STA211 : Eléments fondamentaux pour méthodes supervisées

Vincent Audigier, Ndèye Niang

03 avril, 2025

1 Introduction

Le data-mining vise à identifier des structures au sein de données. On distingue généralement deux types de structures : *les modèles* et *les patterns*. Contrairement aux patterns, les modèles visent à expliquer et/ou prédire une variable réponse, notée ici Y, à partir d'autres variables dites explicatives, notée X. Les méthodes de data-mining *supervisées* sont les approches privilégiées pour la recherche de tels modèles.

Dans le cas où Y est continue, on parlera de problème de *régression*, tandis que l'on parlera de problème de *classification* supervisée (ou de discrimination) si Y est qualitative. Cette distinction est importante car certaines méthodes supervisées ne sont dédiées qu'à un type de variable réponse particulier. On distingue également les méthodes *paramétriques*, des méthodes *non-paramétriques* selon la forme de la fonction de lien entre Y et X : les méthodes paramétriques spécifient une forme a priori pour cette fonction qui peut être modulée en fonction de certains paramètres (e.g. des coefficients de régression pour une régression linéaire). La mise en oeuvre d'une méthode paramétrique consiste alors à trouver les "bons" paramètres afin d'avoir connaissance du lien entre Y et X. Les méthodes non-paramétriques, elles, ne supposent aucune forme spécifique pour la fonction de lien, celle-ci sera alors directement déterminée à partir des données. La Table 1.1 résume la typologie des principales méthodes de data-mining supervisées.

Table 1.1: Différents méthodes supervisées

	Régression	Classification
Paramétrique		
	Régression linéaire	Régression logistique
	ANOVA	Analyse linéaire discriminante
	Modèles additifs généralisés	Modèles à classes latentes
		Modèles additifs généralisés
Non-paramétrique		
	KNN	KNN
	Arbres	Arbres
	Forêts aléatoires	Fôrêts aléatoires
	Réseau de neurones	Réseau de neurones
	Support Vector Machines	Support Vector Machines
	Splines	

L'objectif de ce document est de présenter les stratégies classiques pour déterminer le modèle le plus adapté pour un problème de régression ou de classification. Nous nous plaçons d'un point de vue assez général, sans privilégier de méthodes supervisées particulières, celles-ci pourront notamment être paramétriques ou non. Dans un premier temps nous précisons la notion de modèle, puis nous mettrons en avant le concept de compromis biais-variance qui permettra de poser les bases d'une troisième partie sur le choix de modèles à proprement parler.

2 Apprentissage statistique

2.1 Notations et contexte

On note Y la variable réponse du modèle et X le vecteur de longueur p des p variables explicatives. Ces variables sont observées sur un ensemble de n individus statistiques. $\{(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)\}$ constitue l'*échantillon d'apprentissage*. On souhaite alors apprendre de l'échantillon réalisé $\{(x_1, y_1), \dots, (x_n, y_n)\} \in \mathcal{X} \times \mathcal{Y}$ le lien qui existe entre l'entrée X et la sortie Y. $\mathcal{X}$ et $\mathcal{Y}$ sont des espaces quelconques de dimensions p et 1 respectivement. Dans un contexte de régression, $\mathcal{Y} \subset \mathbb{R}$, tandis que $\mathcal{Y} = \{m_1, \dots, m_k\}$ dans un contexte de classification.

Deux objectifs peuvent alors être considérés : la *prédiction* ou l'*estimation*. Etant donnée une nouvelle entrée $\mathbf{x}_0 \in \mathcal{X}$ (non présente dans l'échantillon d'apprentissage), la prédiction consiste à prédire la sortie $\hat{y}_0 \in \mathcal{Y}$ correspondante, la "plus proche possible" de y_0 (que l'on ne connaît pas). L'estimation consiste quant à elle à approcher au mieux la fonction qui relie Y à X.

Les deux objectifs ne sont pas nécessairement équivalents. Les sous-sections suivantes reviennent sur cette différence.

2.1.1 Régression

Dans un cadre de régression, le modèle peut en toute généralité s'écrire

$$Y = f(X) + \varepsilon \tag{2.1}$$

où $f(X) = \mathbb{E}[Y|X]$ est l'espérance de Y sachant X, ε est une variable de bruit telle que $\mathbb{E}[\varepsilon] = 0$ et on notera $\mathbb{V}\mathrm{ar}[\varepsilon] = \sigma^2$. On cherchera alors à approcher f (inconnue) par une fonction notée $\hat{f}$, ceci en s'appuyant sur l'échantillon $\{(x_1, y_1), \ldots, (x_n, y_n)\}$. Pour construire une **prédiction** pour une nouvelle entrée $\mathbf{x}_0$, on appliquera alors la fonction $\hat{f}$ à cette nouvelle entrée.

Afin de définir une "bonne prédiction", il faut se donner un critère mathématique. Pour cela, on utilise ce que l'on appelle une *fonction de contraste* qui à un prédicteur $\hat{f}$ et un couple (x_0, y_0) associe une "distance" entre $\hat{f}(x_0)$ et y_0. On la note γ.

Par exemple, il est fréquent de s'intéresser à la fonction de contraste $\gamma\left(\hat{f}, (x, y)\right) = \left(\hat{f}(x) - y\right)^2$, appelée *contraste des moindres carrés*.

Par ailleurs, il se peut que cette erreur soit petite pour certaines entrées $\mathbf{x}_0$ et grande pour d'autres. Ainsi, on s'intéresse plutôt à la *fonction de perte*, notée $P_\gamma(\hat{f})$ qui n'est rien d'autre que la moyenne des valeurs de la fonction de contraste sur l'ensemble des nouvelles observations "possibles". Formellement,

$$P_\gamma(\hat{f}) = \mathbb{E}_{X,Y}[\gamma\left(\hat{f}, (X, Y)\right)]$$

soit dans le cadre de la fonction de contraste des moindres carrés

$$P_\gamma(\hat{f}) = \mathbb{E}_{X,Y}[\left(Y - \hat{f}(X)\right)^2]. \tag{2.2}$$

NB : la notation $\mathbb{E}_{X,Y}$ permet de préciser ici que l'espérance mathématique (moralement, la "moyenne") porte sur les valeurs de (X, Y), c'est-à-dire ici les nouvelles observations qui n'ont pas servies à construire $\hat{f}$.

Notons que la fonction de perte reste une quantité aléatoire car elle dépend de l'échantillon d'apprentissage. Autrement dit, pour un échantillon d'apprentissage donné, $P_\gamma(\hat{f})$ quantifie la moyenne des erreurs attendues, mais si on change l'échantillon d'apprentissage, alors cette fonction de perte prendra une valeur différente. Ainsi, il est naturel de s'intéresser au *risque* de notre prédicteur, qui n'est rien d'autre que la moyenne des pertes sur tous les échantillons d'apprentissage possibles, nous y reviendrons en Section 3.

Pour obtenir la meilleure prédiction de Y possible à partir d'un échantillon d'apprentissage, l'idée est donc de minimiser (2.2).

Dans un but d'**estimation**, on cherchera cette fois à minimiser l'écart entre f et $\hat{f}$ selon (dans le cas des moindres carrés)

$$\mathbb{E}[\left(f(X) - \hat{f}(X)\right)^2]. \tag{2.3}$$

En réalité, dans un cadre de régression, ces deux critères sont **équivalents**. On pourra se reporter à Arlot and Bach (2015) p.3 pour la démonstration.

