

# STA211 : Eléments fondamentaux pour méthodes supervisées

Vincent Audigier, Ndèye Niang-Keita

03 avril, 2019

- 1 Introduction
- 2 Apprentissage statistique
  - 2.1 Notations et contexte
    - 2.1.1 Régression
    - 2.1.2 Classification
  - 2.2 Choisir un ``bon'' modèle
    - 2.2.1 Remarque pour le cas binaire
- 3 Compromis biais - variance
- 4 Choix de modèles
  - 4.1 Critères pénalisés
    - 4.1.1 AIC
    - 4.1.2 BIC
    - 4.1.3 Exploration de l'ensemble des modèles
    - 4.1.4 Limites
  - 4.2 Approches empiriques
    - 4.2.1 Découpage test et apprentissage
    - 4.2.2 Validation croisée
    - 4.2.3 Mesure objective de l'erreur de généralisation
- 5 Conclusion
- Références

## 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: Différentes méthodes supervisées

Régression	Classification
<b>Paramétrique</b>	
Régression linéaire	Régression logistique
ANOVA	Analyse linéaire discriminante
Modèles additifs généralisés	Modèles à classes latentes
	Modèles additifs généralisés
<b>Non-paramétrique</b>	
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  $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\} \in \mathcal{X} \times \mathcal{Y}$  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  $\mathbf{x}_0$  (non présente dans l'échantillon d'apprentissage), la prédiction consiste à prédire la sortie  $\hat{y}_0$  correspondante, la "plus proche" possible de  $y_0$ . L'estimation consiste quant à elle à approcher au mieux le lien qui relie  $Y$  à  $X$ . Les deux objectifs ne sont pas

nécessairement équivalents.

## 2.1.1 Régression

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

$$Y = f(X) + \varepsilon \quad (2.1)$$

où  $\varepsilon$  est une variable de bruit telle que  $\mathbb{E}[\varepsilon] = 0$  et on notera  $\text{Var}[\varepsilon] = \sigma^2$ . On cherchera alors à approcher  $f$  (inconnue) par une fonction notée  $\hat{f}$ . Pour construire une prédition pour une nouvelle entrée  $\mathbf{x}_0$ , on appliquera alors la fonction  $\hat{f}$  à cette nouvelle entrée. Pour obtenir la meilleure prédition de  $Y$  possible, on cherchera généralement  $\hat{f}$  telle que l'erreur de généralisation soit la plus faible possible :

$$\mathbb{E}[(Y - \hat{f}(X))^2] \quad (2.2)$$

. Autrement dit, on cherchera à minimiser l'écart quadratique entre  $Y$  et  $\hat{Y}$ .

Dans un but d'estimation, on cherchera cette fois à minimiser l'écart entre  $f$  et  $\hat{f}$  selon

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

En réalité, dans un cadre de régression, ces deux critères sont équivalents.

## 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  $\varepsilon$ ). 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  $P(Y = m_q | X)$ . 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, dans un but de prédition, on cherchera à minimiser l'erreur de généralisation :

$$P(\hat{f}(X) \neq Y) \quad (2.4)$$

où  $\hat{f}$  est un prédicteur de  $Y$  (appelé classifieur) construit à partir de l'échantillon d'apprentissage.

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

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

où  $\hat{p}(Y = m_q | X)$  est l'estimation de la probabilité que  $Y = m_q$  sachant que  $X = \mathbf{x}$ . Contrairement au cas de la régression, il n'est pas équivalent de bien estimer  $P(Y = m_q | X)$  ou de bien prédire  $Y$ . En effet, supposons que la variable  $Y$  ait deux modalités (0 et 1), alors la prédition est la même si  $\hat{p}(Y = 1 | X) = 0.51$  ou  $\hat{p}(Y = 1 | X) = 0.99$ , pourtant, selon la valeur de  $P(Y = m_q | X)$  (inconnue)

l'estimation par  $\hat{p}$  l'estimation peut être très bonne ou très mauvaise.

## 2.2 Choisir un ``bon'' modèle

Que ce soit dans un contexte de régression ou de classification, les critères précédents ne sont calculables que si on connaît le vrai modèle, ce qui n'est évidemment pas le cas en pratique. Néanmoins, il est assez facile d'obtenir une estimation de l'erreur de généralisation à partir des données. En effet, celle-ci peut être estimée par le critère des moindres carrés, dans un cadre de régression, ou par le taux mauvais classement, dans un cadre de classification.

