des règles par partition, évoqué à plusieurs reprises par les sections précédentes.

Exemple 7 (Règle par partition) Soit \mathcal{A} une partition mesurable de \mathcal{X} , finie ou dénombrable. Pour tout $x \in \mathcal{X}$, on note $\mathcal{A}(x)$ l'élément de la partition qui contient x et pour tout $x_1, \dots, x_n \in \mathcal{X}$:

$$W_i^{\mathcal{A}}(x_{1\dots n}; x) := \begin{cases} \frac{\mathbb{1}_{x_i \in \mathcal{A}(x)}}{N_{\mathcal{A}(x)}(x_{1\dots n})} & \text{si } N_{\mathcal{A}(x)}(x_{1\dots n}) > 0 \\ 0 & \text{sinon.} \end{cases}$$

$$\text{où } \forall A \in \mathcal{A}, \quad N_A(x_{1\dots n}) := \text{Card}\{j \in \{1, \dots, n\} / x_j \in A\}.$$

Alors, la règle par moyenne locale associée à $(W_i^{\mathcal{A}})_{1 \leq i \leq n}$ coïncide avec la règle par partition définie aux exemples 1 ($\widehat{f}_{\mathcal{A}}^{p-r}$, en régression) et 3 ($\widehat{f}_{\mathcal{A}}^{p-c}$, en classification) en section 2. De plus, pour tout $x, x_1, \dots, x_n \in \mathcal{X}$, on a :

$$\sum_{i=1}^n W_i^{\mathcal{A}}(x_{1\dots n}; x) = \mathbb{1}_{N_{\mathcal{A}(x)}(x_{1\dots n}) > 0} \leq 1. \quad (42)$$

L'équation (42) montre bien qu'il s'agit d'une moyenne (sauf sur les éléments de \mathcal{A} pour lesquels on ne dispose d'aucune observation). Le caractère local est défini par la partition \mathcal{A} : x et x_i sont « voisins » lorsqu'ils appartiennent au même élément de \mathcal{A} .

La section 3 propose une manière d'analyser les règles par partition, en utilisant le fait qu'elles minimisent un risque empirique sur un modèle. Le fait de les voir comme

des règles par moyenne locale fournit une autre approche (complémentaire) pour les comprendre, que détaille la section 5.3.

D'autres exemples de règles par moyenne locale sont indiqués en section 5.4 (k -plus proches voisins) et 5.5 (noyau).

5.2 Consistance : théorème de Stone

Le théorème ci-dessous, initialement dû à Stone, fournit des conditions suffisantes simples pour la consistance faible d'une règle par moyenne locale en classification 0–1 lorsque $\mathcal{X} = \mathbb{R}^p$.

Théorème 2 (Théorème de Stone) *On suppose que $\mathcal{X} = \mathbb{R}^p$ est muni d'une norme $\|\cdot\|$. Pour tout $n \geq i \geq 1$ entiers, soit $W_{i,n} : \mathcal{X}^{n+1} \rightarrow \mathbb{R}^+$ une application mesurable. Soit P une loi sur $\mathcal{X} \times \{0,1\}$, P_X sa première marginale et X, X_1, X_2, \dots une suite (infinie) de variables aléatoires indépendantes de loi commune P_X . On suppose que P_X et les $W_{i,n}$ vérifient les quatre conditions suivantes.*

(a) *Il existe une constante $c_a > 0$ telle que :*

$$\sum_{i=1}^n W_{i,n}(X_{1\dots n}; X) \xrightarrow[n \rightarrow +\infty]{L^1} 1 \quad \text{et} \quad \sum_{i=1}^n W_{i,n}(X_{1\dots n}; X) \leq c_a$$

presque sûrement.

(b) *Pour tout $a > 0$:*

$$\lim_{n \rightarrow +\infty} \mathbb{E} \left[\sum_{i=1}^n W_{i,n}(X_{1\dots n}; X) \mathbf{1}_{\|X_i - X\| \geq a} \right] = 0.$$

(c)

$$\lim_{n \rightarrow +\infty} \mathbb{E} \left[\max_{1 \leq i \leq n} W_{i,n}(X_{1\dots n}; X) \right] = 0.$$

(d) *Il existe une constante $c_d > 0$ telle que pour toute fonction $f \in L^1(P_X)$:*

$$\mathbb{E} \left[\sum_{i=1}^n W_{i,n}(X_{1\dots n}; X) |f(X_i)| \right] \leq c_d \mathbb{E}[|f(X)|].$$

Alors, en classification binaire avec le risque 0–1, la règle par moyenne locale \hat{f} associée aux poids $(W_{i,n})_{1 \leq i \leq n}$ définie par (40) est (faiblement) consistante pour P .

Avant de démontrer le théorème 2, commentons les quatre conditions suffisantes de consistance faible qu'il fournit.

- La condition (a) indique qu'il s'agit d'une moyenne (ou presque), sa deuxième partie étant technique et vérifiée avec $c_a = 1$ dans tous les exemples qui suivent.
- La condition (b) correspond au côté « local » des poids (la notion de voisinage étant définie par la norme sur \mathbb{R}^p). Toutes les normes sur \mathbb{R}^p étant équivalentes, peu importe le choix de la norme sur \mathbb{R}^p .

- La condition (c) correspond au fait qu'aucune observation particulière n'a un poids strictement positif asymptotiquement : il s'agit *vraiment* d'une moyenne sur un grand nombre d'observations.
- La condition (d) est technique et difficile à interpréter. Elle impose que la mesure empirique pondérée

$$P_n^W := \frac{1}{n} \sum_{i=1}^n W_{i,n}(X_{1\dots n}; X) \delta_{X_i}$$

est « comparable » en moyenne à la loi de X :

$$\forall f \in L^1(P_X), \quad \int |f| dP_n^W \leq c_d \int |f| dP_X.$$

La condition (d) n'est pas nécessaire si η est uniformément continue sur le support de P_X . Dans la démonstration, cette condition est utilisée pour $f = |\eta - \eta'|$ avec η' régulière « proche » de η .

Démonstration du théorème 2

Cette démonstration s'inspire de celle de Devroye *et al.* (1996, théorème 6.3), qui suppose $\sum_i W_{i,n} = 1$ presque sûrement. Le lecteur intéressé peut aussi consulter la démonstration de Biau et Devroye (2015, théorème 17.2), qui s'appuie sur un résultat plus général en régression L^p .

Étape 1 : Décomposition en deux termes La règle \hat{f} est de type « plug-in ». D'après la proposition 4 en section 2.3, il suffit de montrer que :

$$\mathbb{E}|\eta(X) - \hat{\eta}(D_n; X)| \xrightarrow[n \rightarrow +\infty]{} 0.$$

Posons, pour tout $x \in \mathcal{X}$:

$$\eta^*(X_{1\dots n}; x) := \sum_{i=1}^n \eta(X_i) W_{i,n}(X_{1\dots n}; x).$$

D'après l'inégalité triangulaire :

$$\begin{aligned} \mathbb{E}|\eta(X) - \hat{\eta}(D_n; X)| &\leq \mathbb{E}|\eta(X) - \eta^*(X_{1\dots n}; X)| \\ &\quad + \mathbb{E}|\eta^*(X_{1\dots n}; X) - \hat{\eta}(D_n; X)|. \end{aligned} \tag{43}$$

Il suffit donc de montrer que chacun de ces deux termes tend vers zéro quand n tend vers l'infini.

Parenthèse 85 (Interprétation de (43)) Pour tout $x \in \mathcal{X}$, on a :

$$\eta^*(X_{1\dots n}; x) = \mathbb{E}[\hat{\eta}(D_n; x) | X_{1\dots n}] = \hat{\eta}\left((X_i, \eta(X_i))_{1 \leq i \leq n}; x\right).$$

Le premier terme de la borne supérieure dans (43) ressemble donc à une erreur d'approximation, le deuxième à une erreur d'estimation. On peut obtenir une décomposition exacte du risque quadratique en deux termes similaires, voir l'équation (44) et la parenthèse 86 à la fin de cette sous-section.

Étape 2 : Contrôle du premier terme, si η régulière Supposons que η est uniformément continue. Alors, pour tout $\epsilon > 0$, il existe $a > 0$ tel que :

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

Or,

$$\begin{aligned} & |\eta(X) - \eta^*(X_{1\dots n}; X)| \\ &= \left| \sum_{i=1}^n W_{i,n}(X_{1\dots n}; X)(\eta(X) - \eta(X_i)) + \eta(X) \left(1 - \sum_{i=1}^n W_{i,n}(X_{1\dots n}; X) \right) \right| \\ &\leq \sum_{i=1}^n W_{i,n}(X_{1\dots n}; X) |\eta(X) - \eta(X_i)| + \left| 1 - \sum_{i=1}^n W_{i,n}(X_{1\dots n}; X) \right| \end{aligned}$$

car les $W_{i,n}$ sont positifs et $|\eta(X)| \leq 1$. En intégrant ceci, on obtient :

$$\begin{aligned} & \mathbb{E} |\eta(X) - \eta^*(X_{1\dots n}; X)| \\ &\leq \mathbb{E} \left[\sum_{i=1}^n W_{i,n}(X_{1\dots n}; X) |\eta(X) - \eta(X_i)| \right] + \mathbb{E} \left| \sum_{i=1}^n W_{i,n}(X_{1\dots n}; X) - 1 \right| \\ &= \mathbb{E} \left[\sum_{i=1}^n W_{i,n}(X_{1\dots n}; X) \underbrace{\mathbb{1}_{\|X-X_i\| < a} |\eta(X) - \eta(X_i)|}_{\leq \epsilon} \right] \\ &\quad + \mathbb{E} \left[\sum_{i=1}^n W_{i,n}(X_{1\dots n}; X) \mathbb{1}_{\|X-X_i\| \geq a} \underbrace{|\eta(X) - \eta(X_i)|}_{\leq 1} \right] \\ &\quad + \mathbb{E} \left| \sum_{i=1}^n W_{i,n}(X_{1\dots n}; X) - 1 \right| \\ &\leq \epsilon + \mathbb{E} \left[\sum_{i=1}^n W_{i,n}(X_{1\dots n}; X) \mathbb{1}_{\|X-X_i\| \geq a} \right] + (1+\epsilon) \mathbb{E} \left| \sum_{i=1}^n W_{i,n}(X_{1\dots n}; X) - 1 \right|. \end{aligned}$$

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

$$\limsup_{n \rightarrow +\infty} \mathbb{E} |\eta(X) - \eta^*(X_{1\dots n}; X)| \leq \epsilon,$$

et donc que $\mathbb{E} |\eta(X) - \eta^*(X_{1\dots n}; X)|$ converge vers zéro quand n tend vers l'infini.

Étape 3 : Contrôle du premier terme, cas général Lorsque η n'est pas uniformément continue, on utilise le fait que les fonctions continues à support compact sont denses dans $L^1(P_X)$. Par conséquent, pour tout $\epsilon > 0$, il existe $\tilde{\eta} \in L^1(P_X)$ telle que $\tilde{\eta}$ est continue à support compact et :

$$\mathbb{E}|\eta(X) - \tilde{\eta}(X)| = \|\eta - \tilde{\eta}\|_{L^1(P_X)} \leq \epsilon.$$

Comme $\tilde{\eta}$ est uniformément continue et bornée (à l'étape 2, on a utilisé que $|\eta| \leq 1$, mais n'importe quelle autre borne conviendrait), le raisonnement de l'étape 2 s'applique en remplaçant η par $\tilde{\eta}$ d'un côté, et en remplaçant η^* par :

$$\tilde{\eta}^*(X_{1\dots n}; x) := \sum_{i=1}^n \tilde{\eta}(X_i) W_{i,n}(X_{1\dots n}; x)$$

d'un autre côté. On obtient donc :

$$\mathbb{E}|\tilde{\eta}(X) - \tilde{\eta}^*(X_{1\dots n}; X)| \xrightarrow[n \rightarrow +\infty]{} 0.$$

Ainsi, par l'inégalité triangulaire, on a :

$$\begin{aligned} & \mathbb{E}|\eta(X) - \eta^*(X_{1\dots n}; X)| \\ & \leq \underbrace{\mathbb{E}|\eta(X) - \tilde{\eta}(X)|}_{\leq \epsilon} + \underbrace{\mathbb{E}|\tilde{\eta}(X) - \tilde{\eta}^*(X_{1\dots n}; X)|}_{\xrightarrow[n \rightarrow +\infty]{} 0} + \mathbb{E}|\tilde{\eta}^*(X_{1\dots n}; X) - \eta^*(X_{1\dots n}; X)|. \end{aligned}$$

Il reste uniquement à majorer le dernier terme. Or,

$$\begin{aligned} \mathbb{E}|\tilde{\eta}^*(X_{1\dots n}; X) - \eta^*(X_{1\dots n}; X)| &= \mathbb{E}\left|\sum_{i=1}^n W_{i,n}(X_{1\dots n}; X)(\tilde{\eta}(X_i) - \eta(X_i))\right| \\ &\leq \mathbb{E}\left[\sum_{i=1}^n W_{i,n}(X_{1\dots n}; X)|\tilde{\eta}(X_i) - \eta(X_i)|\right] \\ &\leq c_d \mathbb{E}|\tilde{\eta}(X) - \eta(X)| \leq c_d \epsilon, \end{aligned}$$

en utilisant successivement les définitions de $\tilde{\eta}^*$ et η^* , le fait que les $W_{i,n}$ sont positifs, la condition (d) avec $f = |\tilde{\eta} - \eta|$, et pour finir la définition de $\tilde{\eta}$.

Pour conclure sur le premier terme du membre de droite de (43), nous avons prouvé que pour tout $\epsilon > 0$:

$$\limsup_{n \rightarrow +\infty} \mathbb{E}|\eta(X) - \eta^*(X_{1\dots n}; X)| \leq \epsilon + c_d \epsilon.$$

Comme ϵ peut être pris arbitrairement proche de zéro, nous avons prouvé que la quantité $\mathbb{E}|\eta(X) - \eta^*(X_{1\dots n}; X)|$ tend vers 0 quand n tend vers l'infini.

Étape 4 : Contrôle du deuxième terme Pour le deuxième terme du membre de droite de (43), on utilise d'abord l'inégalité de Jensen :

$$\mathbb{E}|\eta^*(X_{1\dots n}; X) - \hat{\eta}(D_n; X)| \leq \sqrt{\mathbb{E}[(\eta^*(X_{1\dots n}; X) - \hat{\eta}(D_n; X))^2]}.$$

De plus, pour tout $i \neq j$ on a :

$$\mathbb{E}\left[\left(Y_i - \eta(X_i)\right)\left(Y_j - \eta(X_j)\right) \mid X, X_{1\dots n}\right] = 0$$

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

$$\begin{aligned} & \mathbb{E}\left[(\eta^*(X_{1\dots n}; X) - \hat{\eta}(D_n; X))^2\right] \\ &= \mathbb{E}\left[\left(\sum_{i=1}^n (\eta(X_i) - Y_i) W_{i,n}(X_{1\dots n}; X)\right)^2\right] \\ &= \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}\left[(\eta(X_i) - Y_i) W_{i,n}(X_{1\dots n}; X) (\eta(X_j) - Y_j) W_{j,n}(X_{1\dots n}; X)\right] \\ &= \sum_{i=1}^n \mathbb{E}\left[\left((\eta(X_i) - Y_i) W_{i,n}(X_{1\dots n}; X)\right)^2\right] \\ &\leq \frac{1}{4} \mathbb{E}\left[\sum_{i=1}^n W_{i,n}(X_{1\dots n}; X)^2\right] \leq \frac{c_a}{4} \mathbb{E}\left[\max_{1 \leq i \leq n} W_{i,n}(X_{1\dots n}; X)\right] \end{aligned}$$

d'après (a), et ce majorant tend vers zéro quand n tend vers l'infini d'après (c). \square

Une lecture attentive de la démonstration du théorème 2 indique que les hypothèses (a), (b) et (d) servent à démontrer qu'il n'y a pas de sous-apprentissage (l' \ll erreur d'approximation \gg tend vers 0), tandis que les hypothèses (a) et (c) garantissent qu'il n'y a pas de surapprentissage (l' \ll erreur d'estimation \gg tend vers 0).

Cas de la régression

La démonstration du théorème 2 démontre la consistante de $\hat{f}_{\hat{\eta}}$ en classification via la consistante de $\hat{\eta}$ en régression pour le risque L^1 d'estimation de η . Une démonstration similaire permet d'obtenir un résultat en régression (lorsque $\mathcal{X} = \mathbb{R}^p$) pour le risque L^q

$$\mathbb{E}[|\hat{f}_{\hat{\eta}} - \eta|^q]$$

avec $q \geq 1$ quelconque, en supposant que $\mathbb{E}|Y|^q < +\infty$ (Biau et Devroye, 2015, chapitre 10).

Réiproquement, si \hat{f} est universellement consistante en régression L^p et si $\sum_i W_{i,n} \leq c_a$, alors les conditions (a), (b), (c), (d) doivent être vérifiées pour toute loi P_X sur \mathcal{X} (Biau et Devroye, 2015, section 10.2).

On peut facilement obtenir une décomposition précise du risque en régression avec le coût quadratique. On a en effet :

$$\begin{aligned} & \mathbb{E}\left[\left(\eta(X) - \widehat{\eta}(D_n; X)\right)^2\right] \\ &= \underbrace{\mathbb{E}\left[\left(\eta(X) - \eta^*(X_{1\dots n}; X)\right)^2\right]}_{\text{biais}} + \underbrace{\mathbb{E}\left[\left(\eta^*(X_{1\dots n}; X) - \widehat{\eta}(D_n; X)\right)^2\right]}_{\text{variance}} \end{aligned} \quad (44)$$

$$= \mathbb{E}\left[\left(\eta(X) - \eta^*(X_{1\dots n}; X)\right)^2\right] + \sum_{i=1}^n \mathbb{E}\left[\sigma^2(X_i) W_{i,n}^2(X_{1\dots n}; X)\right] \quad (45)$$

en notant

$$\sigma^2(X) := \mathbb{E}\left[(Y - \eta(X))^2 \mid X\right]$$

la variance résiduelle. On peut en déduire un résultat de consistance général sous la seule hypothèse que $\mathbb{E}[Y^2] < +\infty$.

Parenthèse 86 (Décomposition biais-variance) Le vocabulaire « biais » et « variance » pour les deux termes de (44) se comprend bien en écrivant une décomposition similaire conditionnellement à $X_{1\dots n}$ et X :

$$\begin{aligned} & \mathbb{E}\left[\left(\eta(X) - \widehat{\eta}(D_n; X)\right)^2 \mid X, X_{1\dots n}\right] \\ &= (\eta(X) - \eta^*(X_{1\dots n}; X))^2 + \mathbb{E}\left[\left(\eta^*(X_{1\dots n}; X) - \widehat{\eta}(D_n; X)\right)^2 \mid X, X_{1\dots n}\right] \\ &= \underbrace{(\eta(X) - \eta^*(X_{1\dots n}; X))^2}_{\text{biais}} + \underbrace{\text{var}(\widehat{\eta}(D_n; X) \mid X, X_{1\dots n})}_{\text{variance}}. \end{aligned} \quad (46)$$

Le premier terme du membre de droite de (46) est le carré du biais (conditionnel) de $\widehat{\eta}(D_n; X)$ comme estimateur de $\eta(X)$, puisque

$$\mathbb{E}[\widehat{\eta}(D_n; X) \mid X, X_{1\dots n}] = \eta^*(X_{1\dots n}; X).$$

Le deuxième terme du membre de droite de (46) est la variance (conditionnelle) de $\widehat{\eta}(D_n; X)$. En toute rigueur, le « terme de biais » de (44) désigne donc la moyenne du carré du biais conditionnel, tandis que le « terme de variance » de (44) correspond à l'espérance de la variance conditionnelle.