2.1.2 Classification

Le modèle (2.1) ne peut pas s'étendre directement au cadre de la classification. En effet, Y étant de nature qualitative, il n'y aurait aucun sens à vouloir écrire que la réponse (Y) serait égale à une somme de deux termes ($f(X)$ et ε). Ainsi, en classification, on s'intéressera plutôt à estimer la probabilité que Y prenne une certaine modalité m_q en fonction de la valeur de X, probabilité dite *a posteriori* et notée $\mathbb{P}\left(Y = m_q|X\right)$. Notons que dans le cas binaire, cette probabilité n'est rien d'autre que $\mathbb{E}[Y|X]$ et on retombe ainsi dans le cadre précédent. La recherche du "meilleur" prediteur, repose alors sur la minimisation d'une fonction de contraste adaptée au nouveau type de la variable cible.

La fonction de contraste usuelle est le contraste 0-1 :

$$\gamma\left(\hat{f}\,, (x, y)\right) = 1_{\hat{f}\,(x) \neq y}$$

où $\hat{f}$ est un prédicteur de Y (appelé *classifieur*) construit à partir de l'échantillon d'apprentissage. Cette fonction compte '0' si y vaut $\hat{f}\,(x)$ (bonne prédiction) et '1' sinon (mauvaise prédiction).

Pour prédire une valeur de Y pour un nouvel individu, on évaluera la probabilité estimée pour $X = \mathbf{x}_0$ pour chaque valeur (modalité) dans $\mathcal{Y}$ et on retiendra, par exemple, la modalité m_q telle que cette probabilité soit la plus grande.

Ainsi, étant donnée la fonction de contraste précédente, dans un but de **prédiction**, on cherchera ici à minimiser la fonction de perte :

$$P_\gamma\left(\hat{f}\right) = \mathbb{E}_{X,Y}[\gamma\left(\hat{f}\,, (X, Y)\right)] = \mathbb{P}\left(\hat{f}\,\left(X\right) \neq Y\right) \tag{2.4}$$

En revanche, dans un but d'**estimation**, on cherchera à minimiser l'écart entre les probabilités a posteriori théoriques et celles estimées à partir de l'échantillon selon

$$\mathbb{E}\left[\sum_{q=1}^{k} |\hat{p}\left(Y = m_q|X\right) - \mathbb{P}\left(Y = m_q|X\right)|\right]. \tag{2.5}$$

où $\hat{p}\left(Y = m_q|X\right)$ est l'estimation de la probabilité que $Y = m_q$ sachant que $X = \mathbf{x}$ (notée $\mathbb{P}\left(Y = m_q|X\right)$).

Contrairement au cas de la régression, il n'est pas équivalent de bien estimer $\mathbb{P}\left(Y = m_q|X\right)$ ou de bien prédire Y. En effet, supposons que la variable Y ait deux modalités (0 et 1), alors la prédiction est la même si $\hat{p}\left(Y = 1|X\right) = 0.51$ ou $\hat{p}\left(Y = 1|X\right) = 0.99$, pourtant, selon la valeur de $\mathbb{P}\left(Y = m_q|X\right)$ (inconnue) l'estimation par $\hat{p}$ peut être très bonne ou très mauvaise. On voit dans cet exemple qu'une bonne estimation **n'est pas nécessaire** pour une bonne prédiction.

2.2 Evaluer la perte pour choisir un modèle

Que ce soit dans un contexte de régression ou de classification, d'estimation ou de prédiction, les critères précédents ((2.2) à (2.5)) ne sont calculables que si on connaît le vrai modèle, ce qui n'est évidemment pas le cas en pratique. Dès lors, on ne peut pas les utiliser directement pour choisir le meilleur modèle. Néanmoins, il est possible d'en obtenir une estimation à partir des données. Pour cela, nous allons constituer un échantillon dit *échantillon test* sur lequel nous calculerons empiriquement les quantités théoriques précédentes. Il s'agira simplement de calculer la fonction de contraste pour tous les individus de l'échantillon test et d'en faire la moyenne. Typiquement, un critère des moindres carrés sera calculé dans un cadre de régression, tandis qu'un taux mauvais classement pourra être calculé dans un cadre de classification.

Par la suite, nous donnons dans un premier temps un exemple pour illustrer en quoi l'utilisation d'un échantillon d'apprentissage seul ne suffit pas et conduit à ce qu'on appelle du *sur-ajustement*. Ensuite, nous évoquons quelques stratégies pour le cas où les modalités de la variable cible sont déséquilibrées (dans un contexte de classification).

2.2.1 Illustration du problème de sur-ajustement

Considérons les données représentées dans le premier graphique en Figure 2.1. Il s'agit d'un jeu de données simulées décrivant 200 individus par 2 variables continues, et une variable de type binaire. On cherche à prédire la variable binaire à partir des deux variables continues selon la méthode des k plus proches voisins (knn). Cette méthode consiste à affecter à chaque couple (x_1, x_2) la classe majoritaire parmi les k voisins les plus proches dans le jeu de données (au sens d'une certaine distance). Ceci permet d'affecter à chaque couple (x_1,x_2), observé ou non, une modalité 0 ou 1. On représente dans le second graphique de la Figure 2.1, la classification obtenue pour $k = 15$ en utilisant une distance euclidienne. On observe une séparation assez nette entre les individus affectés à la classe 1 et ceux affectés à la classe 0. Le taux de mauvais classement calculé sur ces données vaut 15.5%. On peut diminuer ce taux en diminuant le nombre de voisins jusqu'à obtenir une erreur nulle pour 1 voisin (cf 3eme et 4eme graphiques en Figure 2.1). Ce dernier modèle, ajuste très bien les données utilisées, et pour cause, pour chaque individu du jeu de données, on prédit Y par la valeur observée sur l'échantillon ! En quelque sorte, on peut dire que le modèle que l'on a construit avec $k = 1$ est très performant pour les données considérées, mais ceci ne sera pas nécessairement vrai pour un nouvel individu non observé dans cet échantillon et n'a donc pas vraiment d'intérêt. On dira que le modèle ne *généralise* pas bien.

Pour cette raison, la perte ne peut pas être objectivement estimée à partir des données ayant servies à calculer $\hat{f}$, il faut généralement utiliser pour cela un échantillon **indépendant**. Ainsi, pour choisir un modèle on sera généralement amené à considérer deux échantillons : l'*échantillon d'apprentissage*, utilisé pour construire le modèle (i.e. à déterminer $\hat{f}$) et l'*échantillon test*, permettant d'estimer la perte et donc de faire un choix entre plusieurs modèles candidats (e.g. estimation par plus proches voisins en utilisant $k = 1$, 5 ou 15). Ces deux échantillons s'obtiennent en partitionnant les données en deux groupes via un tirage aléatoire (stratifié sur la variable cible). Généralement, on choisira un échantillon d'apprentissage deux fois plus grand que l'échantillon test. Nous reviendrons sur ce point en Section 4.2.