Toutefois, il ne serait pas pertinent de vouloir rechercher la fonction  $\hat{f}$  qui minimise le critère des moindres carré ou le taux de mauvais classement. Pour le comprendre, 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 (estimation de l'erreur de généralisation en classification) 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  $\hat{Y}$  par la valeur observée sur l'échantillon ! En quelque sorte, on peut dire que le modèle que l'on a construit est très performant pour les données considérées, mais ceci ne sera pas nécessairement vrai pour un nouvel individu non observé dans cet échantillon.

Pour cette raison, l'erreur de généralisation ne doit pas être estimée à partir des données ayant servies à calculer  $\hat{f}$ , mais sur 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 l'erreur de généralisation 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. Généralement, on choisira un échantillon d'apprentissage deux fois plus grand que l'échantillon test. Il pourra aussi intégrer une stratification sur la variable cible (cf Cours précédent (<https://par.moodle.lecnam.net/mod/resource/view.php?id=96064>)).

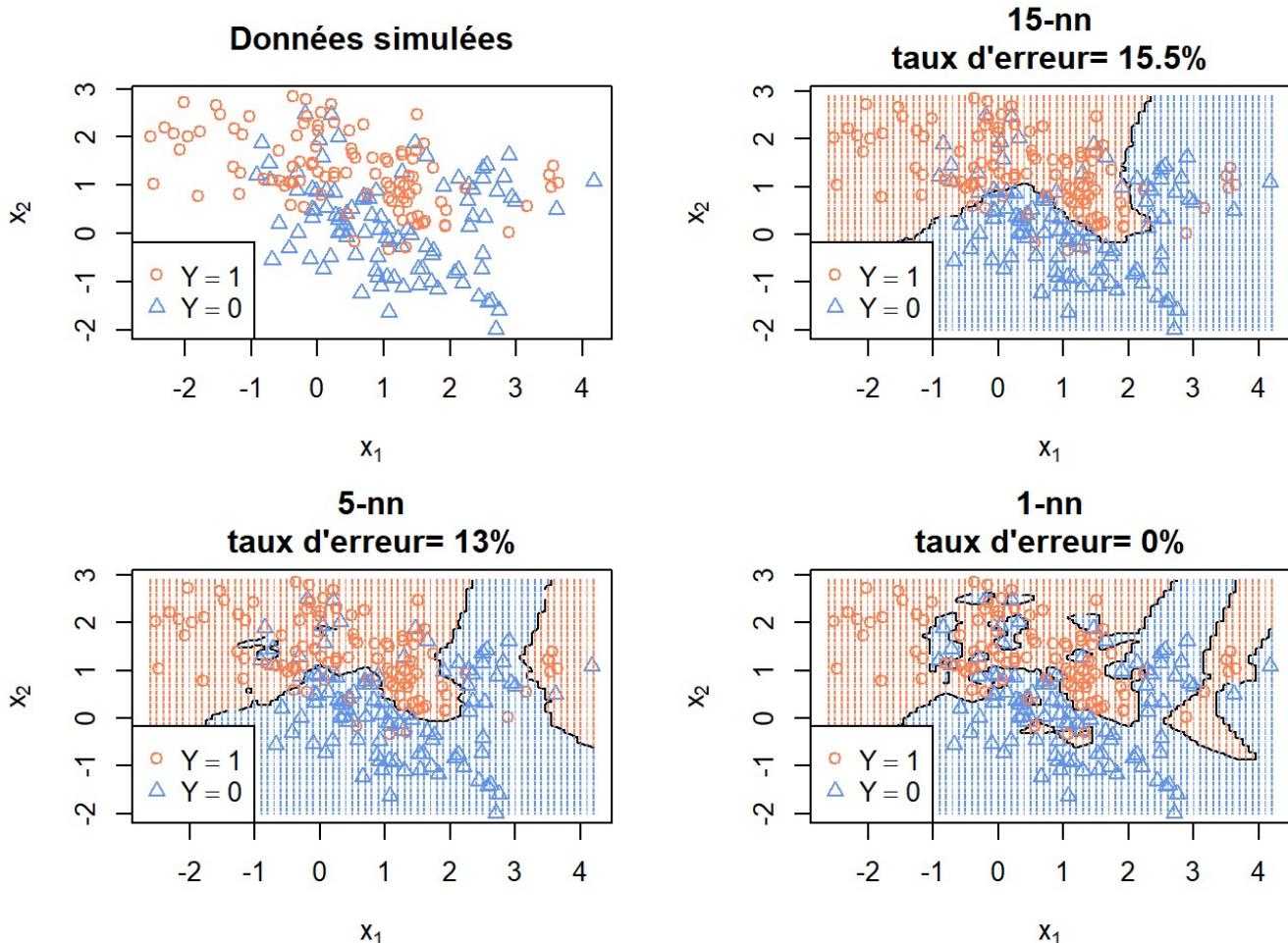


Figure 2.1: Influence de  $k$  sur la classification effectuée par  $k$ -plus proches voisins : données d'origine, classification pour  $k=15$ ,  $k=5$  et  $k=1$