Parenthèse 87 (Analogie entre (44) et (11)) La décomposition (44) est similaire à la décomposition (11) vue en section 3.4 pour un minimiseur du risque empirique. Le terme de biais correspond à l'erreur d'approximation, tandis que le terme de variance correspond à l'erreur d'estimation. Il est d'ailleurs courant de nommer biais l'erreur d'approximation, de nommer variance l'erreur d'estimation, et réciproquement.

Signalons tout de même que pour les règles par partition (qui sont à la fois des minimiseurs du risque empirique et des règles par moyenne locale), (44) ne coïncide pas exactement avec (11). On peut le constater grâce à la reformulation (60)

de (11) démontrée à l'exercice 8. La légère différence entre les deux décompositions est liée à la possibilité que certains éléments de la partition ne contiennent aucune observation.

En supposant de plus qu'on est en *régression homoscédastique*, c'est-à-dire que la variance résiduelle $\sigma^2(X) \equiv \sigma^2$ ne dépend pas de X , on déduit de (45) que le terme de variance est proportionnel à la variance résiduelle :

$$\begin{aligned} & \mathbb{E}\left[\left(\eta(X) - \hat{\eta}(D_n; X)\right)^2\right] \\ &= \mathbb{E}\left[\left(\eta(X) - \eta^*(X_{1\dots n}; X)\right)^2\right] + \sigma^2 \sum_{i=1}^n \mathbb{E}[W_{i,n}^2(X_{1\dots n}; X)]. \end{aligned} \quad (47)$$

Espace \mathcal{X} général

Le théorème 2 et son équivalent en régression sont énoncés lorsque $\mathcal{X} = \mathbb{R}^p$ est muni de la norme euclidienne. On suppose maintenant que (\mathcal{X}, d) est seulement un espace métrique, muni de la tribu borélienne, et l'on remplace la condition (b) du théorème 2 par la condition suivante : pour tout $a > 0$,

$$\lim_{n \rightarrow +\infty} \mathbb{E}\left[\sum_{i=1}^n W_{i,n}(X_{1\dots n}; X) \mathbb{1}_{d(X_i, X) \geq a}\right] = 0.$$

Peut-on généraliser le théorème de Stone à ce cadre ? Oui !

En effet, une lecture attentive de la démonstration du théorème 2 montre qu'elle fonctionne pour un espace métrique quelconque si η est uniformément continue et bornée, ou bien si η peut être approchée dans $L^1(P_X)$ par des fonctions uniformément continues et bornées. Il en est de même en régression L^q ($1 \leq q < +\infty$), la dernière condition devenant « η peut être approchée dans $L^q(P_X)$ par des fonctions uniformément continues et bornées ». Or, l'ensemble des fonctions uniformément continues et bornées⁴¹ est dense dans $L^q(P_X)$ pour tout $q \in [1, +\infty[$ lorsque P_X est extérieurement régulière (Le Gall, 2006, théorème 4.3.1), et toute mesure de probabilité sur un espace métrique est extérieurement régulière (Billingsley, 1999, théorème 1.1). Donc, le théorème de Stone se généralise à tout espace métrique (\mathcal{X}, d) muni de la tribu borélienne.

5.3 Règles par partition

Dans le cas des règles par partition (exemple 7 en section 5.1), le théorème 2 s'applique. On peut en déduire leur consistance universelle sous certaines conditions simples.

⁴¹ Le Gall (2006, théorème 4.3.1) démontre même ce résultat pour les fonctions *lipschitziennes* bornées, qui sont uniformément continues.

Corollaire 1 Soit $(\mathcal{A}_n)_{n \in \mathbb{N}}$ une suite de partitions mesurables de $\mathcal{X} = \mathbb{R}^p$, finies ou dénombrables. On suppose :

$$(b') \sup_{A \in \mathcal{A}_n} \{\text{diam}(A)\} \xrightarrow[n \rightarrow +\infty]{} 0,$$

$$(c') \text{ pour tout } r > 0, \frac{\text{Card}\{A \in \mathcal{A}_n / A \cap \mathcal{B}(0, r) \neq \emptyset\}}{n} \xrightarrow[n \rightarrow +\infty]{} 0,$$

où $\mathcal{B}(0, r)$ est la boule de centre 0 et de rayon r pour une norme $\|\cdot\|$ sur \mathcal{X} et $\text{diam}(A) = \sup_{x_1, x_2 \in A} \|x_1 - x_2\|$. Alors, $\hat{f}_{(\mathcal{A}_n)}^{p-c}$, la règle de classification par partition associée à la suite $(\mathcal{A}_n)_{n \geq 1}$, est faiblement universellement consistante pour le risque 0–1.

Les conditions du corollaire 1 sont en particulier vérifiées pour les règles par partition cubique de pas $h_n \rightarrow 0$ avec $nh_n^p \rightarrow +\infty$ (et ce sont des conditions nécessaires pour avoir la consistance universelle, voir l'exercice 32). Ceci démontre que la consistance universelle (faible) est possible, ce qui est loin d'être évident *a priori* !

Comme indiqué à la fin de la section 5.2, le théorème de Stone est également valable en régression avec le risque L^p , $p \geq 1$ (Biau et Devroye, 2015, théorème 10.1). Par conséquent, sous les hypothèses du corollaire 1, la règle de régression par partition $\hat{f}_{(\mathcal{A}_n)}^{p-r}$ est faiblement universellement consistante pour le risque quadratique (c'est-à-dire, consistante pour toute loi P telle que $\mathbb{E}|Y|^2 < +\infty$), ainsi que pour tous les risques L^p (dès que $\mathbb{E}|Y|^p < +\infty$).

Les conditions imposées par le corollaire 1 à la suite de partitions $(\mathcal{A}_n)_{n \geq 1}$ s'interprètent aisément. La condition (b'), sur le diamètre maximal d'un élément de \mathcal{A}_n , devient $h_n \rightarrow 0$ pour une partition cubique. Elle correspond à l'absence de sous-apprentissage, via la condition (b) du théorème 2. De plus, la condition (b') entraîne que l'erreur d'approximation tend vers zéro, en régression avec le coût quadratique (d'après l'exercice 6) et en classification avec le coût 0–1 (d'après l'exercice 7).

La condition (c'), sur le nombre d'éléments de la partition qui intersectent une boule de rayon r fixé, devient $nh_n^p \rightarrow +\infty$ pour une partition cubique. Elle permet d'éviter le surapprentissage, via la condition (c) du théorème 2. De plus, la condition (c') entraîne que l'erreur d'estimation tend vers zéro en régression avec le coût quadratique (d'après l'exercice 8) et en classification avec le coût 0–1 (d'après l'exercice 18).

L'exercice 33 montre que l'on peut relâcher les conditions (b') et (c') du corollaire 1 sur $(\mathcal{A}_n)_{n \geq 1}$ et P .

Malgré le corollaire 1, les règles par partition cubique ne fonctionnent correctement que lorsque p est petit. En effet, pour avoir simultanément h_n « petit » (disons, inférieur à $1/10$) et nh_n^p « grand » (disons, supérieur à 10), on doit disposer d'au moins 10^{p+1} observations. Si l'on est en dimension $p \geq 10$, une telle condition est totalement irréaliste, même avec des données massives ! On parle de *fléau de la dimension*. Pour apprendre malgré tout quelque chose en grande dimension, il faut faire des hypothèses et utiliser des méthodes spécifiques (Giraud, 2014).

Démonstration du corollaire 1

Il s'agit de vérifier les conditions du théorème de Stone, qui entraînent la consistance en classification 0–1 (théorème 2).

Condition (a) D'après (42), on a

$$1 - \sum_{i=1}^n W_{i,n}^{\mathcal{A}_n}(X_{1\dots n}; X) = \mathbb{1}_{\mathcal{A}_n(X) \cap \{X_1, \dots, X_n\} = \emptyset}$$

donc $\sum_{i=1}^n W_{i,n}^{\mathcal{A}_n}(X_{1\dots n}; X) \leq c_a = 1$ presque sûrement. Reste à vérifier la première partie de la condition (a). Soit $\epsilon > 0$. Il existe alors $r > 0$ tel que $P_X(\mathcal{B}(0, r)^c) < \epsilon$. Notons

$$\tilde{\mathcal{A}}_n(r) = \{A \in \mathcal{A}_n / A \cap \mathcal{B}(0, r) \neq \emptyset\}.$$

Alors,

$$\begin{aligned} \mathbb{E} \left| 1 - \sum_{i=1}^n W_{i,n}^{\mathcal{A}_n}(X_{1\dots n}; X) \right| &= \mathbb{P}(\mathcal{A}_n(X) \cap \{X_1, \dots, X_n\} = \emptyset) \\ &= \sum_{A \in \mathcal{A}_n} P_X(A)(1 - P_X(A))^n \\ &\leq \text{Card}(\tilde{\mathcal{A}}_n(r)) \underbrace{\sup_{t>0} \{t(1-t)^n\}}_{\leq 1/(en)} + \underbrace{\sum_{\substack{A \notin \tilde{\mathcal{A}}_n(r) \\ \leq P_X(\mathcal{B}(0, r)^c) \leq \epsilon}} P_X(A)}_{\leq P_X(\mathcal{B}(0, r)^c) \leq \epsilon} \\ &\leq \frac{\text{Card}(\tilde{\mathcal{A}}_n(r))}{en} + \epsilon \end{aligned}$$

qui tend vers ϵ quand n tend vers l'infini, par hypothèse. Ceci est vrai pour tout $\epsilon > 0$, ce qui conclut la démonstration de (a).

Condition (b) $\sum_{i=1}^n W_{i,n}^{\mathcal{A}_n}(X_{1\dots n}; X) \mathbb{1}_{\|X_i - X\| \geq a}$ est nul presque sûrement (donc aussi en espérance) si $\sup_{A \in \mathcal{A}_n} \text{diam}(A) < a$. Pour tout $a > 0$, ceci a lieu dès que n est assez grand par hypothèse.

Condition (c) Soit $\epsilon > 0$. On choisit r comme pour démontrer (a). Alors,

$$\begin{aligned} \mathbb{E} \left[\max_{1 \leq i \leq n} W_{i,n}^{\mathcal{A}_n}(X_{1\dots n}; X) \right] &\leq \mathbb{E} \left[\sum_{A \in \tilde{\mathcal{A}}_n(r)} \mathbb{1}_{X \in A} \frac{\mathbb{1}_{N_A(X_{1\dots n}) > 0}}{N_A(X_{1\dots n})} + \mathbb{1}_{X \notin \mathcal{B}(0, r)} \right] \\ &\leq \sum_{A \in \tilde{\mathcal{A}}_n(r)} P_X(A) \mathbb{E} \left[\frac{\mathbb{1}_{N_A(X_{1\dots n}) > 0}}{N_A(X_{1\dots n})} \right] + \epsilon. \end{aligned}$$

Il s'agit donc de majorer l'espérance de l'inverse de $N_A(X_{1\dots n})$, qui suit une loi binomiale de paramètres $(n, P_X(A))$, ce que fait le lemme 5 en section 8.5. On obtient alors :

$$\mathbb{E} \left[\max_{1 \leq i \leq n} W_{i,n}^{\mathcal{A}_n}(X_{1\dots n}; X) \right] \leq \frac{2 \text{Card}(\tilde{\mathcal{A}}_n(r))}{n+1} + \epsilon \xrightarrow{n \rightarrow +\infty} \epsilon.$$

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

Condition (d) Enfin, pour tout $f \in L^1(P_X)$ à valeurs positives ou nulles :

$$\begin{aligned}
& \mathbb{E} \left[\sum_{i=1}^n W_{i,n}^{\mathcal{A}_n}(X_{1\dots n}; X) f(X_i) \right] \\
&= \mathbb{E} \left[\sum_{i=1}^n \frac{\mathbb{1}_{X_i \in \mathcal{A}_n(X)} f(X_i)}{N_{\mathcal{A}_n(X)}(X_{1\dots n})} \right] \\
&= \mathbb{E} \left[\sum_{A \in \mathcal{A}_n} \mathbb{1}_{X \in A} \frac{\sum_{i=1}^n f(X_i) \mathbb{1}_{X_i \in A}}{N_A(X_{1\dots n})} \right] \\
&= \sum_{A \in \mathcal{A}_n} \mathbb{P}(X \in A) \mathbb{E} \left[\frac{\sum_{i=1}^n f(X_i) \mathbb{1}_{X_i \in A}}{N_A(X_{1\dots n})} \right] \\
&= \sum_{A \in \mathcal{A}_n} \mathbb{P}(X \in A) \mathbb{E} \left[\frac{\mathbb{1}_{N_A(X_{1\dots n}) > 0}}{N_A(X_{1\dots n})} \mathbb{E} \left[\sum_{i=1}^n f(X_i) \mathbb{1}_{X_i \in A} \mid N_A(X_{1\dots n}) \right] \right] \\
&\leq \sum_{A \in \mathcal{A}_n} \mathbb{P}(X \in A) \mathbb{E}[f(X) \mid X \in A] = \mathbb{E}[f(X)],
\end{aligned}$$

où l'on a utilisé que, conditionnellement à $N_A(X_{1\dots n}) = q > 0$, $\sum_{i=1}^n f(X_i) \mathbb{1}_{X_i \in A}$ a la loi de la somme de q variables indépendantes et de même loi $\mathcal{L}(f(X) \mid X \in A)$. \square

Parenthèse 88 (Espace métrique général et partition) Lorsque (\mathcal{X}, d) est un espace métrique, que peut-on dire sur les règles par partition ? Comme indiqué à la fin de la section 5.2, le théorème de Stone s'applique encore. La démonstration du corollaire 1 se généralise donc à ce cadre, la condition (c') pouvant être remplacée par : pour toute partie compacte K de \mathcal{X} , $\lim_{n \rightarrow +\infty} \frac{\text{Card}\{A \in \mathcal{A}_n / A \cap K \neq \emptyset\}}{n} = 0$.

5.4 Plus proches voisins

La méthode des k plus proches voisins (« k -nearest neighbors » en anglais, souvent abrégé « k -NN ») est un deuxième exemple fondamental de règle par moyenne locale. Nous renvoyons au livre de Biau et Devroye (2015) pour un exposé détaillé sur ce sujet.

Exemple 8 (k plus proches voisins) Soit d une distance sur \mathcal{X} et $k \geq 1$ un entier. Pour tout $x, x_1, \dots, x_n \in \mathcal{X}$, on pose :

$$W_i^{k-\text{PPV}}(x_{1\dots n}; x) := \frac{1}{k} \times \mathbb{1}_{\{x_i \text{ fait partie des } k \text{ plus proches voisins de } x \text{ parmi } x_{1\dots n}\}},$$

où la notion de plus proche voisin est relative à la distance d . La règle par moyenne locale associée aux poids $(W_i^{k-\text{PPV}})_{1 \leq i \leq n}$ est appelée règle des k plus proches voisins (associée à la distance d).

Il s'agit ici d'une vraie moyenne puisque :

$$\forall x, x_1, \dots, x_n \in \mathcal{X}, \quad \sum_{i=1}^n W_i^{k-\text{PPV}}(x_{1\dots n}; x) = 1.$$

Son caractère local est déterminé par la distance d , la notion de voisinage étant plus ou moins étendue suivant la valeur de k .

Parenthèse 89 (Extension de la définition) En général, le nombre de voisins $k = k_n$ dépend de la taille n de l'échantillon. On parle alors de règle des $(k_n)_{n \geq 1}$ plus proches voisins. La distance d est habituellement fixée lorsque n varie (mais rien n'interdit de prendre $d = d_n$). De plus, le fait que d est une distance n'est en réalité pas nécessaire pour que l'exemple 8 définisse une règle par moyenne locale. N'importe quelle mesure de proximité entre éléments de \mathcal{X} peut convenir.

Parenthèse 90 (Cas d'égalité de distances) La définition des k plus proches voisins de x parmi $x_{1\dots n}$ peut poser problème en cas d'égalité entre valeurs de distance $d(x, x_i) = d(x, x_j)$ avec $i \neq j$. Dans ce cas, on suppose que l'on a défini au préalable une règle pour choisir qui de x_i ou x_j est « plus proche » de x (par exemple, celui dont l'indice i ou j est le plus petit). Vu qu'il y a plusieurs manières de le faire, il serait plus rigoureux de parler d'*une* règle des k plus proches voisins plutôt que de *la* règle des k plus proches voisins. L'analyse théorique faite dans ce texte ne dépend pas de ce choix. On conserve donc cette petite ambiguïté qui ne porte pas à conséquence.

Parenthèse 91 (Cas d'égalité en classification) En classification, un autre cas d'égalité potentiellement problématique est celui qui se produit lorsque k est pair et qu'il y a exactement autant d'observations étiquetées 1 que d'observations étiquetées 0 parmi les k plus proches voisins de x . Ici, implicitement (via la définition du plug-in, voir la section 2.3), on a fait le choix d'un classifieur prenant la valeur 0 dans un tel cas d'égalité. D'autres choix sont possibles, sans conséquences au niveau des garanties théoriques. À cause de cette possibilité d'égalité, la littérature théorique sur les k plus proches voisins fait souvent l'hypothèse que k est impair, ce qui évacue le problème.

Les performances des k -plus proches voisins dépendent bien sûr fortement du choix de k . Ainsi, prendre k trop petit conduit en général au surapprentissage, tandis que les grandes valeurs de k (de l'ordre de n) font risquer le sous-apprentissage. Lorsque la taille d'échantillon n tend vers l'infini, ceci se traduit par les conditions suivantes pour obtenir la consistance universelle lorsque $\mathcal{X} = \mathbb{R}^p$ et que d est induite par une norme⁴².

42. On dit que d est induite par la norme $\|\cdot\|$ lorsque, pour tout $x, x' \in \mathcal{X}$, $d(x, x') = \|x - x'\|$.

Corollaire 2 Soit d une distance induite par une norme sur $\mathcal{X} = \mathbb{R}^p$ et $(k_n)_{n \geq 1}$ une suite d'entiers. Alors, la règle de classification des k_n plus proches voisins associée à la distance d est universellement consistante pour le risque 0–1 si et seulement si

$$\lim_{n \rightarrow +\infty} k_n = +\infty \quad \text{et} \quad \lim_{n \rightarrow +\infty} \frac{k_n}{n} = 0. \quad (48)$$

La règle de régression des k_n plus proches voisins associée à la distance d , notée $\widehat{f}_{\text{reg}}^{k_n-\text{PPV}}$, est universellement consistante pour le risque L^p — c'est-à-dire, $\mathbb{E}|\widehat{f}_{\text{reg}}^{k_n-\text{PPV}} - \eta|^p \rightarrow 0$ quand $n \rightarrow +\infty$ pour toute loi P telle que $\mathbb{E}|Y|^p < +\infty$ — si et seulement si la condition (48) est vérifiée.

Le corollaire 2 est une conséquence du théorème de Stone (théorème 2). La condition (a) est toujours vérifiée avec $c_a = 1$. La condition (c) découle immédiatement de l'hypothèse $k_n \rightarrow +\infty$. Vérifier les conditions (b) et (d) demande un peu de travail. En particulier, la condition (d) découle d'un lemme géométrique, appelé lemme de Stone (Devroye *et al.*, 1996, lemme 5.3). Biau et Devroye (2015, sections 10.4 et 19.1) donnent une démonstration détaillée du corollaire 2.