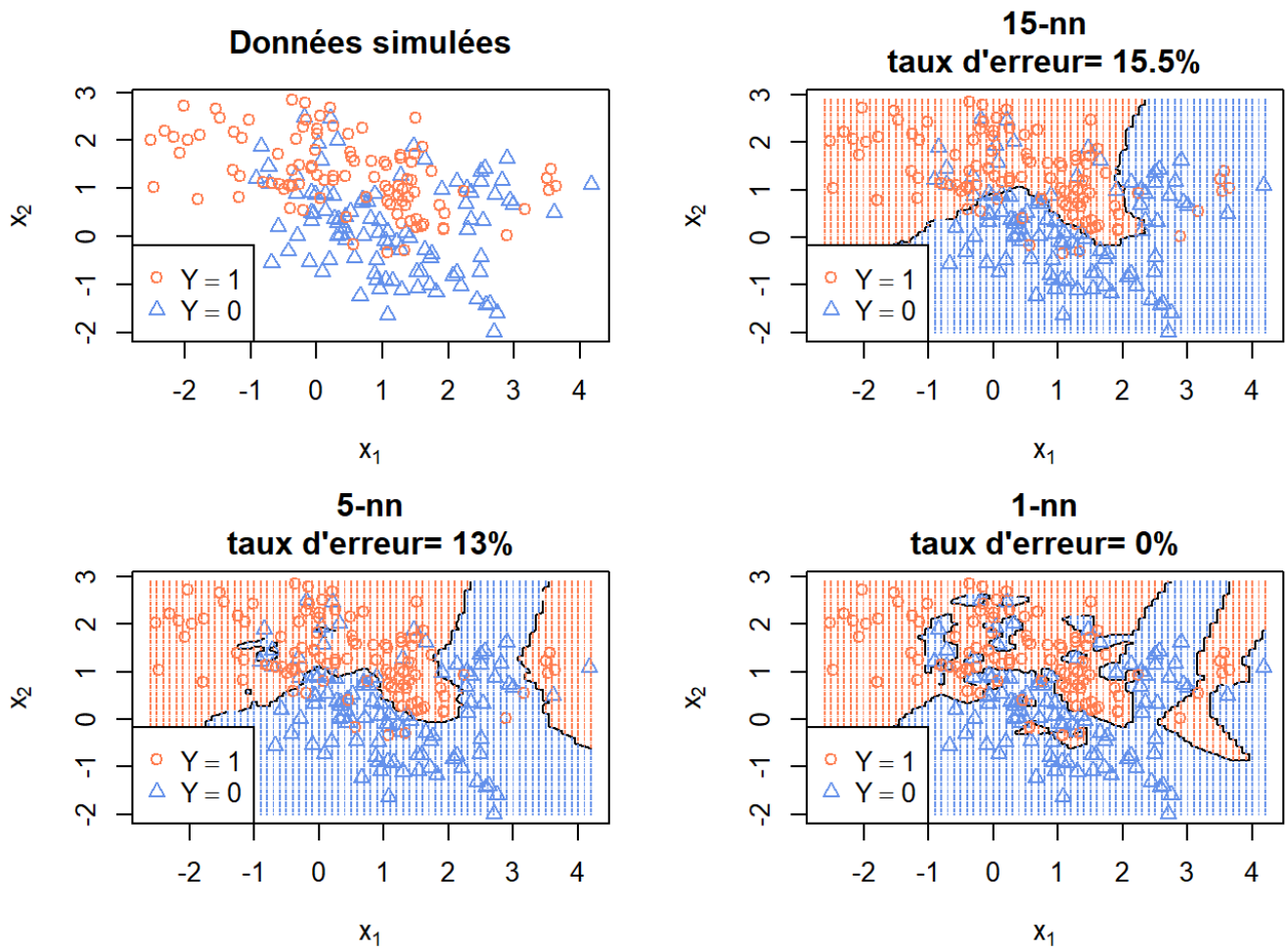

Figure 2.1: Influence de k sur la classification effectuée par k-plus proches voisins : données d'origine, classification pour k=15, k=5 et k=1

2.2.2 Données déséquilibrées

On entend par *données déséquilibrées*, le cas d'une variable réponse qualitative possédant une modalité d'effectif minoritaire par rapport aux autres (ou l'autre dans le cas binaire). Ceci est assez fréquent dans certains contextes, on l'observe notamment dans les données German Crédit où seuls 30% des dossiers correspondent à des impayés. Ce type de données pose potentiellement un problème dans la mesure où le contraste 0-1 est *symétrique*, affectant le même coût à mal classer une modalité majoritaire ou minoritaire. La conséquence de cela étant que comme les individus correspondant à la modalité majoritaire sont plus nombreux dans les données, le modèle tend à s'ajuster de façon privilégiée sur eux et par conséquent ne prédit pas bien les individus correspondants à la classe minoritaire. Dans le cas des données bancaires, ceci est très préjudiciable car on a grand intérêt à prévoir de façon fiable les dossiers qui correspondent à des impayés.

Une façon de remédier à ce problème consiste à utiliser une fonction de contraste dite *asymétrique* affectant un poids différent aux prédictions selon la modalité à prédire. Dans le cas binaire, une telle fonction de contraste s'écrit

$$\gamma\left(\hat{f}, (x, y)\right) = w_0 1_{\hat{f}(x)\neq 0, y=0} + w_1 1_{\hat{f}(x)\neq 1, y=1}$$

où $(w_0, w_1) \in \mathbb{R}^{+^2}$ sont les poids affectés aux deux erreurs possibles (faux positifs ou faux négatifs). On pourra alors choisir un poids plus grand pour la modalités la plus rare. Notons que cette fonction de contraste s'étend naturellement aux cas de $k > 2$ modalités et équivaut au contraste 0-1 si $w_0 = w_1 = 1/2$.

Une autre façon de procéder est de rééchantillonner les données pour se ramener à des données équilibrées. On distingue généralement trois types d'approches : le sur-échantillonnage, le sous-échantillonnage et les méthodes mixtes. Nous les développons par la suite.

2.2.2.1 Sur-échantillonnage

Cette technique consiste à effectuer des tirages avec remises des individus issus de la classe minoritaire de façon à augmenter la fréquence de la modalité sous-représentée. Cette approche n'est généralement pas utilisée en pratique car elle produit des doublons qui peuvent avoir tendance à favoriser le problème observé dans le cas du knn, i.e. que le modèle que l'on construit va devenir très performant pour les individus de la classe minoritaire observés, mais sans pour autant permettre de nouvelles prédictions pertinentes sur des individus jamais observés.

2.2.2.2 Sous-échantillonnage

A l'inverse, on peut tirer un sous-échantillon de l'ensemble des individus constituant la classe majoritaire pour se ramener à des effectifs identiques entre les deux classes. Cette technique peut s'envisager à condition de disposer d'un échantillon de taille suffisante.

2.2.2.3 Combinaison des deux approches : la procédure SMOTE

La procédure SMOTE (Synthetic Minority Over-sampling Technique) est une stratégie très populaire pour gérer le problème du déséquilibrage. Elle consiste à combiner le sous-échantillonnage de la classe majoritaire et le sur-échantillonnage de la classe minoritaire. La spécificité de l'approche est que le sur-échantillonnage n'est pas effectué selon un tirage avec remise, mais selon une procédure un peu plus sophistiquée évitant la production de doublons dans le nouvel échantillon. L'idée principale est de relier les individus de la classe minoritaire par des segments, puis d'effectuer des tirages de nouvelles observations en considérant des points fictifs sur ces segments, correspondant donc à des individus non-observés, mais restant géométriquement proches d'individus observés parmi la classe minoritaire.

Plus précisément, la procédure de sur-échantillonnage est la suivante : pour chaque individu i de la classe minoritaire :

1. identifier les k plus proches voisins (minoritaires). Les auteurs conseillent de choisir $k = 5$.
2. tirer un individu i' au hasard parmi les k
3. pour j de 1 à p

- tirer un nombre u au hasard selon une loi uniforme dans l'intervalle [0;1]
- construire la coordonnée j de l'individu fictif i'' selon

$$x_{i''j} \leftarrow x_{ij} + u \times \left(x_{ij} - x_{i'j}\right)$$

Cette procédure permet de doubler la taille de la classe minoritaire. Il est possible d'aller au-delà en tirant plusieurs individus (sans remise) à l'étape 2). En tirant par exemple 2 individus, on peut multiplier la taille de la classe minoritaire par 3, etc. Pour plus de détails sur la procédure SMOTE, on pourra se référer à l'article Chawla et al. (2002). La procédure est notamment disponible dans le package R *UBL* via la fonction *SmoteClassif*.