En régression, on peut également obtenir des majorations précises de l'erreur (ponctuelle ou intégrée) commise par $\widehat{f}_{\text{reg}}^{k_n-\text{PPV}}$ (Biau et Devroye, 2015, chapitre 14). Pour en donner un bref aperçu, considérons le cadre de la régression homoscédastique avec le coût quadratique, pour laquelle on dispose de la décomposition (47) du risque pour un prédicteur des k_n plus proches voisins. Puisque $W_{i,n} \in \{0, 1/k_n\}$ et $\sum_{i=1}^n W_{i,n} = 1$, on a :

$$\sum_{i=1}^n W_{i,n}^2 = \frac{1}{k_n} \sum_{i=1}^n W_{i,n} = \frac{1}{k_n}.$$

L'erreur d'estimation vaut donc σ^2/k_n et le risque s'écrit :

$$\mathbb{E}\left[\left(\eta(X) - \widehat{f}_{\text{reg}}^{k_n-\text{PPV}}(D_n; X)\right)^2\right] = \mathbb{E}\left[\left(\eta(X) - \eta^*(X_{1\dots n}; X)\right)^2\right] + \frac{\sigma^2}{k_n}.$$

Ceci confirme l'intuition selon laquelle l'erreur d'estimation est une fonction décroissante de k_n . On voit également sur cette formule pourquoi $k_n \rightarrow +\infty$ est une condition nécessaire de consistance universelle dans ce cadre.

On peut aussi majorer l'erreur d'estimation sans l'hypothèse d'homoscédasticité : si la variance résiduelle vérifie $\sigma^2(X) \leq \sigma_{\max}^2$ presque sûrement, alors on déduit de (45) une majoration de l'erreur d'estimation par σ_{\max}^2/k_n .

On peut noter que le corollaire 2 laisse une grande plage de possibilités pour k_n , sans rien dire sur son choix théorique optimal (qui dépend du problème de classification considéré). Le corollaire 2 ne dit pas grand chose non plus pour le choix de k_n en pratique (pour un échantillon de taille n fixé). On peut justifier l'emploi d'une valeur de k_n choisie à l'aide des données, en classification (Devroye *et al.*, 1996, chapitre 26) comme en régression (Biau et Devroye, 2015, chapitre 16). Ce choix peut se faire par validation croisée (Arlot, 2017), ou par une autre procédure de sélection d'estimateurs.

Parenthèse 92 (Espace métrique général) Le corollaire 2 s'étend (en régression L^2 et donc aussi en classification 0–1) au cas où (\mathcal{X}, d) est un espace métrique séparable, lorsque η est bornée et vérifie une condition supplémentaire (Forzani *et al.*, 2012, théorème 4.1).

5.5 Noyau

Un troisième exemple classique de règle par moyenne locale est la famille des règles par noyau. Celles-ci sont souvent appelées estimateurs de Nadaraya-Watson en régression, et sont fortement liées aux estimateurs de Parzen-Rosenblatt de la densité d'un échantillon. Pour plus de détails et des références, le lecteur intéressé est invité à consulter les livres de Devroye *et al.* (1996, chapitre 10, pour la classification) et Györfi *et al.* (2002, chapitre 5, pour la régression).

Exemple 9 (Règle par noyau) On suppose que $\mathcal{X} \subset \mathbb{R}^p$. Soit $K : \mathbb{R}^p \rightarrow \mathbb{R}$ une application mesurable et $h > 0$. Pour tout $x, x_1, \dots, x_n \in \mathcal{X}$, on pose :

$$W_i^{K,h}(x_{1\dots n}; x) = \frac{K\left(\frac{x_i - x}{h}\right)}{\sum_{j=1}^n K\left(\frac{x_j - x}{h}\right)}$$

avec la convention $0/0 = 0$. La règle par moyenne locale associée aux poids $(W_i^{K,h})_{1 \leq i \leq n}$ est appelée *règle par noyau*, de noyau K et de fenêtre (ou largeur de bande) h .

Si K est à valeurs positives ou nulles, il s'agit (presque toujours) d'une vraie moyenne puisque :

$$\forall x, x_1, \dots, x_n \in \mathcal{X}, \quad \sum_{i=1}^n W_i^{K,h}(x_{1\dots n}; x) = \mathbb{1}_{\sum_{j=1}^n K\left(\frac{x_j - x}{h}\right) > 0}.$$

La condition (a) du théorème 2 est donc toujours vérifiée avec $c_a = 1$. Le caractère local d'une règle par noyau est déterminé par le noyau K , la notion de voisinage étant plus ou moins étendue suivant la valeur de h .

Parenthèse 93 (Fenêtre variant avec n) En général, la fenêtre $h = h_n$ dépend de la taille n de l'échantillon. On parle alors de règle par noyau de noyau K et de largeurs de bande $(h_n)_{n \geq 1}$. Le noyau K est habituellement fixé lorsque n varie (mais rien n'interdit de prendre $K = K_n$).

Les exemples les plus courants de noyau K sont :

- le noyau fenêtre $K(x) = \mathbb{1}_{\|x\| \leq 1}$,
- le noyau gaussien $K(x) = \exp(-\|x\|^2)$,
- le noyau de Cauchy $K(x) = 1/(1 + \|x\|^{p+1})$,

— le noyau Epanechnikov $K(x) = (1 - \|x\|^2)\mathbf{1}_{\|x\| \leq 1}$,
 où $\|\cdot\|$ est la norme euclidienne sur \mathbb{R}^p . Remarquons que l'on suppose dans cette section que K est à valeurs positives, mais il est parfois intéressant de considérer un noyau K pouvant prendre des valeurs strictement négatives (Devroye *et al.*, 1996, section 10.1).

Parenthèse 94 (Lien avec l'estimation de densité) On peut relier les règles de classification par noyau à l'estimation de densité par noyau (aussi appelée méthode de Parzen-Rosenblatt). En effet, si $\mathcal{Y} = \{0, 1\}$, on peut réécrire ainsi la règle de régression par noyau, de noyau K et de fenêtre h :

$$\begin{aligned} & \hat{\eta}^{K,h}((x_i, y_i)_{1 \leq i \leq n}; x) \\ &= \frac{\sum_{i=1}^n y_i K\left(\frac{x_i-x}{h}\right)}{\sum_{j=1}^n K\left(\frac{x_j-x}{h}\right)} \\ &= \frac{\hat{p}\hat{g}_1((x_i, y_i)_{1 \leq i \leq n}; x)}{\hat{p}\hat{g}_1((x_i, y_i)_{1 \leq i \leq n}; x) + (1-\hat{p})\hat{g}_0((x_i, y_i)_{1 \leq i \leq n}; x)} \end{aligned} \quad (49)$$

où

$$\hat{p} := \frac{1}{n} \operatorname{Card}\{i \in \{1, \dots, n\} / Y_i = 1\}$$

est un estimateur consistant de $p = \mathbb{P}(Y = 1)$ et, pour $j \in \{0, 1\}$,

$$\hat{g}_j((x_i, y_i)_{1 \leq i \leq n}; x) := \frac{1}{h} \frac{\sum_{i=1}^n K\left(\frac{x_i-x}{h}\right) \mathbf{1}_{Y_i=j}}{\sum_{i=1}^n K(\mathbf{1}_{Y_i=j})}$$

est l'estimateur à noyau classique de g_j , la densité conditionnelle de X sachant $Y = j$ (par rapport à une mesure de référence μ sur \mathcal{X} telle que $\int K d\mu = 1$). Or, on a vu à la remarque 39 en section 2.2 que pour P_X presque tout x :

$$\eta(x) = \frac{pg_1(x)}{(1-p)g_0(x) + pg_1(x)}.$$

La formule (49) permet donc d'interpréter $\hat{\eta}^{K,h}$ comme une estimation par plugin de η fondée sur les estimateurs \hat{p} , \hat{g}_0 et \hat{g}_1 de p , g_0 et g_1 . Ceci fournit d'ailleurs une méthode pour démontrer la consistance de $\hat{\eta}^{K,h}$: plutôt que de passer par le théorème de Stone, on peut commencer par démontrer la consistance de \hat{g}_0 et \hat{g}_1 (Rivoirard et Stoltz, 2009, section 24.1.8).

Comme pour les règles par partition et les k plus proches voisins, les performances des règles par noyau dépendent fortement du noyau K et de la largeur de bande h . Le théorème de Stone (théorème 2) donne une condition suffisante de consistance universelle en classification.

Corollaire 3 Soit $\mathcal{X} \subset \mathbb{R}^p$, $K : \mathcal{X} \rightarrow \mathbb{R}$ une application mesurable, $(h_n)_{n \geq 1}$ une suite de réels strictement positifs. Pour tout $\delta > 0$, on note $\mathcal{B}(0, \delta)$ la boule euclidienne de centre 0 et de rayon δ . On suppose qu'il existe $R > r > 0$ et $b > 0$ tels que

$b\mathbb{1}_{B(0,r)} \leq K \leq \mathbb{1}_{B(0,R)}$, ainsi que :

$$\lim_{n \rightarrow +\infty} h_n = 0 \quad \text{et} \quad \lim_{n \rightarrow +\infty} nh_n^p = +\infty.$$

Alors, la règle de d'apprentissage par noyau (définie à l'exemple 9), de noyau K et de fenêtre $(h_n)_{n \geq 1}$, est universellement consistante pour le risque 0–1 en classification et pour le risque L^p en régression.

Devroye *et al.* (1996, chapitre 10) donnent une démonstration du corollaire 3 en classification, initialement dû à Devroye et Krzyzak. Le cas de la régression L^p est traité par Györfi *et al.* (2002, théorème 5.1).

Parenthèse 95 (Généralisation du corollaire 3) On peut en fait démontrer la consistance universelle *forte* de la règle par noyau associée à K sous des hypothèses plus faibles que celles du corollaire 3 (Devroye *et al.*, 1996, théorème 10.1). Au lieu d'imposer que $b\mathbb{1}_{B(0,r)} \leq K \leq \mathbb{1}_{B(0,R)}$, il suffit d'avoir un noyau K « régulier ». Par exemple, un noyau $K : \mathbb{R}^p \rightarrow \mathbb{R}$, intégrable, uniformément continu et tel que $b\mathbb{1}_{B(0,r)} \leq K$ pour des constantes $b, r > 0$ est régulier. Ainsi, les noyaux fenêtre, gaussien, Cauchy et Epanechnikov sont réguliers, et le corollaire 3 s'applique aux règles de classification associées.

Les conditions sur $(h_n)_{n \geq 1}$ s'interprètent aisément : $h_n \rightarrow 0$ permet d'éviter le sous-apprentissage (via les conditions (b) et (d) du théorème 2), tandis que $nh_n^p \rightarrow +\infty$ permet d'éviter le surapprentissage (via la condition (c) du théorème 2).

Le corollaire 3 laisse un large éventail de possibilités pour $(h_n)_{n \geq 1}$ (et pour K). Un résultat quantitatif plus précis serait nécessaire pour identifier leurs valeurs théoriques optimales (Györfi *et al.*, 2002, section 5.3).

Le choix de h_n et K en pratique (pour un échantillon de taille n fixé) n'est pas non plus résolu par le corollaire 3. Il est cependant possible d'établir la consistance faible d'une règle de classification par noyau lorsque h_n et K sont choisis à l'aide des données (Devroye *et al.*, 1996, chapitre 25), par exemple par validation croisée (Arlot, 2017).

Parenthèse 96 (Espace métrique général et noyaux) On peut définir des règles par noyaux lorsque (\mathcal{X}, d) est un espace métrique quelconque, en remplaçant

$$K\left(\frac{x_i - x}{h}\right) \quad \text{par} \quad K\left(\frac{d(x_i, x)}{h}\right)$$

dans les formules de l'exemple 9, la fonction K étant alors une fonction mesurable $[0, +\infty[\rightarrow \mathbb{R}$. Forzani *et al.* (2012, théorème 5.1) établissent alors la consistance d'une telle règle d'apprentissage (en régression L^2 et donc aussi en classification 0–1), pour un noyau K vérifiant les hypothèses du corollaire 3, lorsque (\mathcal{X}, d) est séparable, η bornée, avec des hypothèses supplémentaires sur η et $(h_n)_{n \geq 1}$.

6 On n'a rien sans rien

Les résultats de la section 5 montrent que la consistance faible universelle est possible en classification avec le risque 0–1. En effet, il existe des règles d'apprentissage \hat{f} telles que :

$$\sup_{P \text{ loi sur } \mathcal{X} \times \mathcal{Y}} \lim_{n \rightarrow +\infty} \mathbb{E}_{D_n \sim P^{\otimes n}} [\ell_P(f_P^*, \hat{f}(D_n))] = 0. \quad (50)$$

D'un point de vue pratique, il serait intéressant de disposer d'informations sur la *vitesse d'apprentissage* de \hat{f} : étant donné $\epsilon > 0$, à partir de quel nombre d'observations⁴³ n_0 peut-on garantir que l'excès de risque moyen est inférieur à ϵ ? La consistance faible universelle (50) garantit que pour tout P et tout $\epsilon > 0$, un tel $n_0(\epsilon, P) < +\infty$ existe. Mais, P étant inconnue et ce $n_0(\epsilon, P)$ non explicite, l'intérêt pratique de ce résultat est limité⁴⁴.

Idéalement, on aimerait disposer d'une *vitesse d'apprentissage universelle*, c'est-à-dire d'un majorant $n_1(\epsilon)$ de $n_0(\epsilon, P)$ valable *pour tout* P (et donc utilisable en pratique). L'existence d'un tel $n_1(\epsilon)$ équivaut à démontrer que \hat{f} est *universellement uniformément consistante* :

$$\lim_{n \rightarrow +\infty} \sup_{P \text{ loi sur } \mathcal{X} \times \mathcal{Y}} \mathbb{E}_{D_n \sim P^{\otimes n}} [\ell_P(f_P^*, \hat{f}(D_n))] = 0. \quad (51)$$

Est-ce possible? Par rapport à la consistance universelle faible (50), il s'agit d'échanger la limite et le supremum.

Les résultats qui suivent démontrent que c'est impossible en classification 0–1, sauf lorsque \mathcal{X} est fini. Autrement dit, on n'a rien sans rien⁴⁵: si l'on veut une information sur la vitesse d'apprentissage, il faut disposer d'informations sur la loi P qui a généré les observations (en plus de l'échantillon D_n lui-même). Le lecteur souhaitant approfondir cette question peut consulter le livre de Devroye *et al.* (1996, chapitre 7).

6.1 À taille d'échantillon fixée

La règle de classification binaire la plus idiote est la règle « pile ou face »⁴⁶, qui attribue à chaque $x \in \mathcal{X}$ une étiquette $\hat{f}^{\text{pf}}(x)$ tirée selon une loi de Bernoulli de paramètre 1/2, indépendamment de x et des valeurs $(\hat{f}^{\text{pf}}(x'))_{x' \neq x}$.

Lorsque \mathcal{X} est infini et que la taille d'échantillon n est fixée, le théorème suivant montre que pour toute règle de classification \hat{f} et tout entier $n \geq 1$, il existe un problème de classification pour lequel \hat{f} ne fait pas mieux que \hat{f}^{pf} en pire cas avec n observations.

43. En anglais, le nombre d'observations n_0 à partir duquel l'excès de risque moyen est inférieur à ϵ est appelé « sample complexity ».

44. En ce sens, la consistance universelle est une garantie *minimale* sur \hat{f} quand on ne dispose d'aucune information sur P : il est bon de l'avoir, mais on ne peut pas s'en satisfaire.

45. En anglais, on utilise l'expression « there is no free lunch » et l'on parle de « no free lunch theorems » pour désigner les résultats théoriques correspondants.

46. La règle pile ou face est une règle randomisée, qui sort donc un peu du cadre introduit en section 1.3 : elle réalise des prévisions aléatoires, et l'on mesure son risque en prenant une moyenne sur l'aléa de prévision. Le théorème 3 s'applique à une telle règle randomisée.

Théorème 3 Soit \mathcal{X} un ensemble infini, $n \geq 1$ un entier et \hat{f} une règle de classification binaire ($\mathcal{Y} = \{0, 1\}$). On note $\mathcal{M}_1(\mathcal{X} \times \{0, 1\})$ l'ensemble des mesures de probabilité sur $\mathcal{X} \times \{0, 1\}$. Alors, avec le coût de classification 0–1 :

$$\sup_{P \in \mathcal{M}_1(\mathcal{X} \times \{0, 1\})} \left\{ \mathbb{E}_{D_n \sim P^{\otimes n}} [\ell_P(f_P^*, \hat{f}(D_n))] \right\} \geq \frac{1}{2}. \quad (52)$$

En particulier, aucune règle de classification n'est uniformément universellement consistante lorsque \mathcal{X} est infini en classification 0–1.

Le théorème 3 a été initialement démontré par Devroye (voir aussi Devroye *et al.*, 1996, théorème 7.1).

Démonstration L'idée de la démonstration est de considérer une loi P pour laquelle \hat{f} ne peut faire mieux que « deviner » les étiquettes des points x non observés (qui sont largement majoritaires dès que le support de X est de taille beaucoup plus grande que n). Le résultat devant être valable pour toute règle \hat{f} , une construction explicite de P serait délicate. Un argument élégant permet d'éviter de le faire : on choisit P aléatoire et l'on démontre une minoration *en moyenne sur P* , pour en déduire une minoration du supremum sur P . Ce type de raisonnement, utilisé dans divers domaines des mathématiques, est appelé *argument probabiliste*.

Voici comment formaliser cet argument. Soit $K \in \mathbb{N}$. L'espace \mathcal{X} étant infini, à bijection près, on peut supposer que $\{1, \dots, K\} \subset \mathcal{X}$.

Restriction du supremum Pour tout $r \in \{0, 1\}^K$, définissons P_r la loi de probabilité sur $\mathcal{X} \times \{0, 1\}$ définie par $\mathbb{P}_{(X, Y) \sim P_r}(X = j \text{ et } Y = r_j) = K^{-1}$ pour tout $j \in \{1, \dots, K\}$. Autrement dit, X est choisi uniformément parmi $\{1, \dots, K\}$ et $Y = r_X$ est une fonction déterministe de X . En particulier, pour tout $r \in \{0, 1\}^K$, la loi P_r est « zéro-erreur » : $\mathcal{R}_{P_r}(f^*) = 0$. On définit alors l'application :

$$F : r \in \{0, 1\}^K \mapsto \mathbb{E}_{D_n \sim P_r^{\otimes n}} [\mathcal{R}_{P_r}(\hat{f}(D_n))].$$

Il s'agit donc de minorer :

$$\sup_{P \in \mathcal{M}_1(\mathcal{X} \times \{0, 1\})} \left\{ \mathbb{E}_{D_n \sim P^{\otimes n}} [\ell_P(f_P^*, \hat{f}(D_n))] \right\} \geq \sup_{r \in \{0, 1\}^K} F(r).$$

Argument probabiliste C'est ici que l'on utilise un raisonnement « probabiliste » :

$$\sup_{r \in \{0, 1\}^K} F(r) \geq \mathbb{E}_{R \sim Q}[F(R)]$$

pour toute loi Q sur $\{0, 1\}^K$. Il suffit donc de minorer cette dernière espérance. L'intérêt d'une telle approche est que l'on n'a pas besoin d'expliciter, pour chaque règle \hat{f} , un $r \in \{0, 1\}^K$ qui pose problème. Le fait que \hat{f} se comporte mal *en moyenne* sur les lois P_r suffit à établir l'*existence* d'un tel r problématique.