3 Compromis biais - variance

La qualité d'un modèle est étroitement liée à la complexité de celui-ci. Par exemple, dans le cas des k plus proches voisins, choisir un grand nombre de voisins conduit à avoir un modèle assez simple (il se traduit par un découpage plutôt simple de l'espace des entrées), tandis qu'un petit nombre de voisins conduit à un modèle d'une plus grande complexité. Il s'agit de trouver un compromis entre un modèle peu complexe, mais qui n'ajuste pas bien les données, et un modèle très complexe qui les ajuste trop, devenant ainsi peu pertinent pour des données ne faisant pas partie de l'échantillon d'apprentissage, autrement dit, ne généralisant pas bien.

Pour l'illustrer, nous découpons le jeu de données German Credit en un échantillon d'apprentissage et un échantillon test (selon un découpage stratifié 2/3 1/3), puis évaluons le taux de mauvais classement pour ces deux échantillons en fonction du nombre de voisins.

Remarque : s'agissant ici d'une illustration, nous nous permettons de simplifier le problème en ne considérant uniquement les variables quantitatives, mais on pourrait utiliser l'ensemble des variables en travaillant sur les composantes d'une analyse factorielle des données mixtes.

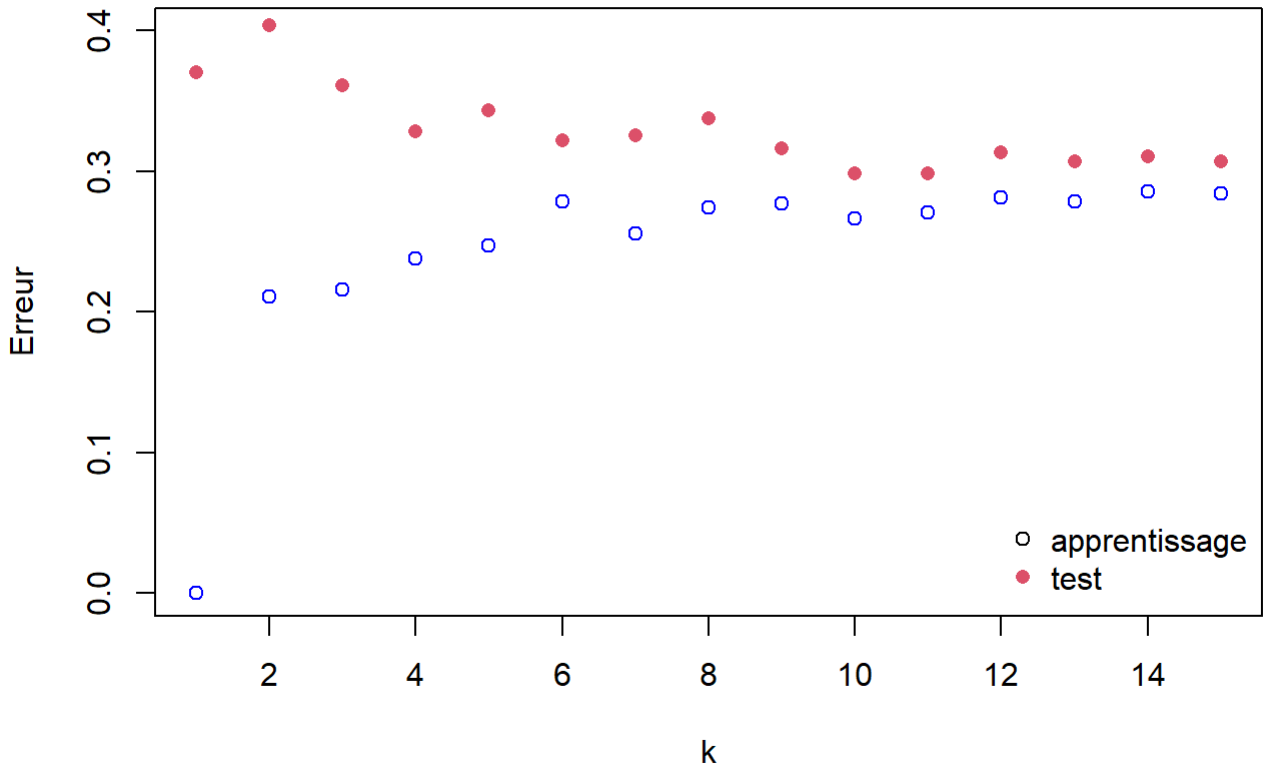

Figure 3.1: Taux de mauvais classement en fonction de la complexité du modèle

On constate que si le modèle est complexe (k petit), alors l'erreur d'apprentissage (i.e. le taux de mauvais classement calculé sur l'échantillon d'apprentissage) est faible, mais l'erreur de test (i.e. le taux de mauvais classement calculé sur l'échantillon test) est grande. Pour un modèle moins complexe (k plus grand), l'erreur d'apprentissage tend à augmenter, mais l'erreur sur l'échantillon test diminue pour atteindre un minimum en $k = 10$. Au-delà, le modèle est trop peu complexe et l'erreur sur l'échantillon test augmente de nouveau. Sur cet exemple, choisir $k = 10$ correspond au modèle proposant le meilleur compromis en termes de complexité parmi les modèles considérés (variant par le choix de k). Notons que ce phénomène est également observé en régression.

Ce compromis porte le nom de *compromis biais-variance*. Le biais fait référence à l'écart entre la fonction de lien f et le modèle choisi. Plus le prédicteur est simple (par exemple en prenant k grand en knn), plus il tend à s'éloigner de la fonction f et donc plus le biais est grand. La variance fait quant à elle référence aux variations du prédicteur en fonction de l'échantillon d'apprentissage considéré. Plus le prédicteur est simple, moins il "varie" en fonction de l'échantillon d'apprentissage. On note que ce compromis est donc étroitement lié à l'échantillon d'apprentissage.

Afin de justifier plus en détail cette dénomination, nous allons considérer le cas où la variable Y est quantitative, mais le même concept prévaut également pour une variable qualitative. Pour cela, on reprend le modèle (2.1), on se donne une estimation de f et on considère une nouvelle entrée $\mathbf{x}_0$ (non utilisée pour

estimer f). Nous allons décomposer le *risque* du prédicteur, autrement dit la moyenne de la fonction de contraste sur les échantillons d'apprentissage, pour une entrée x_0 donnée. Ce risque peut se réécrire comme suit pour une fonction de contraste des moindres carrés :

$$\begin{aligned} E\left[(y_0 - \hat{y}_0)^2\right] &= E\left[\left(f(x_0) + \varepsilon_0 - \hat{f}(x_0)\right)^2\right] \\ &= E\left[\left(\left(f(x_0) - \hat{f}(x_0)\right) + \varepsilon_0\right)^2\right] \\ &= E\left[\left(f(x_0) - \hat{f}(x_0)\right)^2\right] + \sigma^2 \\ &= \underbrace{Bais^2(\hat{f}(x_0)) + Var\left[\hat{f}(x_0)\right]}_{\text{réductible}} + \underbrace{\sigma^2}_{\text{irréductible}} \end{aligned}$$

Cette quantité peut donc être décomposée en trois termes. Le premier terme est $Bais^2(\hat{f}(x_0))$, le biais au carré commis sur la prédiction de y_0. Le biais de $\hat{f}(x_0)$ correspond alors à l'écart moyen entre $f(x_0)$ (i.e. la vraie valeur y_0) et $\hat{f}(x_0)$ pour tous les échantillons d'apprentissage possibles. Le deuxième terme est $Var\left[\hat{f}(x_0)\right]$, la variance de la prédiction pour $\mathbf{x}_0$, i.e. la moyenne des écarts au carré entre $\hat{f}(\mathbf{x}_0)$ et sa valeur moyenne pour tous les échantillons d'apprentissage possibles. Enfin le troisième terme σ^2 est la variance des erreurs.