Choix d'une loi Q On choisit désormais R de loi Q , la distribution uniforme sur $\{0, 1\}^K$, de telle sorte que R_1, \dots, R_K sont indépendantes et de même loi de Bernoulli de paramètre $1/2$. Intuitivement, un tel choix rend la tâche extrêmement difficile pour \hat{f} : hors des points X_i observés, l'échantillon D_n ne donne aucune information sur la loi de Y sachant X , et il est donc impossible d'y faire mieux (en moyenne sur r) que le classifieur « pile ou face ». Formellement, on écrit :

$$\begin{aligned}\mathbb{E}[F(R)] &= \mathbb{P}(\hat{f}(D_n; X) \neq Y) \\ &= \mathbb{P}(\hat{f}((X_i, R_{X_i})_{1 \leq i \leq n}; X) \neq R_X) \\ &= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\{\hat{f}((X_i, R_{X_i})_{1 \leq i \leq n}; X) \neq R_X\}} \mid X, (X_i, R_{X_i})_{1 \leq i \leq n}\right]\right] \\ &\geq \mathbb{E}\left[\mathbb{1}_{X \notin \{x_1, \dots, x_n\}} \mathbb{E}\left[\mathbb{1}_{\{\hat{f}((X_i, R_{X_i})_{1 \leq i \leq n}; X) \neq R_X\}} \mid X, (X_i, R_{X_i})_{1 \leq i \leq n}\right]\right]\end{aligned}$$

Or, sachant $X = x$, $X_i = x_i$, $R_{x_i} = \alpha_i$ ($i = 1, \dots, n$), avec $x \notin \{x_1, \dots, x_n\}$, la quantité $\hat{f}((X_i, R_{X_i})_{1 \leq i \leq n}; X)$ est fixe, égale à 0 ou 1, tandis que $R_X = R_x$ est aléatoire, de loi de Bernoulli de paramètre $1/2$. Donc, l'inégalité ci-dessus se réécrit :

$$\begin{aligned}\mathbb{E}[F(R)] &\geq \mathbb{E}\left[\mathbb{1}_{X \notin \{x_1, \dots, x_n\}} \times \frac{1}{2}\right] \\ &= \frac{1}{2} \mathbb{E}[\mathbb{P}(X_1 \neq x, \dots, X_n \neq x \mid X)] \\ &= \frac{1}{2} \left(1 - \frac{1}{K}\right)^n.\end{aligned}$$

Récapitulons : on a établi que pour tout $K \geq 1$,

$$\sup_{P \in \mathcal{M}_1(\mathcal{X} \times \{0, 1\})} \left\{ \mathbb{E}_{D_n \sim P^{\otimes n}} [\ell_P(f_P^*, \hat{f}(D_n))] \right\} \geq \frac{1}{2} \left(1 - \frac{1}{K}\right)^n.$$

Cette borne inférieure tend vers $1/2$ lorsque K tend vers $+\infty$, d'où le résultat. \square

Parenthèse 97 (\mathcal{X} fini) Si \mathcal{X} est fini, la démonstration du théorème 3 établit que :

$$\sup_{P \in \mathcal{M}_1(\mathcal{X} \times \{0, 1\})} \left\{ \mathbb{E}_{D_n \sim P^{\otimes n}} [\ell_P(f_P^*, \hat{f}(D_n))] \right\} \geq \frac{1}{2} \left(1 - \frac{1}{\text{Card}(\mathcal{X})}\right)^n.$$

Ce minorant est de l'ordre d'une constante strictement positive dès que $\text{Card}(\mathcal{X})$ est de l'ordre de n ou plus grand (à comparer avec le résultat de la proposition 16 en section 6.2). L'hypothèse « \mathcal{X} infini » du théorème 3 peut ainsi se réécrire $\text{Card}(\mathcal{X}) \gg n$.

Remarque 98 Le résultat du théorème 3 peut s'étendre à divers cadres. Tout d'abord, la démonstration s'étend directement au cas où \hat{f} est une règle de classification ran-

domisée⁴⁷, c'est-à-dire, où la sortie $\hat{f}((x_i, y_i)_{1 \leq i \leq n}; x)$ est aléatoire (même quand x et les (x_i, y_i) sont déterministes). Outre la règle « pile ou face » déjà mentionnée, on peut penser à deux manières naturelles de construire une règle de classification randomisée : soit en tirant la sortie selon la loi *a posteriori* d'un classifieur bayésien, soit en utilisant la valeur d'un pseudo-classifieur (à valeurs dans $[0, 1]$) comme valeur de la probabilité d'avoir une sortie égale à 1. De plus, en classification 0–1, on peut imposer que X a une loi à densité et que la fonction de régression η est régulière (exercice 35). On peut également obtenir une borne inférieure pour l'ensemble des lois P telles que le risque de Bayes est égal à $c \in]0, 1/2[$ (exercice 36). Par ailleurs, on peut démontrer un résultat similaire en régression avec le coût quadratique (exercice 37). Enfin, l'estimation du risque de Bayes \mathcal{R}_P^* se heurte aux mêmes limites (Devroye *et al.*, 1996, section 8.5).

6.2 Consistance universelle uniforme en classification lorsque \mathcal{X} est fini

Lorsque \mathcal{X} est fini, le théorème 3 ne s'applique pas. Le résultat suivant montre que la consistance universelle uniforme est alors possible, avec une règle très simple, dite « règle de majorité » : pour tout entier $n \geq 1$, tout $x \in \mathcal{X}$ et tout $(x_i, y_i)_{1 \leq i \leq n} \in (\mathcal{X} \times \{0, 1\})^n$,

$$\hat{f}^{\text{maj}}((x_i, y_i)_{1 \leq i \leq n}; x) := \begin{cases} 1 & \text{si } \text{Card}\{i \in \{1, \dots, n\} / y_i = 1 \text{ et } x_i = x\} \\ & > \text{Card}\{i \in \{1, \dots, n\} / y_i = 0 \text{ et } x_i = x\} \\ 0 & \text{sinon.} \end{cases}$$

Autrement dit, en chaque $x \in \mathcal{X}$, \hat{f}^{maj} réalise un vote majoritaire parmi les Y_i tels que $X_i = x$.

Proposition 16 *On se place en classification binaire avec le coût 0–1 et l'on suppose que \mathcal{X} est fini. Pour tout entier $n \geq 1$, on a :*

$$\sup_{P \in \mathcal{M}_1(\mathcal{X} \times \{0, 1\})} \left\{ \mathbb{E}_{D_n \sim P^{\otimes n}} [\ell_P(f_P^*, \hat{f}^{\text{maj}}(D_n))] \right\} \leq \sqrt{\frac{\ln(2) \text{Card}(\mathcal{X})}{2n}}. \quad (53)$$

En particulier, la règle de majorité \hat{f}^{maj} est uniformément universellement consistante.

La démonstration de la proposition 16 est laissée en exercice au lecteur. On peut notamment remarquer que la règle de majorité est la règle par partition associée à $\mathcal{A} = \{\{x\} / x \in \mathcal{X}\}$ (qui est une partition finie puisque \mathcal{X} est fini). Les résultats de la section 3 s'appliquent donc.

La proposition 16 montre que si \mathcal{X} est fini, alors il est possible d'obtenir une vitesse d'apprentissage universelle, de l'ordre de $n^{-1/2}$. Ce résultat doit toutefois être relativisé car la taille de \mathcal{X} intervient dans la vitesse : la borne (53) n'est utile que lorsque $\text{Card}(\mathcal{X}) = o(n)$, ce qui est cohérent avec la parenthèse 97.

47. On parle également de règle de classification ou de classifieur probabiliste, « probabilistic classifier » en anglais.

La vitesse en $\sqrt{\text{Card}(\mathcal{X})/n}$ est optimale (à constante près) en pire cas⁴⁸, d'après la proposition 17 en section 7 : aucune règle d'apprentissage ne peut faire mieux pour toutes les lois P simultanément. En revanche, avec des hypothèses supplémentaires sur P (par exemple, dans le cas zéro-erreur), une vitesse plus rapide est atteignable, et atteinte par la règle de majorité (voir la proposition 18 en section 7).

6.3 À loi fixée

En classification 0–1 avec \mathcal{X} infini, le théorème 3 n'exclut pas l'existence de vitesse d'apprentissage universelle « à constante près ». Autrement dit, on pourrait avoir une règle de classification \hat{f} et une suite $(u_n)_{n \geq 1}$ (indépendante de la loi P) telles que, pour toute loi P sur $\mathcal{X} \times \{0, 1\}$:

$$\begin{aligned} \exists c(P) \in \mathbb{R}, \quad \forall n \geq 1, \quad \mathbb{E}_{D_n \sim P^{\otimes n}} [\ell_P(f_P^*, \hat{f}(D_n))] &\leq c(P)u_n \\ \text{et} \quad u_n &\xrightarrow[n \rightarrow +\infty]{} 0. \end{aligned} \tag{54}$$

Le résultat ci-dessous montre qu'il est impossible d'avoir (54) dès que \mathcal{X} est infini, même avec une suite $(u_n)_{n \geq 1}$ qui tend très lentement vers zéro — par exemple, $u_n = 1/\ln \ln(n)$.

Théorème 4 *Soit \mathcal{X} un ensemble infini, $n \geq 1$ un entier et \hat{f} une règle de classification binaire ($\mathcal{Y} = \{0, 1\}$). Soit $(a_n)_{n \geq 1}$ une suite de réels positifs, décroissante, convergeant vers zéro et telle que $a_1 \leq 1/16$. Alors, avec le coût de classification 0–1, on a :*

$$\exists P \in \mathcal{M}_1(\mathcal{X} \times \{0, 1\}), \quad \forall n \geq 1, \quad \mathbb{E}_{D_n \sim P^{\otimes n}} [\ell_P(f_P^*, \hat{f}(D_n))] \geq a_n.$$

Devroye *et al.* (1996, théorème 7.2) démontrent le théorème 4. Ce résultat est également valable pour une règle de classification randomisée (voir la remarque 98 en section 6.1).

Dans le même ordre d'idée, on peut signaler que sous les hypothèses du théorème 4, quelle que soit \hat{f} , il existe une règle d'apprentissage \hat{g} universellement consistante et une loi P telles que $\hat{g}(D_n)$ a un risque strictement inférieur à celui de $\hat{f}(D_n)$ pour tout n (Devroye *et al.*, 1996, problème 7.3).

7 Conclusion : enjeux de l'apprentissage

Les théorèmes 3 et 4 ne doivent pas être interprétés de manière pessimiste. Ils démontrent certes qu'il est impossible d'apprendre n'importe quoi (les contre-exemples construits dans la démonstration du théorème 3 correspondant bien à « n'importe quoi »), sauf dans le cadre jouet où \mathcal{X} est fini (proposition 16). Or, justement, les données réelles ne correspondent pas à n'importe quoi ! Elles possèdent une structure (plus ou moins connue *a priori*). Par exemple, très souvent, la fonction de régression est

48. On dit que la règle de majorité est minimax, selon la définition donnée en section 7.

« régulière » (c'est-à-dire, α -hölderienne, lipschitzienne ou de classe C^k , éventuellement par morceaux). Tout l'enjeu de l'apprentissage statistique est d'arriver à exploiter au mieux cette structure (connue ou supposée).

C'est d'ailleurs ce que font déjà implicitement les deux grandes familles de règles d'apprentissage étudiées dans ce texte. Une règle par minimisation du risque empirique sur un modèle S repose sur l'*a priori* que le prédicteur de Bayes est bien approché par le modèle S . Une règle par moyenne locale vise à exploiter une certaine régularité de la fonction de régression, comme on le constate à la lecture de la démonstration du théorème de Stone (théorème 2 en section 5.2).

Sur le plan théorique, ceci ouvre un vaste domaine de recherche (encore très partiellement exploré à ce jour).

Tout d'abord, il faut identifier des « structures » intéressantes, c'est-à-dire qui sont réalistes pour certaines applications et pour lesquelles l'apprentissage est possible.

Ensuite, pour chacune de ces structures, il s'agit d'analyser les performances (en termes de risque) des différentes règles d'apprentissage connues, afin d'identifier lesquelles sont les meilleures (et lesquelles sont à éviter absolument!).

Dans le même ordre d'idées, pour une règle d'apprentissage donnée, il s'agit d'identifier les types de données pour lesquelles elle fonctionne, et ceux pour lesquels elle échoue à apprendre. Au vu des théorèmes 3 et 4 en section 6, il y en a forcément, ce que l'on a tendance à oublier quand on vante les mérites de la nouvelle règle d'apprentissage qu'on est en train de proposer.

Puisque l'on n'a rien sans rien, tout ceci ne peut pas se faire avec une seule règle d'apprentissage. Il est donc important de disposer d'un grand nombre de règles, aussi diverses que possibles.

Enfin, que l'on s'intéresse à des données massives ou pas, il est fondamental d'effectuer tout ceci en prenant pleinement en compte le temps de calcul de chaque règle d'apprentissage considérée. On peut ainsi établir dans différents cadres que si le temps de calcul est contraint, la vitesse d'apprentissage optimale dépend fortement des ressources computationnelles disponibles (Chandrasekaran et Jordan, 2013; Zhang *et al.*, 2014; Wang *et al.*, 2016).

7.1 Minimax

L'approche minimax donne une manière de formaliser ceci.

Définition 8 (Risque minimax) Soit \mathcal{P} un ensemble de lois de probabilité sur $\mathcal{X} \times \mathcal{Y}$. Pour tout entier $n \geq 1$, on définit le risque minimax sur \mathcal{P} pour un échantillon de taille n par :

$$\mathcal{R}_{\text{minimax}}(\mathcal{P}, n) := \inf_{\hat{f} \text{ règle d'apprentissage}} \sup_{P \in \mathcal{P}} \mathbb{E}_{D_n \sim P^{\otimes n}} [\mathcal{R}_P(\hat{f}(D_n)) - \mathcal{R}_P^*].$$

Une règle d'apprentissage qui réalise l'infimum pour tout $n \geq 1$ est dite minimax sur \mathcal{P} .

Autrement dit, le risque minimax est la plus petite valeur possible de l'excès de risque moyen, en pire cas sur \mathcal{P} .

Le théorème 3 peut alors se reformuler ainsi : en classification binaire ($\mathcal{Y} = \{0, 1\}$) avec le coût 0–1, pour tout entier $n \geq 1$,

$$\mathcal{R}_{\text{minimax}}(\mathcal{M}_1(\mathcal{X} \times \{0, 1\}), n) = \frac{1}{2}$$

où l'on rappelle que $\mathcal{M}_1(\mathcal{X} \times \{0, 1\})$ est l'ensemble des mesures de probabilité sur $\mathcal{X} \times \{0, 1\}$. De plus, la règle « pile ou face » est minimax sur $\mathcal{M}_1(\mathcal{X} \times \{0, 1\})$.

Supposer que les données ont une certaine structure revient ici à considérer un ensemble \mathcal{P} de lois beaucoup plus petit que $\mathcal{M}_1(\mathcal{X} \times \{0, 1\})$, de telle sorte que le risque minimax sur \mathcal{P} tend vers zéro lorsque n tend vers l'infini. De manière équivalente, on veut qu'il existe une vitesse d'apprentissage uniforme sur \mathcal{P} . Pour chaque ensemble \mathcal{P} , on peut alors déterminer la valeur du risque minimax et identifier une ou plusieurs règles d'apprentissage qui sont minimax sur \mathcal{P} (au moins à une constante multiplicative près).

Par exemple, en classification binaire, à la suite de la section 3, on peut considérer les ensembles de loi suivants, où $S \subset \mathcal{F}$ désigne une classe de Vapnik-Chervonenkis fixée et $h \in [0, 1]$:

$$\begin{aligned}\mathcal{P}(S) &:= \left\{ P \in \mathcal{M}_1(\mathcal{X} \times \{0, 1\}) / \inf_{f \in S} \mathcal{R}_P(f) = \mathcal{R}_P^* \right\} \\ \mathcal{P}(S, h) &:= \left\{ P \in \mathcal{P}(S) / |2\eta_P(X) - 1| \geq h \text{ p.s.} \right\}.\end{aligned}$$

Dire que $P \in \mathcal{P}(S)$ revient à dire que l'erreur d'approximation de S est nulle. Il est donc naturel de penser à une règle de minimisation du risque empirique sur S pour prendre en compte une telle structure. Le résultat suivant valide cette intuition.

Proposition 17 *On se place en classification 0–1. Des constantes numériques $\kappa_1, \kappa_2 > 0$ existent telles que les deux résultats suivants ont lieu, pour toute classe de Vapnik-Chervonenkis S de dimension $V(S)$ et tout entier $n \geq 1$. D'une part, pour tout $P \in \mathcal{P}(S)$ et toute règle \hat{f}_S minimisant le risque empirique sur S :*

$$\mathbb{E}[\ell(f^*, \hat{f}_S(D_n))] \leq \kappa_1 \sqrt{\frac{V(S)}{n}}.$$

D'autre part, si $V(S) \geq 2$:

$$\mathcal{R}_{\text{minimax}}(\mathcal{P}(S), n) \geq \kappa_2 \sqrt{\frac{V(S)}{n}}.$$

Ainsi, une règle minimisant le risque empirique sur S est minimax sur $\mathcal{P}(S)$ à constante près (voir Massart et Nédélec, 2006, section 1.2.1, pour des références). Remarquons que la majoration du risque de \hat{f}_S obtenue en section 3.7 est légèrement

moins bonne⁴⁹ que celle de la proposition 17 : les deux majorations diffèrent d'un facteur $\sqrt{\ln n}$.

Parenthèse 99 (Pires cas parmi $\mathcal{P}(S)$) La démonstration de la borne inférieure sur le risque minimax sur $\mathcal{P}(S)$ permet d'avoir une bonne idée des lois $P \in \mathcal{P}(S)$ les plus « difficiles » pour toute règle de classification \hat{f} ; voir l'exercice 39.

Les familles de lois $\mathcal{P}(S, h)$ ajoutent une condition supplémentaire, dite « condition de marge » (Boucheron *et al.*, 2005, section 5.2) : la fonction de régression doit rester éloignée de $1/2$, à distance au moins $h/2 > 0$. Au vu de la proposition 2 en section 2.2, ceci rend le problème de classification plus facile⁵⁰. En particulier, dès que n est assez grand, on est à l'abri des pires cas de la proposition 17, à cause desquels la vitesse d'apprentissage sur $\mathcal{P}(S)$ est de l'ordre de $\sqrt{V(S)/n}$. Par exemple, avec $h = 1$, on obtient les lois zéro-erreur de $\mathcal{P}(S)$, qui correspondent clairement à un problème de classification plus « simple ». Le résultat suivant montre que l'on obtient alors des vitesses d'apprentissage bien meilleures (comme on l'a vu en section 3.8 pour le cas zéro-erreur).

Proposition 18 *On se place en classification 0–1. Des constantes numériques $\kappa_3, \kappa_4 > 0$ existent telles que les deux résultats suivants sont valables pour toute classe de Vapnik-Chervonenkis S de dimension $V(S)$, tout entier $n \geq 1$ et tout $h \in]0, 1]$. D'une part, pour tout $P \in \mathcal{P}(S, h)$ et toute règle \hat{f}_S minimisant le risque empirique sur S :*

$$\mathbb{E}[\ell(f^*, \hat{f}_S(D_n))] \leq \begin{cases} \kappa_3 \sqrt{\frac{V(S)}{n}} & \text{si } h \leq \sqrt{\frac{V(S)}{n}} \\ \kappa_3 \frac{V(S)}{nh} \left(1 + \ln \frac{nh^2}{V(S)}\right) & \text{sinon.} \end{cases}.$$

D'autre part, si $V(S) \geq 2$:

$$\mathcal{R}_{\text{minimax}}(\mathcal{P}(S, h), n) \geq \kappa_4 \min\left\{\sqrt{\frac{V(S)}{n}}, \frac{V(S)}{nh}\right\}.$$

Massart et Nédélec (2006) démontrent la proposition 18. Pour la borne inférieure, un cas particulier est traité par l'exercice 40.

Parenthèse 100 (Borne supérieure sous condition de marge)

Massart et Nédélec (2006, corollaire 3) démontrent en fait des bornes supérieures sur le risque de \hat{f}_S plus générales que celles de la proposition 18. Sans supposer que $P \in \mathcal{P}(S)$ — mais avec la condition de marge —, on a que pour tout $\epsilon > 0$,

49. Ce facteur $\sqrt{\ln n}$ peut être supprimé dans le cas général (et pas seulement le cas bien spécifié $P \in \mathcal{P}(S)$ traité par la proposition 17), avec une démonstration plus fine que celle de la section 3.7.

50. Du moins, il devient plus facile d'apprendre un classifieur de risque inférieur à un seuil $\epsilon > 0$ fixé. Pour le statisticien, l'analyse théorique fine devient plus complexe...

il existe une constante $\kappa'_3(\epsilon)$ telle que

$$\mathbb{E}[\ell(f^*, \hat{f}_S(D_n))] \leq (1 + \epsilon)\ell(f^*, S) + \kappa'_3(\epsilon) \frac{V(S)}{nh} \left[1 + \ln\left(\frac{nh^2}{V(S)}\right) \right]$$

si $h > \sqrt{V(S)/n}$. De plus, sans supposer que S est une classe de Vapnik-Chervonenkis, on a :

$$\mathbb{E}[\ell(f^*, \hat{f}_S(D_n))] \leq (1 + \epsilon)\ell(f^*, S) + \frac{\kappa'_3(\epsilon)}{nh} \left(1 + \mathbb{E}[H_S(X_1, \dots, X_n)] \right).$$

La proposition 18 met en avant une propriété remarquable des règles par minimisation du risque empirique sur S dans cet exemple : leur *adaptation* au paramètre de marge h . En effet, quel que soit $h \in [0, 1]$, \hat{f}_S est minimax sur $\mathcal{P}(S, h)$ à constante près (plus éventuellement un facteur $\ln(n)$), alors que \hat{f}_S n'utilise pas la valeur du paramètre h : la *même* règle atteint ici le risque minimax (ou presque) sur $\mathcal{P}(S, h)$ pour tout h . Reste encore un « paramètre » à choisir : le modèle S . La section 3.9 aborde cette question (le problème de sélection de modèles).

Plus généralement, obtenir des règles d'apprentissage adaptatives est un enjeu important. Les familles de loi « intéressantes » sont en effet souvent de la forme \mathcal{P}_α , où $\alpha \in \mathbb{R}$ est un paramètre (par exemple, la régularité hölderienne de la fonction de régression). Ainsi, disposer d'une règle unique qui est (quasi) minimax sur \mathcal{P}_α pour tout α est bien plus utile en pratique que d'avoir une règle différente pour chaque α (ce qui nécessite de connaître α). Bien sûr, l'adaptation n'est possible que si la famille $(\mathcal{P}_\alpha)_{\alpha \in \mathbb{R}}$ n'est pas « trop grande » (puisque n'a rien sans rien), et elle a souvent un coût statistique (la perte d'une constante ou d'un facteur logarithmique par rapport au risque minimax). Le lecteur intéressé peut consulter l'article de Barron *et al.* (1999, section 5), le livre de Tsybakov (2004, chapitre 3) et la thèse de Chagny (2013, section 1.1) pour plus de détails et des références sur le problème statistique de l'adaptation, qui est un domaine de recherche à part entière.

7.2 Autres approches

Le point de vue minimax présente aussi des inconvénients. Le fait de considérer le risque moyen « en pire cas » est souvent critiquable, les lois P les plus « difficiles » étant souvent irréalistes (voir la démonstration du théorème 3 en section 6.1 et la parenthèse 99 ci-dessus). S'il est inutile de chercher la meilleure règle \hat{f} pour une loi P fixée (c'est toujours la règle constante égale à f_P^* , qui n'a aucun intérêt pratique), d'autres approches sont possibles.

Par exemple, l'approche *maxiset* vise à identifier, pour une règle d'apprentissage et une vitesse donnée, quel est l'ensemble des lois pour lesquelles cette règle atteint cette vitesse d'apprentissage (Rivoirard, 2009; Autin, 2012).

On peut aussi prendre un point de vue *bayésien*⁵¹ et s'intéresser au risque en

51. Signalons une différence importante avec le point de vue bayésien décrit dans la parenthèse 10 en section 1.2 : le « paramètre à estimer » est ici la loi P du couple (X, Y) (à partir d'un échantillon

moyenne sur une famille \mathcal{P} de lois, une distribution π sur \mathcal{P} étant donnée *a priori*. L'approche PAC-bayésienne permet d'obtenir ce type de borne. Pour en savoir plus, on peut lire l'article de survol de Boucheron *et al.* (2005, section 6), le mémoire d'Audibert (2010, chapitre 2) et le tutoriel de McAllester (2013).

8 Annexe : outils probabilistes

Cette section rassemble les résultats probabilistes utiles dans les sections qui précédent. À quelques détails près, ces résultats (et leurs démonstrations) sont issus du livre de Boucheron *et al.* (2013), où le lecteur intéressé peut trouver des références bibliographiques et bien d'autres résultats probabilistes utiles en apprentissage statistique.

8.1 Sommes de variables indépendantes bornées : inégalité de Hoeffding

On commence par un résultat de concentration pour des sommes de variables aléatoires indépendantes et bornées (l'inégalité de Hoeffding), qui repose sur le contrôle suivant des moments exponentiels d'une variable aléatoire bornée.

Lemme 3 (Lemme de Hoeffding) *Si ξ est une variable aléatoire à valeurs dans $[a, b] \subset \mathbb{R}$, alors, pour tout $\lambda \in \mathbb{R}$,*

$$\ln \mathbb{E}[e^{\lambda \xi}] \leq \lambda \mathbb{E}[\xi] + \frac{\lambda^2(b-a)^2}{8}.$$

Le lemme 3 a été initialement démontré par Hoeffding.

Démonstration Pour tout $\lambda \in \mathbb{R}$, on pose :

$$\psi(\lambda) = \ln \mathbb{E}[e^{\lambda \xi}].$$

Cette quantité est bien définie pour tout $\lambda \in \mathbb{R}$ car ξ est bornée et l'exponentielle est strictement positive. On note que $\psi(0) = 0$. Par ailleurs, ψ est de classe \mathcal{C}^∞ sur \mathbb{R} ($e^{\lambda \xi}$ est une variable positive bornée et l'exponentielle est \mathcal{C}^∞) et pour tout $\lambda \in \mathbb{R}$,

$$\psi'(\lambda) = \frac{\mathbb{E}[\xi e^{\lambda \xi}]}{\mathbb{E}[e^{\lambda \xi}]}.$$

On note que $\psi'(0) = \mathbb{E}[\xi]$. On obtient enfin que pour tout $\lambda \in \mathbb{R}$,

$$\psi''(\lambda) = \frac{\mathbb{E}[\xi^2 e^{\lambda \xi}]}{\mathbb{E}[e^{\lambda \xi}]} - \left(\frac{\mathbb{E}[\xi e^{\lambda \xi}]}{\mathbb{E}[e^{\lambda \xi}]} \right)^2 = \mathbb{E}[Z_\lambda^2] - (\mathbb{E}[Z_\lambda])^2 = \text{var}(Z_\lambda)$$

D_n), alors que dans la parenthèse 10 il s'agit seulement d'« estimer » Y à partir de l'observation de X .

où Z_λ est une variable aléatoire dont la loi admet pour densité

$$x \mapsto \frac{e^{\lambda x}}{\mathbb{E}[e^{\lambda \xi}]}$$

par rapport à la loi P de ξ . En particulier, Z_λ est une variable à valeurs dans $[a, b]$, donc pour tout $\lambda \in \mathbb{R}$,

$$\psi''(\lambda) = \text{var}(Z_\lambda) \leq \mathbb{E}\left[\left(Z - \frac{a+b}{2}\right)^2\right] \leq \frac{(b-a)^2}{4}.$$

On en déduit le résultat en intégrant deux fois par rapport à λ . \square

Remarque 101 (Variable sous-gaussienne) Une variable aléatoire ξ centrée et telle que

$$\forall \lambda \in \mathbb{R}, \quad \ln \mathbb{E}[e^{\lambda \xi}] \leq \frac{v \lambda^2}{2} \quad (55)$$

est dite *sous-gaussienne* de facteur de variance v (Boucheron *et al.*, 2013, section 2.3). Par exemple, une variable gaussienne centrée de variance σ^2 est sous-gaussienne de facteur de variance σ^2 . Le lemme de Hoeffding démontre qu'une variable centrée bornée est sous-gaussienne de facteur de variance $(b-a)^2/4$.

Théorème 5 (Inégalité de Hoeffding) Soit ξ_1, \dots, ξ_n des variables aléatoires indépendantes et bornées : pour tout $i \in \{1, \dots, n\}$, on a des réels a_i et b_i tels que $a_i \leq \xi_i \leq b_i$ presque sûrement. On pose :

$$S_n = \sum_{i=1}^n \xi_i - \mathbb{E}[\xi_i].$$

Alors, on a :

$$\forall \lambda \in \mathbb{R}, \quad \mathbb{E}[e^{\lambda S_n}] \leq \exp\left(\frac{\lambda^2}{8} \sum_{i=1}^n (b_i - a_i)^2\right) \quad (56)$$

$$\forall \epsilon > 0, \quad \mathbb{P}(S_n \geq \epsilon) \leq \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) \quad (57)$$

$$\text{et} \quad \forall \epsilon > 0, \quad \mathbb{P}(|S_n| \geq \epsilon) \leq 2 \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Le théorème 5 a été initialement démontré par Hoeffding.

Démonstration L'inégalité (56) est une conséquence directe du lemme de Hoeffding et de l'indépendance des ξ_i .

Pour en déduire l'inégalité (57), on applique la méthode de Cramér-Chernoff :

pour tout $\epsilon > 0$ et tout $\lambda > 0$,

$$\begin{aligned}\mathbb{P}(S_n \geq \epsilon) &= \mathbb{P}\left(e^{\lambda S_n} \geq e^{\lambda \epsilon}\right) \\ &\leq e^{-\lambda \epsilon} \mathbb{E}[e^{\lambda S_n}] \\ &\leq \exp\left(\frac{\lambda^2}{8} \sum_{i=1}^n (b_i - a_i)^2 - \lambda \epsilon\right)\end{aligned}$$

donc

$$\mathbb{P}(S_n \geq \epsilon) \leq \exp\left(\inf_{\lambda > 0} \left\{ \frac{\lambda^2}{8} \sum_{i=1}^n (b_i - a_i)^2 - \lambda \epsilon \right\}\right).$$

On obtient le résultat en choisissant

$$\lambda = \frac{4\epsilon}{\sum_{i=1}^n (b_i - a_i)^2} > 0.$$

La troisième inégalité se déduit de (57) (en l'appliquant aux ξ_i puis aux $-\xi_i$, avant d'utiliser une borne d'union). \square

L'inégalité (56) signifie que S_n est sous-gaussienne de facteur de variance $\sum_{i=1}^n (b_i - a_i)^2 / 4$. La méthode de Cramér-Chernoff (Boucheron *et al.*, 2013, section 2.2 et théorème 2.8) — qui s'applique bien plus généralement — permet d'en déduire l'inégalité de concentration⁵² (57).

Remarque 102 (Reformulation statistique) L'inégalité (57) est présentée sous un angle probabiliste : on fixe d'abord un seuil $\epsilon > 0$ et l'on cherche ensuite à majorer la probabilité que S_n dépasse ce seuil. On peut la reformuler en un résultat plus directement utile pour un statisticien : pour tout $\delta \in]0, 1[$, avec probabilité au moins $1 - \delta$, on a :

$$S_n < \mathbb{E}[S_n] + \sqrt{\frac{1}{2} \sum_{i=1}^n (a_i - b_i)^2 \ln\left(\frac{1}{\delta}\right)}.$$

Supposons que les variables ξ_i sont de même loi. Alors, pour un niveau de risque δ donné, on a une borne supérieure sur $S_n - \mathbb{E}[S_n]$ de l'ordre de $\sqrt{n \ln(1/\delta)}$. Ce résultat est similaire au théorème limite central : les déviations typiques de S_n autour de son espérance sont (au plus) de l'ordre de \sqrt{n} . Il y a cependant deux différences fondamentales entre ces deux résultats. D'une part, l'inégalité de Hoeffding est *non-asymptotique* : elle est valable pour tout entier n fixé, ce qui est très utile. D'autre part, la constante $n(b_1 - a_1)^2 / 2$ apparaissant dans le terme de déviation n'est qu'un *majorant* de la variance de S_n : le théorème limite central est donc plus précis sur ce point. Si l'on veut un résultat non-asymptotique faisant apparaître la variance de S_n , on peut utiliser l'inégalité de Bernstein (Boucheron *et al.*, 2013, section 2.8).

52. On parle ici de concentration de S_n car (57) majore la probabilité que S_n dévie d'au moins ϵ de son *espérance* (qui est nulle), quel que soit $\epsilon > 0$. Une inégalité traitant des déviations de S_n au-delà d'un seuil t éloigné de son espérance serait appelée *inégalité de déviation*.

8.2 Maximum de variables sous-gaussiennes

Il est souvent nécessaire de majorer l'espérance du maximum de K variables aléatoires. La proposition suivante permet de le faire pour des variables aléatoires sous-gaussiennes (par exemple, la variable S_n du théorème 5).

Proposition 19 *Soit Z_1, \dots, Z_K des variables aléatoires sous-gaussiennes de facteur de variance v , c'est-à-dire, telles que :*

$$\forall k \in \{1, \dots, K\}, \quad \mathbb{E}[Z_k] = 0 \quad \text{et} \quad \forall \lambda \in \mathbb{R}, \quad \ln \mathbb{E}[e^{\lambda Z_k}] \leq \frac{v\lambda^2}{2}.$$

Alors,

$$\mathbb{E}\left[\max_{1 \leq k \leq K} Z_k\right] \leq \sqrt{2v \ln(K)}.$$

Un point remarquable est que la proposition 19 s'applique sans *aucune hypothèse sur la dépendance* entre les Z_k . Un résultat similaire s'en déduit pour la partie positive et la valeur absolue des Z_k (exercice 43).

Par ailleurs, la borne obtenue dans la proposition 19 est précise dans un cas particulier au moins : lorsque Z_1, \dots, Z_K sont des variables normales centrées réduites indépendantes (voir l'exercice 44).

Démonstration Un argument élémentaire « à la Pisier » fournit une démonstration courte et élégante, et peut même être utilisé pour obtenir un résultat beaucoup plus général (Boucheron *et al.*, 2013, section 2.5). Posons

$$M = \max_{1 \leq k \leq K} Z_k.$$

Pour tout $\lambda > 0$, par l'inégalité de Jensen,

$$e^{\lambda \mathbb{E}[M]} \leq \mathbb{E}[e^{\lambda M}] = \mathbb{E}\left[\max_{1 \leq k \leq K} e^{\lambda Z_k}\right] \leq \sum_{k=1}^K \mathbb{E}[e^{\lambda Z_k}] \leq K e^{\lambda^2 v / 2}.$$

On a donc

$$\mathbb{E}[M] \leq \inf_{\lambda > 0} \left\{ \frac{\ln(K)}{\lambda} + \frac{\lambda v}{2} \right\},$$

d'où le résultat en prenant $\lambda = \sqrt{2 \ln(K) / v}$. \square

Dans le cas particulier des moyennes de Rademacher, on obtient le résultat suivant.

Lemme 4 *Soit $\mathcal{B} \subset \mathbb{R}^n$ un ensemble fini (déterministe) et $(\varepsilon_i)_{i=1, \dots, n}$ des variables de Rademacher indépendantes. Alors,*

$$\mathbb{E}\left[\max_{\beta \in \mathcal{B}} \sum_{i=1}^n \beta_i \varepsilon_i\right] \leq \sqrt{2 \ln(\text{Card}(\mathcal{B})) \max_{\beta \in \mathcal{B}} \sum_{i=1}^n \beta_i^2}.$$

Démonstration Le lemme de Hoeffding (lemme 3) montre que pour tout $\beta \in \mathcal{B}$, $\sum_{i=1}^n \beta_i \varepsilon_i$ est sous-gaussienne de facteur de variance $\sum_{i=1}^n \beta_i^2$. Le résultat s'en déduit avec la proposition 19. \square

Autrement dit, avec la notation introduite en section 3.7, le lemme 4 démontre :

$$\mathbb{E}[\text{Rad}_n(\mathcal{B})] \leq \sqrt{2 \ln(\text{Card}(\mathcal{B}))} \times \max_{\beta \in \mathcal{B}} \|\beta\|_2.$$

8.3 Inégalité de Mc Diarmid

L'inégalité suivante, dûe à McDiarmid, démontre qu'une fonction de n variables aléatoires indépendantes se concentre autour de son espérance, si cette fonction varie peu lorsqu'une seule de ses entrées varie⁵³.

Théorème 6 (Inégalité de Mc Diarmid) *On considère Ξ un espace mesurable, ξ_1, \dots, ξ_n des variables aléatoires indépendantes à valeurs dans Ξ et $F : \Xi^n \rightarrow \mathbb{R}$ une fonction mesurable. Notons $Z = F(\xi_1, \dots, \xi_n)$. Si pour tous $x, x' \in \Xi^n$ et $i \in \{1, \dots, n\}$,*

$$|F(x_1, \dots, x_{i-1}, x'_i, x_{i+1}, \dots, x_n) - F(x)| \leq c_i, \quad (58)$$

alors, pour tout $\epsilon \geq 0$,

$$\mathbb{P}(Z - \mathbb{E}[Z] \geq \epsilon) \leq \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n c_i^2}\right). \quad (59)$$

Boucheron *et al.* (2013, théorème 6.2) démontrent le théorème 6. L'inégalité de McDiarmid peut être vue comme une généralisation de l'inégalité de Hoeffding : si $F(\xi_1, \dots, \xi_n) = \sum_{i=1}^n \xi_i$ avec $\xi_i \in [a_i, b_i]$ presque sûrement, la condition (58) est vérifiée avec $c_i = b_i - a_i$ et l'inégalité (59) est alors équivalente à (57).

8.4 Inégalité de symétrisation

L'inégalité de symétrisation suivante est utile pour majorer le risque d'un minimiseur du risque empirique, comme on le fait en section 3.7. Plus généralement, elle motive l'introduction des moyennes de Rademacher (18).