La variance des erreurs est irréductible car elle dépend seulement des données, tandis que les deux autres termes dépendent de $\hat{f}$ et peuvent donc être potentiellement diminués en choisissant une autre estimation de f. Plus le modèle sera complexe, plus le biais de $\hat{f}$ sera faible, au détriment de la variance qui va augmenter. Ainsi, choisir la complexité d'un modèle en se basant sur l'erreur de test revient à rechercher un bon compromis biais-variance.

La figure3.2 résume le comportement typique de l'erreur de test et d'apprentissage en fonction de la complexité du modèle. L'erreur d'apprentissage tend à diminuer quand la complexité du modèle augmente. Au contraire, plus le modèle est complexe, plus l'erreur de test augmente.

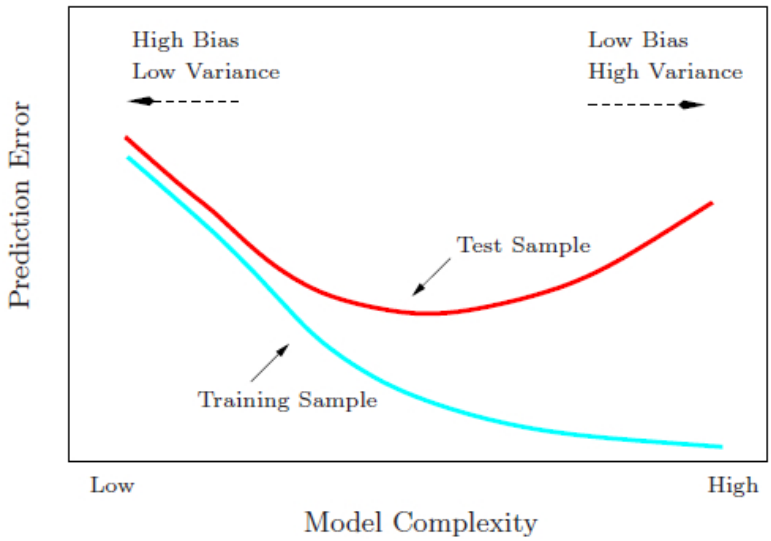

Figure 3.2: Erreur de test et d'apprentissage en fonction de la complexité du modèle (Source : Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. The Elements of Statistical Learning. Springer Series in Statistics. New York, NY, USA: Springer New York Inc.)

Nous avons supposé que la variable réponse Y était quantitative, mais la recherche d'un bon compromis biais-variance est tout aussi centrale quand Y est qualitative.

4 Choix de modèles

La gamme des modèles à disposition pour prédire une variable réponse est large : on peut choisir plusieurs familles de méthodes (modèle linéaire, k plus proches voisins, arbres de régression, etc.) et au sein de ces familles, plusieurs choix sont envisageables (e.g. choix des variables explicatives dans le modèle linéaire, choix de k pour les plus proches voisins, nombres de noeuds dans l'arbre, etc.). La section précédente a permis d'illustrer que le meilleur modèle n'est pas nécessairement celui qui ajuste au mieux les données d'apprentissage, mais que celui-ci doit aussi être d'une complexité limitée de façon à assurer un bon compromis biais/variance. Nous présentons maintenant les stratégies classiques pour effectuer un bon choix de modèle.

4.1 Critères pénalisés

L'estimation de la fonction de lien f s'effectue généralement en minimisant le risque empirique $\frac{1}{n}\sum_{i=1}^{n}\gamma(\hat{f}\,,(x_i,y_i)$ obtenu à partir de l'échantillon d'apprentissage. Comme cette quantité tend à diminuer au fur et à mesure que la complexité du modèle augmente, une stratégie classique est d'ajouter à ce critère une quantité d'autant plus grande que le modèle est complexe. Les critères AIC (Akaïke Information Criterion) et BIC (Bayesian Information Criterion) sont les représentants les plus classiques de ce type de critères, ils permettent de trouver un compromis pour des modèles paramétriques issus d'une même famille.

Dans cette partie, on supposera que la famille de modèles considérée est paramétrée par θ. Par exemple, en régression logistique, $\theta = (\beta_0, \ldots, \beta_p)$ est l'ensemble des coefficients de régression, on cherche alors à modéliser la distribution d'une variable réponse Y (binaire) conditionnellement à des variables explicatives X. Celle-ci admet une fonction de densité, notée $g(y;\theta)$. Une façon classique d'estimer θ est alors de procéder par maximum de vraisemblance. La vraisemblance évaluée en $\hat{\theta}$ permet de mesurer l'adéquation d'un modèle aux données puisqu'elle représente la probabilité d'avoir observé l'échantillon sous le modèle. En pratique, on utilise plutôt la log-vraisemblance, plus commode pour les calculs.

Notons ici que l'existence d'une densité ouvre la voie à d'autres fonctions de contraste que les moindres carrés, en particulier le contraste *log-vraisemblance* : en maximisant la log-vraisemblance, on ne fait que minimiser son opposé, on retombe bien ainsi sur la minimisation d'une fonction de contraste particulière.

4.1.1 AIC

L'AIC est défini selon

$$AIC = -2lnL(\hat{\theta}) + 2k$$

où k correspond au nombre de paramètres du modèle.

Ce critère repose sur des hypothèses asymptotiques. On peut montrer que plus n est grand, plus l'AIC maximisera la vraisemblance pour de futures données. Il est donc un critère à privilégier dans une optique de prédiction. En revanche, rien n'assure que le bon modèle sera sélectionné s'il fait partie de la famille considérée.

4.1.2 BIC

Le BIC est défini selon

$$BIC = -2lnL(\hat{\theta}) + ln(n)k$$

Ce critère repose sur des justifications Bayésiennes non développées ici. Par rapport à l'AIC, on voit que quand n est grand, la pénalité est plus forte que pour l'AIC ce qui conduit à conserver des modèles plus parcimonieux (i.e. comportant moins de paramètres). On peut montrer que si le vrai modèle fait partie de la famille considérée, alors si n tend vers l'infini, le bon modèle est sélectionné presque sûrement. Ceci en fait un critère à privilégier dans une optique d'estimation (Saporta (2006)).

L'AIC et le BIC sont les critères pénalisés les plus classiques, mais ils en existent beaucoup d'autres. Citons par exemple le R^2 ajusté ou le C_p de Mallows.

4.1.3 Exploration de l'ensemble des modèles

Les critères précédents permettent de comparer des modèles d'une même famille bien qu'ayant un nombre de paramètres différents. Néanmoins, pour une famille de modèles donnée, par exemple les modèles de régression linéaire, le nombre de modèles candidats est rapidement trop grand pour qu'il soit possible de calculer les critères précédents pour chacun d'entre eux ($2^p - 1$ pour un modèle de régression pour p variables). Typiquement, on delà d'une dizaine de variables, il ne sera pas envisageable de considérer l'intégralité des modèles possible. Dès lors, on utilise des algorithmes qui vont parcourir l'ensemble des modèles pour en retenir un candidat, qui ne sera pas nécessairement le meilleur, mais qui sera au moins un bon modèle.

Ces algorithmes sont appelés *méthodes pas à pas*. Ils consistent à partir d'un certain modèle, puis à ajouter, ou à enlever des variables explicatives de façon successive. Parmi eux, la méthode *descendante* consiste à partir d'un modèle incluant l'ensemble des variables explicatives, puis à éliminer une variable de façon à optimiser un critère (par exemple l'AIC), on recommence alors jusqu'à ce qu'il ne soit plus possible d'optimiser le critère en éliminant une nouvelle variable ou si toutes les variables sont éliminées. La méthode *ascendante* consiste au contraire, à partir du modèle vide (sans variables explicatives) puis à intégrer la variable qui permet d'optimiser le critère choisi. On procède itérativement jusqu'à ce qu'il ne soit plus possible d'optimiser le critère, ou quand toutes les variables ont déjà été intégrées. Un schéma synthétique de la méthode ascendante est présenté en Figure 4.1. Enfin, la méthode *progressive* est une méthode mixte, qui suit le même principe que la méthode ascendante, sauf que l'on peut éliminer des variables déjà introduites. Cette approche est celle communément adoptée car elle permet d'éviter les redondances dans les variables explicatives.