Proposition 20 *Soit Z_1, \dots, Z_n des vecteurs aléatoires indépendants de la forme $Z_i = (Z_{i,t})_{t \in \mathcal{T}}$. On suppose que pour tout $i \in \{1, \dots, n\}$ et $t \in \mathcal{T}$, $\mathbb{E}[Z_{i,t}] = 0$. Soit $\varepsilon_1, \dots, \varepsilon_n$ des variables de Rademacher indépendantes, indépendantes de Z_1, \dots, Z_n . Alors,*

$$\begin{aligned} \mathbb{E}\left[\sup_{t \in \mathcal{T}} \left\{\sum_{i=1}^n Z_{i,t}\right\}\right] &\leq 2\mathbb{E}\left[\sup_{t \in \mathcal{T}} \left\{\sum_{i=1}^n \varepsilon_i Z_{i,t}\right\}\right] \\ \text{et} \quad \frac{1}{2}\mathbb{E}\left[\sup_{t \in \mathcal{T}} \left|\sum_{i=1}^n \varepsilon_i Z_{i,t}\right|\right] &\leq \mathbb{E}\left[\sup_{t \in \mathcal{T}} \left|\sum_{i=1}^n Z_{i,t}\right|\right] \leq 2\mathbb{E}\left[\sup_{t \in \mathcal{T}} \left|\sum_{i=1}^n \varepsilon_i Z_{i,t}\right|\right]. \end{aligned}$$

53. En anglais, l'inégalité de McDiarmid est souvent appelée « bounded-difference inequality ».

Démonstration Soit Z'_1, \dots, Z'_n une copie indépendante de Z_1, \dots, Z_n . Par conséquent, les vecteurs $Z_i - Z'_i$ sont indépendants et symétriques, de même loi que les vecteurs $\varepsilon_i(Z_i - Z'_i)$. On a alors :

$$\begin{aligned} \mathbb{E}\left[\sup_{t \in \mathcal{T}}\left\{\sum_{i=1}^n Z_{i,t}\right\}\right] &= \mathbb{E}\left[\sup_{t \in \mathcal{T}}\left\{\sum_{i=1}^n (Z_{i,t} - \mathbb{E}[Z'_{i,t}])\right\}\right] \\ &\leq \mathbb{E}\left[\sup_{t \in \mathcal{T}}\left\{\sum_{i=1}^n (Z_{i,t} - Z'_{i,t})\right\}\right] \\ &\quad (\text{par l'inégalité de Jensen, puisque } Z \mapsto \sup_{t \in \mathcal{T}} Z_t \text{ est convexe}) \\ &= \mathbb{E}\left[\sup_{t \in \mathcal{T}}\left\{\sum_{i=1}^n \varepsilon_i(Z_{i,t} - Z'_{i,t})\right\}\right] \\ &\leq \mathbb{E}\left[\sup_{t \in \mathcal{T}}\left\{\sum_{i=1}^n \varepsilon_i Z_{i,t}\right\}\right] + \mathbb{E}\left[\sup_{t \in \mathcal{T}}\left\{\sum_{i=1}^n -\varepsilon_i Z'_{i,t}\right\}\right] \\ &= 2\mathbb{E}\left[\sup_{t \in \mathcal{T}}\left\{\sum_{i=1}^n \varepsilon_i Z_{i,t}\right\}\right], \end{aligned}$$

ce qui prouve la première inégalité. La troisième inégalité s'en déduit en posant :

$$\tilde{Z}_i = ((Z_{i,t})_{t \in \mathcal{T}}, (-Z_{i,t})_{t \in \mathcal{T}}).$$

On démontre de même la deuxième inégalité :

$$\begin{aligned} \mathbb{E}\left[\sup_{t \in \mathcal{T}}\left|\sum_{i=1}^n \varepsilon_i Z_{i,t}\right|\right] &= \mathbb{E}\left[\sup_{t \in \mathcal{T}}\left|\sum_{i=1}^n \varepsilon_i (Z_{i,t} - \mathbb{E}[Z'_{i,t}])\right|\right] \\ &\leq \mathbb{E}\left[\sup_{t \in \mathcal{T}}\left|\sum_{i=1}^n \varepsilon_i (Z_{i,t} - Z'_{i,t})\right|\right] \\ &\quad (\text{par l'inégalité de Jensen, puisque } Z \mapsto \sup_{t \in \mathcal{T}} |Z_t| \text{ est convexe}) \\ &= \mathbb{E}\left[\sup_{t \in \mathcal{T}}\left|\sum_{i=1}^n (Z_{i,t} - Z'_{i,t})\right|\right] \\ &\leq 2\mathbb{E}\left[\sup_{t \in \mathcal{T}}\left|\sum_{i=1}^n Z_{i,t}\right|\right]. \end{aligned}$$

□

Parenthèse 103 (Proposition 20 avec \mathcal{T} infini) La démonstration de la proposition 20 reste valable lorsque \mathcal{T} est infini, dès lors que tous les suprema sur $t \in \mathcal{T}$ écrits dans l'énoncé restent mesurables (ce qui est vrai notamment lorsque \mathcal{T} est dénombrable, ou bien sous des conditions de séparabilité).

Parenthèse 104 (Minoration sans valeurs absolues) La minoration de la deuxième partie de la proposition 20 n'est pas valable en général si l'on retire les valeurs absolues dans le supremum. En revanche, on peut démontrer que :

$$\mathbb{E} \left[\sup_{t \in \mathcal{T}} \sum_{i=1}^n \varepsilon_i Z_{i,t} \right] \leq \mathbb{E} \left[\sup_{t \in \mathcal{T}} \sum_{i=1}^n Z_{i,t} \right] + \mathbb{E} \left[\sup_{t \in \mathcal{T}} \sum_{i=1}^n (-Z_{i,t}) \right].$$

8.5 Espérance de l'inverse d'une variable binomiale

On termine cette section par un lemme, issu du livre de Györfi *et al.* (2002, lemme 4.1), qui est utile pour l'analyse des règles par partition (notamment la démonstration du corollaire 1 en section 5.3).

Lemme 5 Soit Z une variable aléatoire binomiale de paramètres $n \geq 1$ et $p \in]0, 1]$. Alors, en posant par convention $\frac{0}{0} = 0$, on a :

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

Démonstration On a :

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

où la dernière égalité découle de la formule du binôme. \square

9 Annexe : exercices

Nous proposons enfin une série d'exercices qui complètent les résultats présentés précédemment, en suivant l'ordre de présentation des sections précédentes.

9.1 Régression et classification

Exercice 1 Déterminer l'ensemble des prédicteurs de Bayes, la valeur du risque de Bayes et l'excès de risque en régression avec le coût valeur absolue.

Exercice 2 On se place en régression avec le coût valeur absolue, et l'on suppose que la loi de Y sachant X est (presque sûrement) symétrique autour de son espérance $\eta(X)$. Préciser alors l'ensemble des prédicteurs de Bayes et la formule de l'excès de risque.

Exercice 3 Démontrer la proposition 3.

Exercice 4 Démontrer un analogue de la proposition 4 pour un coût asymétrique c_w quelconque, avec la règle par plug-in correspondante $\hat{f}_{\hat{\eta},w}$ (voir la parenthèse 40).

Exercice 5 On suppose que P est une loi « zéro-erreur » en classification binaire, c'est-à-dire que $\eta(X) \in \{0, 1\}$ presque sûrement. Montrer que, pour le coût 0–1 en classification, on a :

$$\ell(f^*, \hat{f}_{\hat{\eta}}(D_n)) \leq 4\mathbb{E}\left[\left(\hat{\eta}(D_n; X) - \eta(X)\right)^2 \mid D_n\right].$$

Plus généralement, lorsque P vérifie la condition de marge

$$|2\eta_P(X) - 1| \geq h \quad \text{p.s.},$$

montrer que l'on a, pour le coût 0–1 en classification :

$$\ell(f^*, \hat{f}_{\hat{\eta}}(D_n)) \leq \frac{4}{h}\mathbb{E}\left[\left(\hat{\eta}(D_n; X) - \eta(X)\right)^2 \mid D_n\right].$$

9.2 Minimisation du risque empirique

9.2.1 Erreur d'approximation et erreur d'estimation

Exercice 6 (Partition en régression : erreur d'approximation) Soit \mathcal{A} une partition mesurable de \mathcal{X} , finie ou dénombrable, et $S_{\text{reg}}^{\text{part}}(\mathcal{A})$ le modèle associé. Démontrer que, pour le risque quadratique en régression, l'erreur d'approximation de $S_{\text{reg}}^{\text{part}}(\mathcal{A})$ est égale à l'excès de risque du prédicteur

$$\eta_{\mathcal{A}} : x \in \mathcal{X} \mapsto \eta_{\mathcal{A}(x)} \quad \text{où} \quad \forall A \in \mathcal{A}, \quad \eta_A := \mathbb{E}[\eta(X) \mid X \in A]$$

et l'on rappelle que $\mathcal{A}(x)$ désigne l'unique élément de \mathcal{A} qui contient x . En déduire :

$$\ell(f^*, S_{\text{reg}}^{\text{part}}(\mathcal{A})) = \sum_{A \in \mathcal{A}} \mathbb{P}(X \in A) \text{var}(\eta(X) | X \in A).$$

On suppose que η est L -lipschitzienne relativement à une distance d sur \mathcal{X} , c'est-à-dire :

$$\forall x, x' \in \mathcal{X}, \quad |\eta(x) - \eta(x')| \leq Ld(x, x').$$

Démontrer que :

$$\ell(f^*, S_{\text{reg}}^{\text{part}}(\mathcal{A})) \leq \frac{L^2}{4} \mathbb{E}[\text{diam}_d(\mathcal{A}(X))^2] \leq \frac{L^2}{4} \sup_{A \in \mathcal{A}} (\text{diam}_d(A))^2,$$

où pour tout $A \subset \mathcal{X}$

$$\text{diam}_d(A) := \sup_{x, x' \in A} d(x, x')$$

désigne le diamètre de A . Pour finir, on suppose que X est de loi uniforme sur $\mathcal{X} = [0, 1]$, que η est de classe C^1 et l'on considère $\mathcal{A}^{\text{cub}}(h)$ la partition cubique de pas $h > 0$. Démontrer que :

$$\ell(f^*, S_{\text{reg}}^{\text{part}}(\mathcal{A}^{\text{cub}}(h))) \underset{h \rightarrow 0}{\sim} \frac{h^2}{12} \int_0^1 (\eta'(x))^2 dx.$$

Exercice 7 (Partition en classification : erreur d'approximation) Soit \mathcal{A} une partition mesurable de \mathcal{X} , finie ou dénombrable, et $S_{\text{class}}^{\text{part}}(\mathcal{A})$ le modèle associé. Démontrer que, pour le risque 0–1 en classification, l'erreur d'approximation de $S_{\text{class}}^{\text{part}}(\mathcal{A})$ est égale à l'excès de risque du classifieur

$$f_{\mathcal{A}, \text{class}}^*: x \in \mathcal{X} \mapsto \mathbf{1}_{\mathbb{P}(Y=1 | X \in A) > \frac{1}{2}}$$

et qu'elle vérifie :

$$\begin{aligned} & \ell(f^*, S_{\text{class}}^{\text{part}}(\mathcal{A})) \\ &= \sum_{A \in \mathcal{A}} \mathbb{P}(X \in A) \min \left\{ \mathbb{E}\left[(2\eta(X) - 1)_+ | X \in A\right], \mathbb{E}\left[(2\eta(X) - 1)_- | X \in A\right] \right\} \\ &\leq 2\sqrt{\ell(f^*, S_{\text{reg}}^{\text{part}}(\mathcal{A}))}. \end{aligned}$$

En déduire que si η est L -lipschitzienne relativement à une distance d sur \mathcal{X} , alors :

$$\ell(f^*, S_{\text{class}}^{\text{part}}(\mathcal{A})) \leq L \sup_{A \in \mathcal{A}} \text{diam}_d(A).$$

Exercice 8 (Partition en régression : erreur d'estimation) Soit \mathcal{A} une partition mesurable de \mathcal{X} , finie ou dénombrable, et $\hat{f}_{\mathcal{A}}^{\text{p-r}}$ la règle de régression par partition associée. Montrer que :

$$\begin{aligned} & \mathbb{E}\left[\left(\hat{f}_{\mathcal{A}}^{\text{p-r}}(D_n; X) - f^*(X)\right)^2 \mid D_n\right] \\ &= \mathbb{E}\left[\left(\eta_{\mathcal{A}}(X) - f^*(X)\right)^2\right] + \mathbb{E}\left[\left(\hat{f}_{\mathcal{A}}^{\text{p-r}}(D_n; X) - \eta_{\mathcal{A}}(X)\right)^2 \mid D_n\right], \end{aligned} \quad (60)$$

où $\eta_{\mathcal{A}}$ est définie à l'exercice 6. L'équation (60) correspond à la décomposition (11) de l'excès de risque de $\hat{f}_{\mathcal{A}}^{\text{p-r}}$ (qui est un minimiseur du risque empirique) en erreur d'approximation et erreur d'estimation.

Pour tout $A \in \mathcal{A}$, on rappelle la notation :

$$N_A(X_{1\dots n}) = \text{Card}\{i \in \{1, \dots, n\} / X_i \in A\}.$$

On définit également la variance résiduelle sachant X :

$$\sigma^2(X) := \mathbb{E}\left[\left(Y - \eta(X)\right)^2 \mid X\right].$$

Montrer que l'espérance de l'erreur d'estimation de $\hat{f}_{\mathcal{A}}^{\text{p-r}}$ s'écrit :

$$\begin{aligned} & \mathbb{E}\left[\left(\hat{f}_{\mathcal{A}}^{\text{p-r}}(D_n; X) - \eta_{\mathcal{A}}(X)\right)^2\right] \\ &= \sum_{A \in \mathcal{A}} \mathbb{P}(X \in A) (1 - \mathbb{P}(X \in A))^n \eta_A^2 \\ & \quad + \sum_{A \in \mathcal{A}} \mathbb{P}(X \in A) \mathbb{E}\left[\frac{\mathbf{1}_{N_A(X_{1\dots n})>0}}{N_A(X_{1\dots n})}\right] \text{var}(Y \mid X \in A), \end{aligned} \quad (61)$$

avec $\text{var}(Y \mid X \in A) = \left(\mathbb{E}[\sigma^2(X) \mid X \in A] + \text{var}(\eta(X) \mid X \in A)\right).$

Majorer l'espérance de l'erreur d'estimation de $\hat{f}_{\mathcal{A}}^{\text{p-r}}$ en démontrant que, pour tout $A \in \mathcal{A}$:

$$\begin{aligned} \mathbb{P}(X \in A) (1 - \mathbb{P}(X \in A))^n &\leq \min\left\{\frac{1}{ne}, \mathbb{P}(X \in A)\right\} \\ \mathbb{P}(X \in A) \mathbb{E}\left[\frac{\mathbf{1}_{N_A(X_{1\dots n})>0}}{N_A(X_{1\dots n})}\right] &\leq \min\left\{\frac{2}{n+1}, \mathbb{P}(X \in A)\right\}. \end{aligned}$$

On suppose désormais que les fonctions η et σ sont bornées (ce qui est toujours le cas si Y est bornée).

Lorsque \mathcal{A} est finie, montrer que :

$$\mathbb{E}\left[\left(\hat{f}_{\mathcal{A}}^{\text{p-r}}(D_n; X) - \eta_{\mathcal{A}}(X)\right)^2\right] \leq 3(\|\eta\|_{\infty}^2 + \|\sigma\|_{\infty}^2) \frac{\text{Card}(\mathcal{A})}{n}.$$

Lorsque $\mathcal{X} = \mathbb{R}^p$ et que $\mathcal{A} = \mathcal{A}_n$ (finie ou dénombrable) vérifie l'hypothèse (c') du corollaire 1 en section 5.3, montrer que l'espérance de l'erreur d'estimation de $\hat{f}_{\mathcal{A}_n}^{\text{p-r}}$ tend vers zéro quand n tend vers l'infini.

Remarque 105 (Commentaires sur l'exercice 8) Considérons le cas de la régression homoscédastique (c'est-à-dire que la variance résiduelle $\sigma^2(X) = \sigma^2$ ne dépend pas de X) et supposons qu'il existe $\mathcal{A}'_n \subset \mathcal{A} = \mathcal{A}_n$ tel que

$$\mathbb{P}\left(X \in \bigcup_{A \in \mathcal{A}'_n} A\right) = 1 - o\left(\frac{1}{n}\right) \quad \text{et} \quad \inf_{A \in \mathcal{A}'_n} \mathbb{P}(X \in A) \gg \frac{1}{n}.$$

Alors, le premier terme de la décomposition (61) est négligeable par rapport à $1/n$. De plus, pour le second terme, Arlot (2008, lemme 3) démontre que pour toute partie mesurable A de \mathcal{X} telle que $\mathbb{P}(X \in A) > 0$,

$$\mathbb{P}(X \in A) \mathbb{E}\left[\frac{\mathbf{1}_{N_A(X_1 \dots n) > 0}}{N_A(X_1 \dots n)}\right] \underset{n \rightarrow +\infty}{\sim} \frac{1}{n}.$$

Enfin, si η est uniformément continue (relativement à une distance d sur \mathcal{X}) et si $\sup_{A \in \mathcal{A}_n} \text{diam}_d(A)$ tend vers 0 quand n tend vers $+\infty$, on a :

$$\sup_{A \in \mathcal{A}_n} \text{var}(\eta(X) | X \in A) \xrightarrow[n \rightarrow +\infty]{} 0.$$

Par conséquent, l'espérance de l'erreur d'estimation de \hat{f}_A^{p-r} est de l'ordre de

$$\frac{\sigma^2 \text{Card}(\mathcal{A}'_n)}{n}.$$

Cette formule donne l'ordre de grandeur typique de l'erreur d'estimation de \hat{f}_S en régression homoscédastique : niveau de bruit (σ^2), multiplié par le nombre de paramètres du modèle (la dimension de S comme espace vectoriel, ici, $\text{Card}(\mathcal{A}'_n)$), divisés par le nombre d'observations (n). Un tel résultat est exact (et facile à démontrer) en régression sur un plan d'expérience déterministe (Arlot et Bach, 2009, équation (6)).

9.2.2 Majoration générale de l'erreur d'estimation

Exercice 9 Démontrer la proposition 8 en section 3.5. Plus généralement, montrer que pour tout modèle $S \subset \mathcal{F}$, $\rho \geq 0$ et $\hat{f}_{S,\rho}$ une règle par ρ -minimisation du risque empirique sur S , on a :

$$\mathbb{E}[\ell(f^*, \hat{f}_{S,\rho}) - \ell(f^*, S)] \leq \rho + \mathbb{E}\left[\sup_{f \in S} \{\mathcal{R}_P(f) - \hat{\mathcal{R}}_n(f)\}\right].$$

9.2.3 Cas d'un modèle fini

Exercice 10 Montrer que sous les conditions de la proposition 9 en section 3.6, pour tout $x, \rho \geq 0$, si $\hat{f}_{S,\rho}$ est une règle par ρ -minimisation du risque empirique sur S , on a :

$$\mathbb{P}\left(\ell(f^*, \hat{f}_{S,\rho}(D_n)) < \ell(f^*, S) + \rho + (b-a)\sqrt{\frac{2[x + \ln(\text{Card } S)]}{n}}\right) \geq 1 - e^{-x}.$$

Exercice 11 Soit $S \subset \mathcal{F}$ un modèle fini, $\rho \geq 0$ et $\hat{f}_{S,\rho}$ une règle par ρ -minimisation du risque empirique sur S . On suppose que le risque \mathcal{R}_P et le risque empirique $\widehat{\mathcal{R}}_n$ sont définis tous les deux avec le même coût c . Alors, si

$$\forall f \in S, \quad \mathbb{E}|c(f(X), Y)| < +\infty,$$

on a presque sûrement :

$$\limsup_{n \rightarrow +\infty} \ell(f^*, \hat{f}_{S,\rho}(D_n)) \leq \ell(f^*, S) + \rho.$$

Si de plus une constante v existe telle que

$$\forall f \in S, \quad \text{var}\left(c(f(X), Y)\right) \leq v,$$

alors, pour tout $\delta \in (0, 1]$:

$$\mathbb{P}\left(\ell(f^*, \hat{f}_{S,\rho}(D_n)) < \ell(f^*, S) + \rho + \sqrt{\frac{v \text{Card } S}{n\delta}}\right) \geq 1 - \delta.$$

Exercice 12 Démontrer la proposition 10 en section 3.6. Sous les mêmes hypothèses, montrer que plus généralement, pour tout $\rho \geq 0$ et toute règle $\hat{f}_{S,\rho}$ par ρ -minimisation du risque empirique sur S , on a :

$$\mathbb{E}\left[\ell(f^*, \hat{f}_{S,\rho}(D_n))\right] \leq \ell(f^*, S) + \rho + (b - a)\sqrt{\frac{\ln(\text{Card } S)}{2n}}.$$

9.2.4 Cas d'un modèle quelconque

Exercice 13 On se place en classification avec le coût 0–1. Soit $S \subset \mathcal{F}$ un modèle « symétrique », c'est-à-dire tel que pour tout $f \in S$, on a également $1 - f \in S$. En utilisant la proposition 20 en section 8.4, montrer que l'inégalité de symétrisation (17) est fine (à un facteur numérique près) :

$$\frac{1}{2}\mathbb{E}\left[\sup_{f \in S}\left\{\frac{1}{n} \sum_{i=1}^n \varepsilon_i c(f(X_i), Y_i)\right\}\right] \leq \mathbb{E}\left[\sup_{f \in S}\{\mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f)\}\right].$$

Exercice 14 Démontrer les affirmations faites après la définition 5 :

1. Un modèle S fini est toujours une classe de Vapnik-Chervonenkis, de dimension $V(S) \leq \ln_2(\text{Card } S)$.
2. Si $\mathcal{X} = \mathbb{R}^2$, le modèle

$$S = \{\mathbb{1}_A / A \subset \mathbb{R}^2 \text{ convexe}\}$$

n'est pas une classe de Vapnik-Chervonenkis.