Pour plus de précisions sur les approches pas à pas, le lecteur pourra par exemple consulter le livre Cornillon and Matzner-Lober (2010).

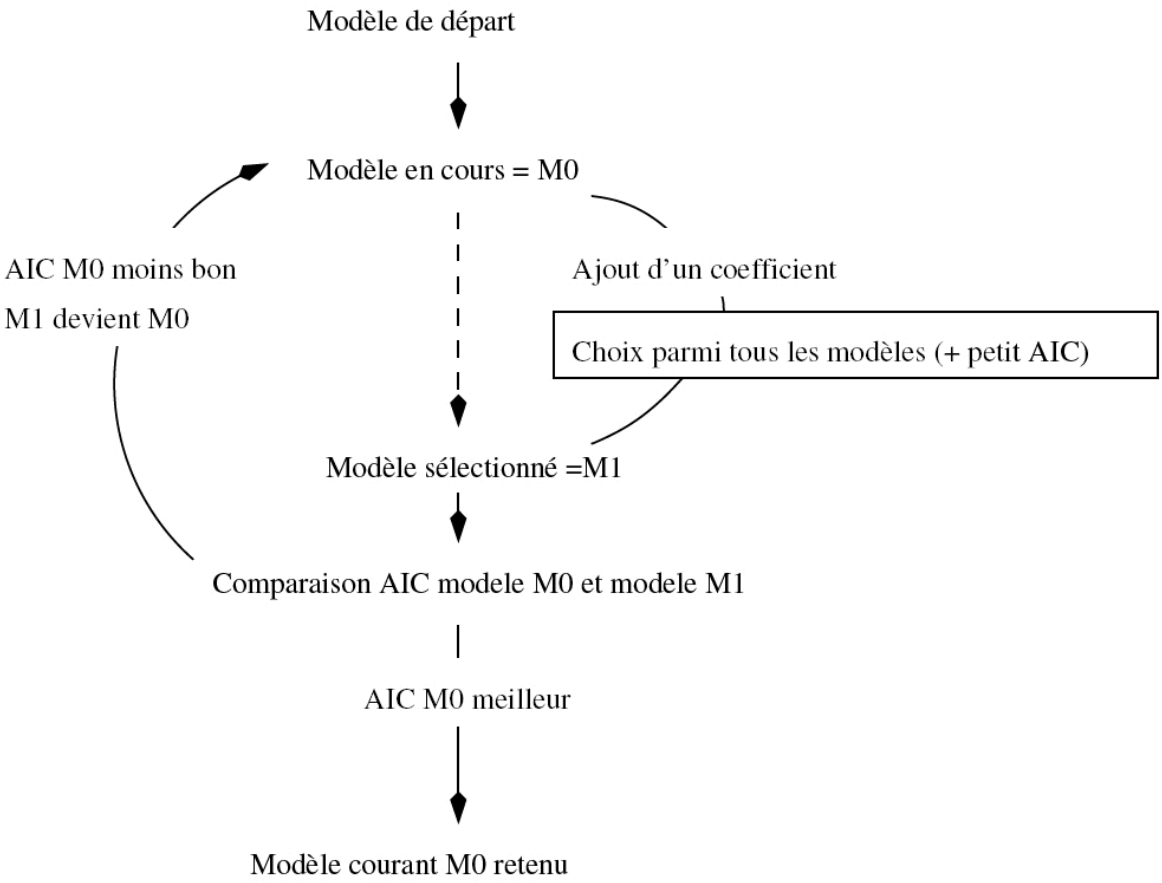

Figure 4.1: Algorithme de sélection pas à pas ascendante. Source : Cornillon et Matzner-Lober 2011

4.1.4 Limites

Ces critères sont d'usage très courant mais ne peuvent pas être envisagés pour n'importe quel modèle. En effet, il faut tout d'abord un modèle paramétrique afin de pouvoir déterminer la vraisemblance (ce n'est pas le cas des k plus proches voisins par exemple). Par ailleurs, ils ne permettent pas de comparer des modèles issus de familles différentes. De plus, le nombre de paramètres ne traduit pas toujours la complexité du modèle (c.f. théorie de l'apprentissage de V. Vapnik). Par exemple, pour un modèle de régression linéaire à p variables explicatives, le nombre de paramètres vaut $p+1$, mais si on utilise une méthode de régression avancée ceci n'est plus vrai. La régression ridge par exemple consiste, une fois les coefficients de régression obtenus, à modifier leur valeur de façon à se rapprocher d'un modèle plus simple (celui où tous les coefficients seraient nuls), tout en gardant un modèle à p variables. Pour une telle méthode, le nombre de paramètres reste le même qu'en régression linéaire classique, mais la complexité est plus faible. Enfin, s'ils permettent de faire un choix parmi une famille de modèles, ils ne permettent généralement pas d'en apprécier les capacités prédictives.

4.2 Approches empiriques

Leur principe est de constituer un échantillon d'apprentissage pour ajuster les différents modèles et un échantillon test pour n'en sélectionner qu'un seul.

4.2.1 Découpage test et apprentissage

Cette stratégie a déjà été partiellement évoquée en Section 2.2 avec la méthode des k plus proches voisins où les modèles envisageables variaient par le nombre de voisins considérés ($k=1$ ou $k=2$, etc). Le principe est de constituer un échantillon test et un échantillon d'apprentissage. Pour chaque k on ajuste alors le modèle (illustré dans les graphiques de la Figure 2.1) à partir de l'échantillon d'apprentissage, ce qui nous amène à M modèles candidats. Plus généralement, il est possible de considérer des modèles issus de

familles différentes (*e.g.* k plus proches voisins, régression logistique, analyse discriminante, etc). Pour choisir entre ces M modèles, on évalue la perte à partir de l'échantillon test. On retient alors le modèle pour lequel l'erreur en test est la plus faible (comme évoqué en Section 3). **Une fois ce modèle choisi on le ré-ajustera en considérant l'intégralité des données à disposition**.

Néanmoins, l'erreur calculée sur l'échantillon test reste potentiellement sensible au découpage effectué. Afin de l'illustrer, on représente en Figure 4.2 les erreurs de tests et d'apprentissage pour différents découpages (apprentissage/test).

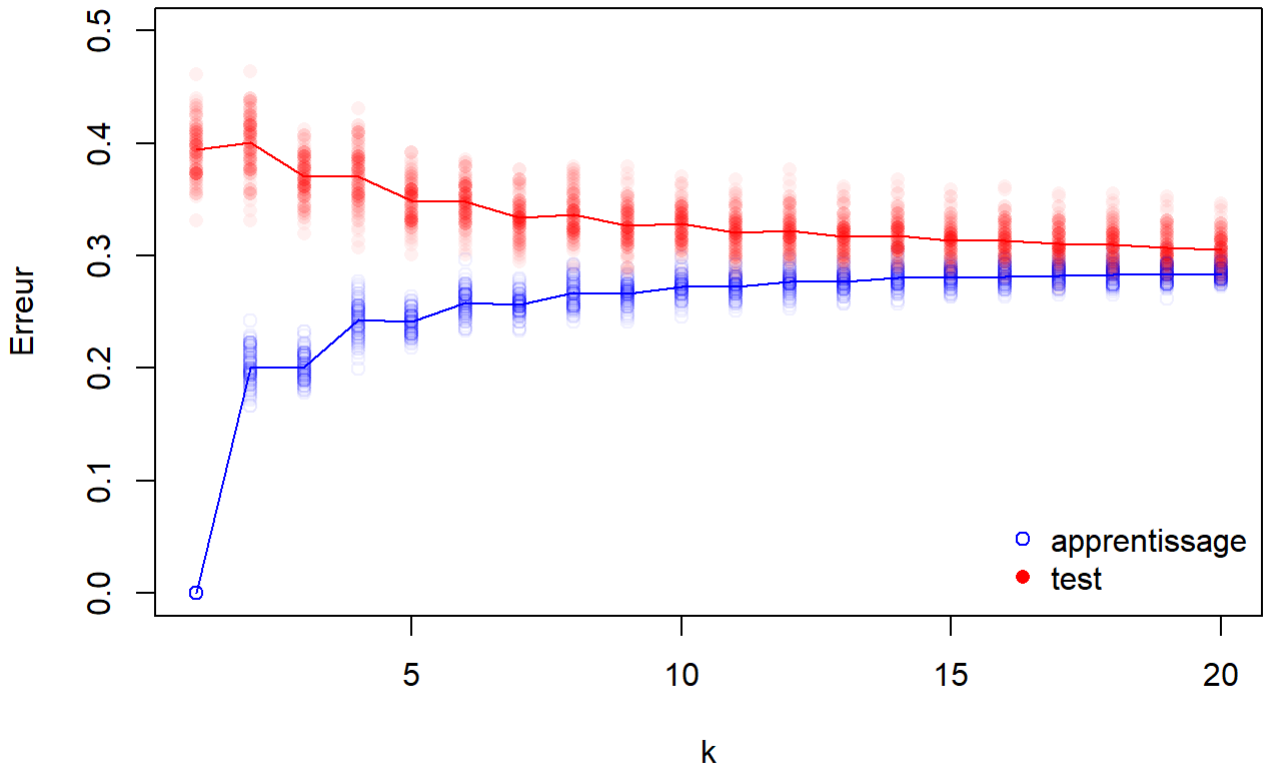

Figure 4.2: Taux de mauvais classement en fonction de la complexité du modèle pour 100 découpages de l'échantillon. Les erreurs moyennes sont représentées par une ligne.

On peut alors préférer retenir le modèle qui minimise l'erreur moyenne obtenues sur différents découpages apprentissage/test plutôt que ne considérer que l'un d'entre eux.

Aussi, si l'échantillon est de taille modeste (disons inférieure à 1000 pour fixer les idées), on risque d'effectuer un choix peu pertinent (car nécessairement, les jeux d'apprentissage et de test seront de taille encore plus petite). Pour remédier à cela, on sera amené à utiliser une approche de validation croisée.

4.2.2 Validation croisée

Les techniques de validation croisée consistent à découper l'échantillon en K blocs de même taille (i.e. avec le même nombre d'observations). L'ensemble des $K-1$ premiers sert alors d'échantillon d'apprentissage et le Kème sert d'échantillon test. On obtient ainsi une première estimation de la perte. On recommence ensuite $K-1$ fois l'opération en utilisant le $K-2$eme bloc comme échantillon test et les autres comme échantillon d'apprentissage, etc. Ainsi, on obtient K erreurs de test que l'on moyenne ce qui nous donne une estimation du risque plus précise que celle obtenue par un simple découpage test-apprentissage.

On choisit généralement $K=10$, mais ce paramètre peut être choisi plus petit de façon à gagner en temps de calcul, ou plus grand de façon à gagner en précision. Dans le cas où on choisit $K=n$ on parle de *leave-one-out*, tandis que l'on parle de validation par $K-fold$ sinon (Tufféry (2007), Saporta (2006)).

4.2.3 Mesure objective de la perte

Le découpage en échantillon en deux parties (apprentissage - test) permet de choisir un modèle mais il ne permet pas d'avoir une estimation sans biais de la perte associée au modèle final. En effet, comme on s'est servi de l'échantillon test pour choisir le modèle, celui-ci peut être considéré comme une partie du jeu d'apprentissage. Pour évaluer la performance du modèle de façon objective, il faut pouvoir évaluer cette perte sur un échantillon indépendant, d'où la nécessité de disposer d'un 3eme échantillon. Ainsi, on considérera généralement : un échantillon d'*apprentissage* (pour ajuster les modèles), un échantillon de *validation* (pour choisir entre eux) et enfin un échantillon *test* (pour mesurer objectivement la perte). Si le nombre d'observations le permet, ce découpage pourra être fait typiquement en proportion 1/2, 1/4, 1/4 (voir Hastie, Tibshirani, and Friedman (2009), page 222). La procédure de sélection est alors la suivante :

1. on ajuste les différents modèles sur l'échantillon d'apprentissage
2. on choisit le "meilleur" modèle à partir de la perte calculée sur l'échantillon de validation
3. on ajuste le meilleur modèle sur les données rassemblant échantillon d'apprentissage et de validation
4. on évalue la perte sur l'échantillon test
5. on ré-ajuste ce modèle sur l'ensemble des 3 échantillons

Par exemple, supposons que l'on considère les 3 modèles de régression suivants :

- $\mathcal{M}_1 : Y = f^{(1)}(X) + \epsilon^{(1)} = \beta_0^{(1)} + \beta_1^{(1)} \times X_1 + \beta_2^{(1)} \times X_2 + \epsilon^{(1)}$
- $\mathcal{M}_2 : Y = f^{(2)}(X) + \epsilon^{(2)} = \beta_0^{(2)} + \beta_1^{(2)} \times X_1 + \beta_2^{(2)} \times X_2 + \beta_{12}^{(2)} \times X_1 X_2 + \epsilon^{(2)}$
- $\mathcal{M}_3 : Y = f^{(3)}(X) + \epsilon^{(3)} = \beta_0^{(3)} + \beta_1^{(3)} \times X_1^2 + \beta_2^{(3)} \times X_2 + \epsilon^{(3)}$

L'étape 1) consiste à estimer les paramètres de chacun de ces modèles à partir de l'échantillon d'apprentissage. On obtient alors trois estimations de fonctions de lien notées $\widehat{f}^{(1)}$,$\widehat{f}^{(2)}$,$\widehat{f}^{(3)}$. Par exemple, si $\widehat{\beta}_0^{(1)}$, $\widehat{\beta}_1^{(1)}$, $\widehat{\beta}_2^{(1)}$ sont les estimations des paramètres $\beta_0^{(1)}$, $\beta_1^{(1)}$, $\beta_2^{(1)}$ du modèle $\mathcal{M}_1$, on a $\widehat{f}^{(1)}(X) = \widehat{\beta}_0^{(1)} + \widehat{\beta}_1^{(1)} \times X_1 + \widehat{\beta}_2^{(1)} \times X_2$.

L'étape 2) consiste à choisir entre les 3 modèles. Pour cela on utilise l'échantillon de validation pour comparer les 3 valeurs de perte et on retient le modèle conduisant à la plus faible.

Admettant que le modèle $\mathcal{M}_1$ soit celui qui minimise cette perte, l'étape 3) consiste à estimer de nouveau la fonction de lien $f^{(1)}$ en considérant dorénavant l'échantillon d'apprentissage **et** celui de validation.

L'étape 4) consiste à calculer la perte pour cette nouvelle fonction de lien. Cette erreur sert simplement à fournir une mesure objective de la perte et peut être vue comme un indice de qualité du modèle construit.

L'étape 5) permet de fournir un modèle de qualité potentiellement supérieure, car s'appuyant sur un échantillon d'apprentissage plus grand (constitué de l'ensemble des individus disponibles). Toutefois, ayant utilisé l'intégralité des données pour le construire, il n'est plus possible d'en estimer la perte associée. On se contentera simplement de la borne supérieure obtenue à l'étape 4).

Cet exemple n'a pour objectif que de fixer les idées sur la **procédure qu'il convient de suivre**. En pratique, on aura tout intérêt à considérer un ensemble de modèles bien plus grand, et non limité à des modèles issus d'une même famille.