Exercice 15 Pour toute collection \mathcal{A} de parties mesurables de \mathcal{X} , on définit le modèle :

$$S_{\mathcal{A}} = \{\mathbf{1}_A / A \in \mathcal{A}\}.$$

Démontrer que $S_{\mathcal{A}}$ est une classe de Vapnik-Chervonenkis, déterminer sa dimension $V(S_{\mathcal{A}})$ et encadrer $\mathcal{C}(S_{\mathcal{A}}, k)$ pour tout $k \geq 1$ dans les cas suivants.

1. $\mathcal{X} = \mathbb{R}$ et \mathcal{A} est l'ensemble des demi-droites de la forme $]-\infty, a]$ avec $a \in \mathbb{R}$.
2. $\mathcal{X} = \mathbb{R}$ et \mathcal{A} est l'ensemble des demi-droites de \mathbb{R} .
3. $\mathcal{X} = \mathbb{R}$ et \mathcal{A} est l'ensemble des intervalles de \mathbb{R} .
4. $\mathcal{X} = \mathbb{R}^p$ et $\mathcal{A} = \{]-\infty, a_1] \times \dots \times]-\infty, a_p] / a_1, \dots, a_p \in \mathbb{R}\}$.
5. $\mathcal{X} = \mathbb{R}^p$ et \mathcal{A} est l'ensemble des pavés de \mathbb{R}^p .
6. $\mathcal{X} = \mathbb{R}^p$ et \mathcal{A} est l'ensemble des demi-espaces de \mathbb{R}^p .

Pour ce dernier cas, on peut commencer par résoudre l'exercice 16 ci-dessous.

Exercice 16 Soit \mathcal{G} un espace vectoriel de fonctions mesurables $\mathcal{X} \rightarrow \mathbb{R}$. Démontrer que si \mathcal{G} est de dimension finie r , alors

$$S = \{\mathbf{x} \mapsto \mathbf{1}_{g(\mathbf{x}) \geq 0} / g \in \mathcal{G}\}$$

est une classe de Vapnik-Chervonenkis de dimension $V(S) \leq r$.

Exercice 17 On se place en classification binaire avec le coût 0–1, $\mathcal{X} = \mathbb{R}^p$ et l'on considère \hat{f} une règle de classification qui minimise le risque empirique sur le modèle

$$S = \{\mathbf{1}_A / A \text{ demi-espace de } \mathbb{R}^p\}.$$

Montrer qu'on a toujours :

$$\mathbb{E}[\ell(f^*, \hat{f}_S(D_n))] \leq \ell(f^*, S) + 2\sqrt{2} \sqrt{\frac{p+1}{n} \ln\left(\frac{en}{p+1}\right)}.$$

Si de plus on suppose que le support de P_X est inclus dans un sous-espace vectoriel de \mathbb{R}^d de dimension $d < p$, montrer qu'on a alors :

$$\mathbb{E}[\ell(f^*, \hat{f}_S(D_n))] \leq \ell(f^*, S) + 2\sqrt{2} \sqrt{\frac{d+1}{n} \ln\left(\frac{en}{d+1}\right)}.$$

Exercice 18 (Partition en classification : entropie combinatoire) Soit \mathcal{A} une partition mesurable de \mathcal{X} , finie ou dénombrable, et $S_{\text{class}}^{\text{part}}(\mathcal{A})$ le modèle par partition associé. On rappelle que son entropie combinatoire empirique $H_{S_{\text{class}}^{\text{part}}(\mathcal{A})}$ est définie par l'équation (22) en section 3.7. Soit $X, X_1, \dots, X_n, \dots$ une suite de variables indépendantes et de même loi P_X sur \mathcal{X} .

Démontrer que pour tout entier $n \geq 1$, on a :

$$H_{S_{\text{class}}^{\text{part}}(\mathcal{A})}(X_1, \dots, X_n) = \ln(2) \sum_{A \in \mathcal{A}} \mathbb{1}_{\{\exists i \in \{1, \dots, n\} / X_i \in A\}}.$$

En déduire que

$$\begin{aligned} \frac{1}{n} \mathbb{E}[H_{S_{\text{class}}^{\text{part}}(\mathcal{A})}(X_1, \dots, X_n)] &= \frac{\ln(2)}{n} \sum_{A \in \mathcal{A}} \left[1 - (1 - \mathbb{P}(X \in A))^n \right] \\ &\xrightarrow[n \rightarrow +\infty]{} 0, \end{aligned} \quad (62)$$

puis que l'espérance de l'erreur d'estimation d'une règle de classification par partition sur $S_{\text{class}}^{\text{part}}(\mathcal{A})$ tend vers zéro quand n tend vers l'infini, quelle que soit la loi P des observations. Si $\mathcal{X} = \mathbb{R}^p$, montrer que c'est encore le cas lorsque la partition $\mathcal{A} = \mathcal{A}_n$ varie avec n et vérifie l'hypothèse (c') du corollaire 1 en section 5.3.

Exercice 19 En classification avec le coût 0–1, démontrer que l'application $D_n \in (\mathcal{X} \times \mathcal{Y})^n \mapsto \text{Rad}_n(\mathcal{B}_S(D_n))$ vérifie les conditions de l'inégalité de Mc Diarmid. En déduire une inégalité de concentration pour $\text{Rad}_n(\mathcal{B}_S(D_n))$ autour de son espérance.

Exercice 20 Montrer que

$$\begin{aligned} \mathbb{E} \left[\sup_{f \in S} |\mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f)| \right] &\leq 2 \mathbb{E} \left[\sup_{f \in S} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i c(f(X_i), Y_i) \right| \right] \\ &= 2 \text{Rad}_n(\mathcal{B}_S(D_n) \cup -\mathcal{B}_S(D_n)). \end{aligned}$$

On se place désormais en classification 0–1. Expliquer pourquoi remplacer $\mathcal{B}_S(D_n)$ par $\mathcal{B}_S(D_n) \cup -\mathcal{B}_S(D_n)$ revient à remplacer le modèle S par :

$$\overline{S} := S \cup \{1 - f / f \in S\}.$$

Majorer l'entropie combinatoire empirique de \overline{S} en fonction de celle de S , ainsi que la dimension de Vapnik-Chervonenkis de \overline{S} en fonction de celle de S .

Exercice 21 On suppose le coût c borné : $c(f(X), Y) \in [a, b]$ presque sûrement pour tout $f \in S$. En utilisant l'inégalité de Mc Diarmid (théorème 6 en section 8.3), démontrer que pour tout $x \geq 0$,

$$\mathbb{P} \left(\sup_{f \in S} \{ \mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f) \} < \mathbb{E} \left[\sup_{f \in S} \{ \mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f) \} \right] + (b-a) \sqrt{\frac{x}{2n}} \right) \geq 1 - e^{-x}.$$

En déduire une majoration de $\ell(f^*, \widehat{f}_S)$ valable avec grande probabilité : d'abord dans le cas général, puis en classification 0–1 lorsque S est une classe de Vapnik-Chervonenkis (en utilisant l'exercice 20).

Exercice 22 Étendre l'analyse de Vapnik-Chervonenkis détaillée en section 3.7 au cas d'un coût asymétrique c_w quelconque.

9.2.5 Choix d'un modèle

Exercice 23 Démontrer une version plus générale du lemme 2 : soit \mathcal{E} un ensemble et $\mathcal{C}, \mathcal{R}, A, B$ des applications $\mathcal{E} \rightarrow \mathbb{R}$ telles que,

$$\forall x, x' \in \mathcal{E}, \quad [\mathcal{C}(x) - \mathcal{R}(x)] - [\mathcal{C}(x') - \mathcal{R}(x')] \leq B(x) + A(x'). \quad (63)$$

Alors, pour tout $\rho > 0$ et tout

$$\hat{x} \in \mathcal{E} \quad \text{tel que} \quad \mathcal{C}(\hat{x}) \leq \inf_{x \in \mathcal{E}} \mathcal{C}(x) + \rho,$$

on a

$$\mathcal{R}(\hat{x}) - A(\hat{x}) \leq \inf_{x \in \mathcal{E}} \{\mathcal{R}(x) + B(x)\} + \rho.$$

On peut remarquer que la condition (63) implique la condition (27).

9.3 Coûts convexes en classification

Exercice 24 Soit $S \subset \overline{\mathcal{F}}$ et $\Phi : \mathbb{R} \rightarrow \mathbb{R}$ une fonction L_Φ -lipschitzienne, bornée sur S (c'est-à-dire, $|\Phi(Yf(X))| \leq B$ p.s.) et telle que $\Phi(0) = 0$. Soit \hat{f}_S un minimiseur du Φ -risque empirique sur S :

$$\hat{f}_S \in \operatorname{argmin}_{f \in S} \left\{ \hat{\mathcal{R}}_n^\Phi(f) \right\} \quad \text{où} \quad \hat{\mathcal{R}}_n^\Phi(f) := \frac{1}{n} \sum_{i=1}^n \Phi(Y_i f_\theta(X_i)).$$

En généralisant l'approche de Vapnik-Chervonenkis détaillée en section 3.7, majorer l'espérance de l'erreur d'estimation de \hat{f}_S à l'aide de la moyenne de Rademacher

$$\operatorname{Rad}_n \left(\left\{ (f(X_i))_{1 \leq i \leq n} / f \in \mathcal{F} \right\} \right).$$

Pour cela, on peut utiliser le principe de contraction (Boucheron *et al.*, 2005, théorème 3.3) : si $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ est une fonction L -lipschitzienne telle que $\varphi(0) = 0$ et si $A \subset \mathbb{R}^n$ est un ensemble borné, alors

$$\operatorname{Rad}_n \left(\left\{ (\varphi(a_1), \dots, \varphi(a_n)) / (a_1, \dots, a_n) \in A \right\} \right) \leq L \operatorname{Rad}_n(A).$$

L'exercice 24 suppose la fonction Φ bornée, ce qui n'est le cas d'aucune des fonctions Φ convexes classiques. L'exercice 25 ci-dessous montre qu'on peut s'y ramener pour le coût charnière (qui est 1-lipschitzien) et le coût quadratique tronqué.

Exercice 25 Soit $\Phi : \mathbb{R} \rightarrow \mathbb{R}$ la fonction définie par $\Phi(u) = (1-u)_+$ (comme pour définir le coût charnière). Soit $g \in \overline{\mathcal{F}}$ et

$$\bar{g} : x \in \mathcal{X} \mapsto \min \left\{ 1, \max \{-1, g(x)\} \right\}$$

le pseudo-classifieur « tronqué » associé. Montrer que

$$\mathcal{R}_P^\Phi(\bar{g}) \leq \mathcal{R}_P^\Phi(g) \quad \text{et} \quad \widehat{\mathcal{R}}_n^\Phi(\bar{g}) \leq \widehat{\mathcal{R}}_n^\Phi(g).$$

Autrement dit, on ne perd rien à se limiter à un modèle $S \subset \mathcal{F}$ contenant uniquement des fonctions bornées par 1 (en valeur absolue).

Montrer qu'il en est de même avec le risque quadratique tronqué, c'est-à-dire, lorsque $\Phi(u) = (1 - u)_+^2$.

Exercice 26 On définit la loi de $(X, Y) \in \mathbb{R}^p \times \{-1, 1\}$ comme suit (dans l'esprit de la remarque 39 en section 2.2). D'une part, $\mathbb{P}(Y = 1) = q \in [0, 1]$. D'autre part, sachant Y , X suit une loi gaussienne, de moyenne $\boldsymbol{\mu}_Y \in \mathbb{R}^p$ et de matrice de covariance Σ inversible et indépendante de Y . Démontrer qu'on a alors :

$$\begin{aligned} \log \frac{\mathbb{P}(Y = 1 | X = \mathbf{x})}{\mathbb{P}(Y = -1 | X = \mathbf{x})} &= a + \mathbf{b}^\top \mathbf{x} \\ \text{avec} \quad &\begin{cases} a = \log \frac{q}{1-q} + \frac{1}{2} \boldsymbol{\mu}_{-1}^\top \Sigma^{-1} \boldsymbol{\mu}_{-1} - \frac{1}{2} \boldsymbol{\mu}_1^\top \Sigma^{-1} \boldsymbol{\mu}_1 \\ \mathbf{b} = \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_{-1}) \end{cases} \end{aligned} \quad (64)$$

En déduire que :

$$\mathbb{P}(Y = 1 | X = \mathbf{x}) = \tau(a + \mathbf{b}^\top \mathbf{x}) \quad \text{avec} \quad \tau(u) = \frac{1}{1 + e^{-u}}.$$

Généraliser ce résultat au cas où $\mathcal{L}(X | Y = 1)$ et $\mathcal{L}(X | Y = -1)$ appartiennent à une même famille exponentielle de statistique suffisante X , c'est-à-dire, si ces deux lois ont pour densités respectives

$$f_1(\mathbf{x}) = c(x) \exp(\boldsymbol{\theta}_1^\top \mathbf{x} - A(\boldsymbol{\theta}_1)) \quad \text{et} \quad f_{-1}(\mathbf{x}) = c(x) \exp(\boldsymbol{\theta}_{-1}^\top \mathbf{x} - A(\boldsymbol{\theta}_{-1}))$$

avec $\boldsymbol{\theta}_1, \boldsymbol{\theta}_{-1} \in \mathbb{R}^p$, $c : \mathcal{X} \rightarrow [0, +\infty[$ et $A : \mathbb{R}^p \rightarrow \mathbb{R}$.

Exercice 27 Pour le coût quadratique, démontrer l'équivalence entre l'inégalité fournie par le théorème 1 (section 4.5) et la proposition 4 (section 2.2), en faisant bien attention aux passages entre les conventions $\mathcal{Y} = \{0, 1\}$ et $\mathcal{Y} = \{-1, 1\}$.

Exercice 28 Justifier l'ensemble des formules indiquées dans la table 1.

Exercice 29 Soit Φ convexe et calibrée pour la classification. Soit P une loi sur $\mathcal{X} \times \{-1, 1\}$ vérifiant la « condition de marge » suivante pour un réel $h > 0$:

$$\mathbb{P}(|2\zeta(X) - 1| \geq h) = 1.$$

Montrer que pour tout pseudo-classifieur $g \in \overline{\mathcal{F}}$, on a alors :

$$\frac{\Psi(h)}{h} \left[\mathcal{R}_P^{0-1}(\text{signe}(g)) - \mathcal{R}_P^{0-1*} \right] \leq \mathcal{R}_P^\Phi(g) - \mathcal{R}_P^{\Phi*}.$$

Dans le cas du risque quadratique, comparer ceci au résultat de l'exercice 5.

9.4 Moyenne locale

Exercice 30 Démontrer les affirmations faites dans l'exemple 7 en section 5.1.

Exercice 31 On se place en régression avec le coût quadratique.

- (a) Démontrer les équations (44) et (45).
- (b) On suppose de plus que la variance résiduelle ne dépend pas de X (cas homoscédastique). Démontrer l'équation (47).

Exercice 32 (Partition cubique et consistance universelle) Soit $(h_n)_{n \geq 0}$ une suite de réels strictement positifs. Montrer que la règle de classification par partition cubique associée $\hat{f}_{(h_n)}^{\text{cub}-c}$ est universellement consistante si et seulement si $h_n \rightarrow 0$ et $nh_n^p \rightarrow +\infty$ quand n tend vers l'infini.

Exercice 33 (Partition en classification : consistance) Soit $(\mathcal{A}_n)_{n \geq 1}$ une suite de partitions finies ou dénombrables de $\mathcal{X} = \mathbb{R}^p$ et P une loi sur $\mathcal{X} \times \{0, 1\}$. On suppose les conditions suivantes vérifiées :

- (i) $\text{diam}(\mathcal{A}_n(X)) \xrightarrow[n \rightarrow +\infty]{(p)} 0$
- (ii) $N_{\mathcal{A}_n(X)} \xrightarrow[n \rightarrow +\infty]{(p)} +\infty$.

Montrer que la règle par partition $\hat{f}_{(\mathcal{A}_n)}^{p-c}$ associée est faiblement consistante pour P en classification pour le risque 0–1.

Montrer également que ceci généralise le corollaire 1 : si (b') a lieu, alors la condition (i) ci-dessus est vérifiée. Si (c') a lieu, alors la condition (ii) ci-dessus est vérifiée.

Exercice 34 Démontrer les affirmations de la parenthèse 88.

9.5 On n'a rien sans rien

Exercice 35 On suppose que $\mathcal{X} = \mathbb{R}$. Montrer que le théorème 3 est encore valable quand on restreint le supremum aux lois P telles que :

- (a) X a une loi à densité sur \mathbb{R} .
- (b) X a une loi à densité sur \mathbb{R} et la fonction de régression η est de classe \mathcal{C}^∞ sur \mathbb{R} .

Exercice 36 Soit \mathcal{X} un ensemble infini, $n \geq 1$ un entier, $c \in [0, 1/2]$ et \hat{f} une règle de classification binaire ($\mathcal{Y} = \{0, 1\}$). Avec le coût de classification 0–1, déterminer la valeur de :

$$\sup_{P \text{ loi sur } \mathcal{X} \times \{0,1\} / \mathcal{R}_P^* = c} \left\{ \mathbb{E}_{D_n \sim P^{\otimes n}} [\ell_P(f_P^*, \hat{f}(D_n))] \right\}. \quad (65)$$

Exercice 37 On se place en régression sur $\mathcal{X} = \mathbb{R}$ avec le coût quadratique. Montrer que pour tout entier $n \geq 1$:

$$\inf_{\hat{f}} \left\{ \sup_{P \text{ loi sur } \mathbb{R} \times [0,1]} \left\{ \mathbb{E}_{D_n \sim P^{\otimes n}} [\ell_P(f_P^*, \hat{f}(D_n))] \right\} \right\} \in \left[\frac{1}{16}, \frac{1}{4} \right] \quad (66)$$

où l'infimum est pris sur l'ensemble des règles de régression. Déterminer la valeur exacte de l'infimum.

Exercice 38 Démontrer la proposition 16. Pour cela, on peut notamment remarquer que \hat{f}^{maj} est une règle par partition (donc une règle minimisant le risque empirique), ce qui permet d'utiliser les résultats de la section 3.

9.6 Conclusion

Exercice 39 Démontrer la proposition 17 en section 7.1. Pour établir la borne inférieure, on peut démontrer le résultat plus général suivant. Pour tout $L \in [0, 1/2]$, on note :

$$\mathcal{P}_L(S) := \{P \in \mathcal{P}(S) / \mathcal{R}_P^* = L\}.$$

On a alors :

$$\forall L \in \left]0, \frac{1}{2}\right[, \quad \forall n \geqslant \frac{V(S) - 1}{2L} \max\left(9, \frac{1}{(1 - 2L)^2}\right),$$

$$\mathcal{R}_{\minimax}(\mathcal{P}_L(S), n) \geqslant e^{-8} \sqrt{\frac{L(V(S) - 1)}{24n}}.$$

Pour ce faire, on peut s'inspirer de la démonstration du théorème 3 en section 6.1 en considérant pour tout $r \in \{0, 1\}^{V(S)-1}$ la loi P_r définie comme suit. On choisit $x_1, \dots, x_{V(S)}$ qui réalisent le supremum de l'entropie combinatoire empirique H_S . On fixe des paramètres $p \in [0, 1/(V(S) - 1)]$ et $c \in [0, 1/2]$, et l'on pose :

$$P_X(x_i) = \begin{cases} p & \text{pour } i \in \{1, \dots, V(S) - 1\}, \\ 1 - (V(S) - 1)p & \text{pour } i = V(S), \end{cases}$$

$$\eta(x_i) = \begin{cases} \frac{1}{2} + c(2r_i - 1) & \text{pour } i \in \{1, \dots, V(S) - 1\}, \\ 0 & \text{pour } i = V(S). \end{cases}$$

Un choix judicieux de p et c permet d'obtenir la minoration annoncée.

Exercice 40 Démontrer la proposition 18 en section 7.1, dans le cas particulier $h = 1$ (classification zéro-erreur). Plus précisément, pour tout entier $n \geqslant V(S) - 1$, établir que :

$$\mathcal{R}_{\minimax}(\mathcal{P}(S, 1), n) \geqslant \frac{V(S) - 1}{2en} \left(1 - \frac{1}{n}\right).$$

Pour ce faire, on peut s'inspirer de la démonstration du théorème 3 en section 6.1, en considérant pour tout $r \in \{0, 1\}^{V(S)-1}$ la loi P_r définie comme suit. On choisit $x_1, \dots, x_{V(S)}$ qui réalisent le supremum de l'entropie combinatoire empirique H_S . On

pose alors :

$$P_X(x_i) = \begin{cases} \frac{1}{n} & \text{pour } i \in \{1, \dots, V(S) - 1\}, \\ 1 - \frac{V(S)-1}{n} & \text{pour } i = V(S), \end{cases}$$

$$\eta(x_i) = \begin{cases} r_i & \text{pour } i \in \{1, \dots, V(S) - 1\}, \\ 0 & \text{pour } i = V(S). \end{cases}$$

Exercice 41 Justifier les affirmations de la parenthèse 45 en section 2.3 : borne supérieure sur le risque 0–1 d'un minimiseur du risque empirique sur $S_{\text{class}}^{\text{lin}}$, borne inférieure minimax correspondante, et borne inférieure minimax en régression.

9.7 Outils probabilistes

Exercice 42 Sous les hypothèses de la proposition 19, majorer l'espérance de $\max_{1 \leq k \leq K} Z_k$, en commençant par majorer $\mathbb{P}(Z_k \geq \epsilon)$ pour tous k et ϵ , puis en utilisant une borne d'union. Comparer au résultat de la proposition 19.

Exercice 43 Sous les hypothèses de la proposition 19, démontrer les deux inégalités suivantes :

$$\mathbb{E}\left[\max_{1 \leq k \leq K} (Z_k)_+\right] \leq \sqrt{2v \ln(K+1)}$$

$$\mathbb{E}\left[\max_{1 \leq k \leq K} |Z_k|\right] \leq \sqrt{2v \ln(2K)}.$$

Exercice 44 Si Z_1, \dots, Z_K sont des variables normales centrées réduites indépendantes, on a :

$$\mathbb{E}\left[\max_{1 \leq k \leq K} Z_k\right] \underset{K \rightarrow +\infty}{\sim} \sqrt{2 \ln(K)}.$$

Exercice 45 Démontrer les affirmations de la parenthèse 104 en section 8.4.

Remerciements

Cet texte fait suite à un cours donné dans le cadre des Journées d'Études en Statistique 2016. Il s'agit d'une version préliminaire du chapitre 2 du livre *Apprentissage statistique et données massives*, édité par Frédéric Bertrand, Myriam Maumy-Bertrand, Gilbert Saporta et Christine Thomas-Agnan, à paraître aux éditions Technip.

Ce texte doit beaucoup à plusieurs cours que j'ai donnés avant celui des JES 2016, par ordre chronologique : à l'École Centrale Paris (avec Gilles Stoltz), à l'Université Paris-Sud (avec Francis Bach) et à l'École Normale Supérieure (avec Jean-Yves Audibert, Francis Bach, Olivier Catoni, Guillaume Obozinski et Gilles Stoltz). Je tiens à remercier tous les collègues avec qui j'ai travaillé sur ces cours, ainsi que les étudiants dont les questions m'ont aidé à enrichir leur contenu. Je remercie également les participants des JES 2016 pour leurs questions et commentaires. Enfin, ces notes doivent

beaucoup aux commentaires et critiques de Francis Bach, Gérard Biau, Matthieu Lerasle et Guillaume Obozinski, qui ont généreusement accepté de relire certaines parties de ce texte (voire sont intégralité, merci Matthieu!), ainsi qu'à de nombreux collègues (notamment au laboratoire de mathématiques d'Orsay) qui m'ont apporté leur aide sur des points plus précis ; je les en remercie beaucoup.

Références

- Sylvain ARLOT : *Resampling and Model Selection*. Thèse de doctorat, Université Paris-Sud, décembre 2007. Disponible à l'adresse <http://tel.archives-ouvertes.fr/tel-00198803/>.
- Sylvain ARLOT : *V-fold cross-validation improved : V-fold penalization*, février 2008. arXiv :0802.0566v2.
- Sylvain ARLOT : Validation croisée, 2017. Preprint hal.
- Sylvain ARLOT et Francis BACH : Data-driven calibration of linear estimators with minimal penalties. In Y. BENGIO, D. SCHUURMANS, J. LAFFERTY, C. K. I. WILLIAMS et A. CULOTTA, éditeurs : *Advances in Neural Information Processing Systems 22*, pages 46–54, 2009.
- Sylvain ARLOT et Alain CELISSE : A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79, 2010.
- Jean-Yves AUDIBERT : *PAC-Bayesian aggregation and multi-armed bandits*. Habilitation à diriger des recherches, Université Paris-Est, octobre 2010. Disponible à l'adresse <https://tel.archives-ouvertes.fr/tel-00536084>.
- Jean-Yves AUDIBERT et Alexandre TSYBAKOV : Fast learning rates for plug-in classifiers. *Annals of Statistics*, 35(2):608–633, 2007.
- Florent AUTIN : *Contributions aux problèmes d'estimation et de test non paramétriques*. Habilitation à diriger des recherches, Université Aix-Marseille, décembre 2012. Disponible à l'adresse <https://www.cmi.univ-mrs.fr/~autin/HDR-AUTIN.pdf>.
- Andrew BARRON, Lucien BIRGÉ et Pascal MASSART : Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113(3):301–413, 1999.
- Peter L. BARTLETT, Michael I. JORDAN et Jon D. McAULIFFE : Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.
- Frédéric BERTRAND, Jean-Jacques DROESBEKE, Gilbert SAPORTA et Christine THOMAS-AGNAN, éditeurs. *Choix et agrégation de modèles*. Technip, 2016.

- Gérard BIAU et Luc DEVROYE : *Lectures on the Nearest Neighbor Method*. Springer Series in the Data Sciences. Springer, 2015.
- Peter J. BICKEL et Kjell A. DOKSUM : *Mathematical statistics—basic ideas and selected topics*. Vol. 1. Prentice Hall, deuxième édition, 2001.
- Patrick BILLINGSLEY : *Convergence of probability measures*. Wiley Series in Probability and Statistics : Probability and Statistics. John Wiley & Sons, Inc., New York, deuxième édition, 1999. A Wiley-Interscience Publication.
- Lucien BIRGÉ et Pascal MASSART : Minimal penalties for Gaussian model selection. *Probability Theory and Related Fields*, 138(1-2):33–73, 2007.
- Stéphane BOUCHERON, Olivier BOUSQUET et Gábor LUGOSI : Theory of classification : a survey of some recent advances. *ESAIM. Probability and Statistics*, 9:323–375 (electronic), 2005.
- Stéphane BOUCHERON, Gábor LUGOSI et Pascal MASSART : *Concentration Inequalities : A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford, 2013.
- Gaëlle CHAGNY : *Estimation adaptative avec des données transformées ou incomplètes. Application à des modèles de survie*. Thèse de doctorat, Université Paris Descartes, juillet 2013. Disponible à l'adresse <https://tel.archives-ouvertes.fr/tel-00863141>.
- Venkat CHANDRASEKARAN et Michael I. JORDAN : Computational and statistical tradeoffs via convex relaxation. *Proceedings of the National Academy of Sciences*, 110(13):E1181–E1190, 2013.
- Pierre-André CORNILLON et Éric MATZNER-LØBER : *Régression avec R*. Pratique R. Springer, 2011.
- Luc P. DEVROYE, László GYÖRFI et Gábor LUGOSI : *A Probabilistic Theory of Pattern Recognition*, volume 31 de *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1996.
- David L. DONOHO et Iain M. JOHNSTONE : Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.
- Jean-Jacques DROESBEKE, Gilbert SAPORTA et Christine THOMAS-AGNAN : *Méthodes robustes en statistique*. Éditions Technip, Paris, 2015.
- Bradley EFRON : The estimation of prediction error : covariance penalties and cross-validation. *Journal of the American Statistical Association*, 99(467):619–642, 2004. With comments and a rejoinder by the author.
- Liliana FORZANI, Ricardo FRAIMAN et Pamela LLOP : Consistent nonparametric regression for functional data under the Stone-Besicovitch conditions. *IEEE Transactions on Information Theory*, 58(11):6697–6708, Nov 2012.

- Johannes FÜRNKRANZ : Round robin classification. *Journal of Machine Learning Research (JMLR)*, 2(4):721–747, 2002. Special issue on machine learning approaches to shallow parsing.
- Christophe GIRAUD : *Introduction to High-Dimensional Statistics*, volume 139 de *Monographs on Statistics and Applied Probability*. Chapman and Hall/CRC, Boca Raton, FL, 2014.
- Tilmann GNEITING, Fadoua BALABDAOUI et Adrian E. RAFTERY : Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 69(2):243–268, 2007.
- Tilmann GNEITING et Adrian E. RAFTERY : Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.
- Gregory GRIFFIN, Alex HOLUB et Pietro PERONA : Caltech-256 object category dataset. Rapport technique, Caltech Technical Report, avril 2007. Disponible à l'adresse <http://resolver.caltech.edu/CaltechAUTHORS:CNS-TR-2007-001>.
- Yann GUERMEUR : *SVM Multiclasses, Théorie et Applications*. Habilitation à diriger des recherches, Université Henri Poincaré - Nancy I, novembre 2007. Disponible à l'adresse <https://tel.archives-ouvertes.fr/tel-00203086>.
- László GYÖRFI, Michael KOHLER, Adam KRZYŻAK et Harro WALK : *A Distribution-free Theory of Nonparametric Regression*. Springer Series in Statistics. Springer-Verlag, New York, 2002.
- Michael KEARNS, Yishay MANSOUR, Andrew Y. NG et Dana RON : An Experimental and Theoretical Comparison of Model Selection Methods. *Machine Learning Journal*, 7:7–50, 1997.
- Jean-François LE GALL : Intégration, probabilités et processus aléatoires, 2006. Cours donné à l'école normale supérieure, disponible à l'adresse <https://www.math.u-psud.fr/~jflegall/IPPA2.pdf>.
- Yann LECUN, Léon BOTTOU, Yoshua Bengio et Patrick HAFFNER : Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324, 1998.
- Pascal MASSART : *Concentration Inequalities and Model Selection*, volume 1896 de *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.
- Pascal MASSART et Élodie NÉDÉLEC : Risk bounds for statistical learning. *The Annals of Statistics*, 34(5):2326–2366, 2006.

David A. McALLESTER : A PAC-Bayesian tutorial with a dropout bound, juillet 2013.
arXiv :1307.2118.

Sebastian NOWOZIN, Peter V. GEHLER, Jeremy JANCARY et Christoph H. LAMPERT, éditeurs. *Advanced Structured Prediction*. From Neural Information Processing. MIT Press, 2014.

Vincent RIVOIRARD : *Contributions à l'estimation non-paramétrique. Du rôle de la parcimonie en statistique à la calibration théorique et pratique d'estimateurs*. Habilitation à diriger des recherches, University Paris-Sud, décembre 2009. Disponible à l'adresse https://www.ceremade.dauphine.fr/~rivoirar/HDR_Rivoirard.pdf.

Vincent RIVOIRARD et Gilles STOLTZ : *Statistique en action*. Vuibert, 2009.

Matthieu SOLNON : *Apprentissage statistique multi-tâches*. Thèse de doctorat, Université Pierre et Marie Curie - Paris VI, novembre 2013. Disponible à l'adresse <http://hal.inria.fr/tel-00911498>.

Ambuj TEWARI et Peter L. BARTLETT : On the consistency of multiclass classification methods. *Journal of Machine Learning Research (JMLR)*, 8:1007–1025, 2007.

Alexandre B. TSYBAKOV : *Introduction à l'estimation non-paramétrique*, volume 41 de *Mathématiques & Applications [Mathematics & Applications]*. Springer-Verlag, Berlin, 2004.

Aad W. VAN DER VAART et Jon A. WELLNER : *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

Vladimir VAPNIK : *The nature of statistical learning theory*. Springer, 2000.

Emmanuel VIENNET : Réseaux à fonctions de base radiales. In Younès BENNANI, éditeur : *Apprentissage connexionniste*, I2C Hermès, pages 105–122. Lavoisier, 2006. Disponible à l'adresse <https://hal.archives-ouvertes.fr/hal-00085092/>.

Tengyao WANG, Quentin BERTHET et Richard J. SAMWORTH : Statistical and computational trade-offs in estimation of sparse principal components. *The Annals of Statistics*, 44(5):1896–1930, 2016.

Min-Ling ZHANG et Zhi-Hua ZHOU : A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, août 2014.

Yuchen ZHANG, Martin J. WAINWRIGHT et Michael I. JORDAN : Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In *JMLR W& CP (COLT 2014)*, volume 35, pages 921–948, 2014.

Index

- adaboost, 61
- adaptation, 95–96
- AIC, 54, 56
- analyse discriminante
 - linéaire, 62, 63
- apprentissage supervisé, 2
- arbre de décision, 13, 18, 31
- BIC, 55
- boosting, 61
- calibration, 66
- calibration pour la classification 0–1, 65–67
- classification non supervisée, 16
- classification supervisée, 16–26, 30–32, 40–46, 57–70, 72–96, 104, 108–115
 - calibrée, 66
 - condition de marge, *voir* marge
 - multiclasse, 3, 16–18, 23, 67
 - multiétiquette, 3, 4, 16
 - (par) partition, *voir* partition
 - règle de majorité, 91
- séparation linéaire, *voir* discrimination linéaire
- zéro-erreur, 20, 47–48, 70, 95, 104, 114
- classifieur, 16
- classifieur bayésien, 90
- classifieur probabiliste, *voir* règle d'apprentissage randomisée
- combinaison de prédicteurs, *voir* agrégation
- compromis biais-variance, 52, 77–78
- condition de marge, *voir* marge
- consistance, 10–11, 44, 69
 - en sélection, *voir* sélection de modèles
 - faible, 10, 38, 72–77
 - forte, 10
- fonction universelle, 11, 71, 77, 79, 83, 86–87, 92, 113
- fonction universelle uniforme, 87–91, 114
- coût (fonction de), 5
 - pour des exemples, *voir aussi* risque 0–1, 19–21
 - asymétrique, 21–23, 104, 110
 - convexe, 57–70, 111–112
 - Huber, 15
 - intervalle de prévision, 15
 - L^p , 15
 - quadratique, 13–15, 24
 - quadratique seuillé, 15
 - Tukey-biweight, 15
 - valeur absolue, 14, 15, 104
- C_p , 54, 56
- décision (théorie de la), 7, 10
- degrés de liberté, 56
- discrimination, *voir* classification supervisée
- linéaire, 25, 31–32, 109
- données aberrantes, 15
- entropie combinatoire empirique, 41, 45, 46, 48, 56, 95, 109, 110, 114
- erreur d'approximation, 32–33, 51–53, 62, 70, 74, 77–78, 94, 104–105
- erreur d'estimation, 33–48, 51–53, 74, 77–78, 84, 105–111
- erreur de généralisation, *voir* risque
- erreur de prévision, *voir* risque
- estimateur, 9
- estimation de densité, 85–86
 - par noyau, *voir* Parzen-Rosenblatt
- excès de risque, 7, 10
- fléau de la dimension, 80
- fonction de régression, 11, 17, 58
- histogramme, *voir* partition
- Hoeffding (inégalité de), 96

k plus proches voisins, 71, 82–84
 LDA, *voir* analyse discriminante linéaire
 marge, 59
 condition de, 70, 94–95, 104, 112
 maxiset, 96
 McDiarmid (inégalité de), 100, 110
 M-estimateur, 27
 minimax, 91, 93–96
 minoration, 20, 88, 113–115
 risque, 93
 minimisation du risque empirique, 26–58, 62, 67, 70, 92, 94–95
 minimisation approchée, 27, 35, 38, 46, 107–108
 minimisation structurelle du risque, 55, 56
 modèle, 27
 moindres carrés, 27
 régression, 13
 régression linéaire, 29
 moyenne locale, 71–87, 92, 113
 multi-tâches, 11, 17
 Nadaraya-Watson (estimateur de), *voir*
 noyau (règle d'apprentissage)
 noyau (règle d'apprentissage), 84–87
 PAC-bayésienne (approche), 96
 partition, *voir aussi* arbre de décision
 règle de classification, 18, 23, 30, 38–39, 46, 72, 79–82, 91, 102, 109, 113, 114
 erreur d'approximation, 79, 105
 erreur d'estimation, 80, 109
 partition cubique, 46, 79–80, 113
 règle de régression, 12–13, 28, 30, 72, 78–82, 102
 erreur d'approximation, 79, 104–106
 erreur d'estimation, 80, 105–107
 partition cubique, 13, 52–53, 79–80, 105
 Parzen-Rosenblatt (estimation de la densité), 85
 pénalisation, *voir* sélection de modèles plug-in, 18, 23–26, 57, 58, 63, 71, 86, 104, 112
 plus proches voisins, *voir* *k* plus proches voisins
 prédicteur, 5
 (de) Bayes, 7, 8, 14, 16, 19, 21, 61, 63–67, 104
 prévision, 2, 6
 prévision structurée, 17
 processus empirique, 36
 pseudo-classifieur, 58, 90
 Rademacher
 complexité de Rademacher, 40, 45, 56, 110
 moyenne de Rademacher, 39, 40, 100, 111
 variable de Rademacher, 39
 rééchantillonnage, 40
 règle d'apprentissage, 9–10
 pile ou face, 88, 93
 randomisée, 88, 90, 92
 règle de classification, 16
 régression, 2, 3, 11–15, 28–30, 77–80, 82–87, 104, 113
 affine, 29
 homoscédastique, 78, 83, 113
 linéaire, 2, 28–30, 50
 (par) moindres carrés, *voir* moindres carrés
 multi-tâches, 11
 multivariée, 11
 (par) partition, *voir* partition
 polynomiale, 30
 régression logistique, 61–63, 70
 régressogramme, *voir* partition, règle de régression
 réseau à fonction de base radiale, 30
 réseau de neurones, 23, 30, 31, 43
 risque, 5, 9
 0–1, 19, 59, 67–70
 (de) Bayes, 7, 8, 14, 19–21, 47, 63–65, 90, 104
 bayésien, 8

- charnière, 61, 67, 69–70, 111
- conditionnel, 6
- empirique, 26
- excès de risque, 7, 10
- exponentiel, 61, 64, 67, 69, 70
- logistique, 61, 64, 67, 69–70
- minimax, *voir* minimax
- (des) moindres carrés, *voir* risque quadratique
- moyen, 9
- Φ -risque, 46, 58, 67–70
- Φ -risque conditionnel, 63
- quadratique, 13, 61, 64, 67, 69, 112
- quadratique tronqué, 61, 67, 69, 112

- Sauer (lemme de), 42
- score de classification, 58
- sélection d'estimateurs, 84, 87
 - principe d'estimation sans biais du risque, 54, 56
- sélection de modèles, 48–57
 - consistance en sélection, 49
 - inégalité oracle, 49, 55
 - optimalité asymptotique, 49, 55
 - pénalisation, 45, 56–57
 - pénalité covariance, 56
 - pénalité idéale, 56
- sélection de variables, 29, 32, 48, 50, 55
- sous-apprentissage, 51–53, 71, 77, 79, 83, 87
- Stone (théorème de), 71–80, 82, 83, 86, 113
- surapprentissage, 42, 50–53, 57, 71, 77, 79, 83, 87
- SVM, 17, 31, 32, 59, 61, 70
- symétrisation, 39, 101, 115

- validation croisée, 54, 84, 87
- Vapnik-Chervonenkis
 - classe de Vapnik-Chervonenkis, 41–45, 48, 94–95, 108–110
 - dimension de Vapnik-Chervonenkis, 41–43, 56, 108–110
- variable
 - explicative, 4