4.3 Complément pour la classification supervisée

En présence d'une variable réponse possédant seulement deux modalités (0, 1), on évalue les performances prédictives du modèle en considérant souvent d'autres indicateurs que le taux de mauvais classement, en particulier l'AUROC (Area Under the Receiver Operating Characteristic), ou simplement AUC (Area under the

curve), aire sous la courbe en français. Cet indicateur synthétise les deux types d'erreur qui peuvent être commises par le modèle : prédire Y par 1 alors que $Y = 0$ (faux positif) ou prédire Y par 0 alors que $Y = 1$ (faux négatif).

Plus précisément, une prédiction pour une variable binaire peut s'obtenir en retenant la modalité 1 si $\hat{p}\left(Y = 1|X\right) > 0.5$ et 0 sinon, mais plus généralement, on peut choisir n'importe quel seuil s ($s \in [0;1]$) tel que $\hat{p}\left(Y = 1|X\right) > s$ plutôt que le seuil 0.5. La courbe ROC représente $1 - \mathbb{P}\left(\hat{f}\left(X\right) = 0|Y = 1\right)$ en fonction de $\mathbb{P}\left(\hat{f}\left(X\right) = 1|Y = 0\right)$, i.e. le taux de vrais positifs en fonction du taux de faux positifs pour différents seuils s. Un exemple de courbe ROC est indiqué en Figure 4.3.

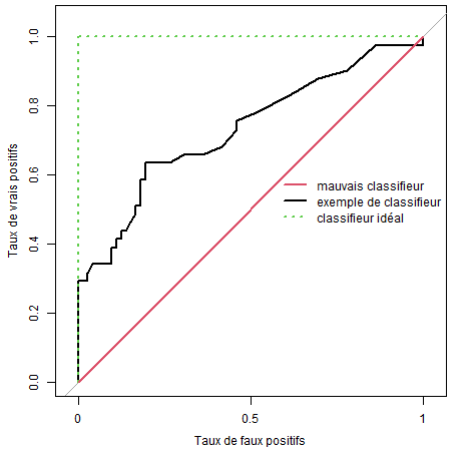

Figure 4.3: Exemple de courbe ROC

Si $s \leq 0$, alors Y est toujours prédit par 0, donc il y a 0% de faux positifs et 0% de vrais positifs. Au contraire, si $s \geq 1$ alors il y a 100% de faux positifs et 100% de vrais positifs. De ceci résultent les deux points de coordonnées respectives (0,0) et (1,1) sur la Figure 4.3. Les autres points de la courbe ROC (représentée en noir sur la Figure 4.3) correspondent aux points obtenus pour des valeurs intermédiaires de s. La droite reliant ces deux points possède une AUC de 0.5 et correspond à un modèle tel que $\hat{f}\left(X\right) = 0.5$ quelque soit la valeur de X, i.e. un modèle tirant à pile ou face la modalité prédite et donc sans pouvoir prédictif. Au contraire, un modèle dont les probabilités de succès estimées $\hat{p}\left(Y = 1|X\right)$ vaudraient systématiquement 1 quand Y vaut 1 et 0 sinon aurait une courbe ROC formant un angle droit dont l'AUC vaut 1 (courbe en pointillés en Figure 4.3).

La courbe ROC permet de comparer les performances de prédicteurs différents d'un point de vue global, mais les courbes ROC peuvent se croiser, ce qui implique que deux prédicteurs peuvent avoir un même AUC, mais qu'à taux de faux positifs identiques, l'un peut avoir un taux de vrais positifs plus élevé et est donc ponctuellement meilleur. Pour comparer des classifieurs localement, on pourra utiliser l'AUC partiel dont le principe est de calculer l'aire sous la courbe définie pour un seuil s variant entre s_{min} et s_{max} plutôt qu'un seuil variant dans $[0;1]$. Par exemple, si on souhaite choisir un classifieur avec un faible taux de faux positifs, on choisira $s_{min} = 1 - \Delta$ et $s_{max} = 1$ avec Δ plus ou moins grand en fonction à assurer un taux de faux positifs suffisamment petit.

Notons enfin que l'indice de Gini, défini comme égal à (2$\times$ AUC)-1 est une autre mesure équivalente à l'AUC souvent utilisée pour évaluer la performance d'un modèle.

5 Conclusion

Construire des modèles a toujours été une activité des statisticiens. Un modèle est un résumé global des relations entre variables, permettant de comprendre des phénomènes, et d'émettre des prévisions. A ce titre, G. Box aimait à rappeler que *tous les modèles sont faux, certains sont utiles*. Il serait en effet illusoire

de croire que la relation réelle entre une variable réponse et des variables explicatives peut être retrouvée. Celle-ci est en générale trop complexe pour pouvoir être identifier par les différentes familles de méthodes supervisées. Le problème est donc plus de trouver un bon modèle, que de rechercher le modèle vrai.

Les modèles envisageables sont vastes, mais il n'est pas possible ni souhaitable de tous les envisager. Un premier choix peut être fait a priori en fonction des besoins (construire un modèle interprétable ou non) et de la connaissance que l'on peut avoir sur les données. Par exemple, si on sait que la relation entre les variables explicatives et la réponse n'est pas linéaire, il n'est pas pertinent d'aller mettre en oeuvre des méthodes purement linéaires.

Enfin, il faut noter qu'une alternative au choix de modèles est d'employer des techniques d'agrégation comme le bagging ou le boosting. Leur principe est de combiner les prédictions des différents modèles plutôt que de choisir entre toutes. Ces approches seront développées ultérieurement dans ce cours.

Références

Arlot, Sylvain. 2011. "Sélection de Modèles Et Sélection d'estimateurs Pour l'apprentissage Statistique (Cours Peccot)." https://www.imo.universite-paris-saclay.fr/~sylvain.arlot/enseign/2011Peccot/peccot_notes_cours1.pdf (https://www.imo.universite-paris-saclay.fr/~sylvain.arlot/enseign/2011Peccot/peccot_notes_cours1.pdf).

Arlot, Sylvain, and Francis Bach. 2015. "Apprentissage Statistique M2 Probabilités Et Statistiques, Université Paris-Sud. Cours 1 : Théorie de l'apprentissage Statistique: De Vapnik à La Localisation." https://www.imo.universite-paris-saclay.fr/~sylvain.arlot/enseign/2015Orsay/Cours1.pdf (https://www.imo.universite-paris-saclay.fr/~sylvain.arlot/enseign/2015Orsay/Cours1.pdf).

Chawla, Nitesh V, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. "SMOTE: Synthetic Minority over-Sampling Technique." *Journal of Artificial Intelligence Research* 16: 321–57. https://arxiv.org/pdf/1106.1813.pdf (https://arxiv.org/pdf/1106.1813.pdf).

Cornillon, P. A., and E. Matzner-Lober. 2010. *Régression Avec r*. Pratique r. Springer.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc. https://hastie.su.domains/ElemStatLearn/printings/ESLII_print12_toc.pdf (https://hastie.su.domains/ElemStatLearn/printings/ESLII_print12_toc.pdf).

Saporta, G. 2006. *Probabilités, Analyse Des Données Et Statistique*. Editions Technip.

Tufféry, S. 2007. *Data Mining Et Statistique décisionnelle: L'intelligence Des Données*. Editions Technip.

Wikistat. 2016. "Apprentissage Machine / Statistique — Wikistat." http://wikistat.fr/pdf/st-m-Intro-ApprentStat.pdf (http://wikistat.fr/pdf/st-m-Intro-ApprentStat.pdf).

———. 2023. "Qualité de Prévision Et Risque — Wikistat." http://wikistat.fr/pdf/st-m-app-risque.pdf (http://wikistat.fr/pdf/st-m-app-risque.pdf).