## 2.2.1 Remarque pour le cas binaire

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 risques 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  $\hat{p}(Y = 1|X) > 0.5$  et 0 sinon, mais plus généralement, on peut choisir n'importe quel seuil  $s$  ( $s \in [0; 1]$ ) tel que  $\hat{p}(Y = 1|X) > s$  plutôt que le seuil 0.5. La courbe ROC représente  $1 - P(\hat{f}(X) = 0|Y = 1)$  en fonction de  $P(\hat{f}(X) = 1|Y = 0)$ , 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 2.2.

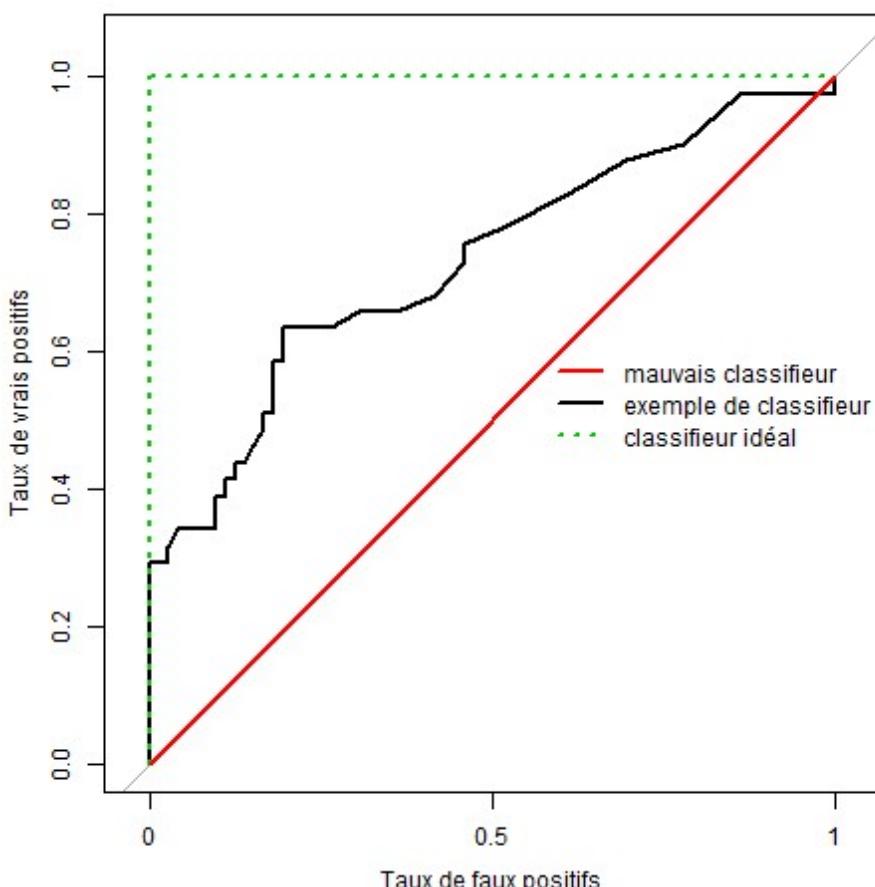


Figure 2.2: Exemple de courbe ROC

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 2.2. Les autres points de la courbe ROC (représentée en noir sur la Figure 2.2) 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  $\hat{f}(X) = 0.5$  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  $\hat{p}(Y = 1|X)$  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 2.2).

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 classificateurs 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  $\Delta$  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  $(2 \times \text{AUC}) - 1$  est une autre mesure équivalente à l'AUC souvent utilisée pour évaluer la performance d'un modèle.

# 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.

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. NB : 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 soit sur les composantes d'une analyse factorielle des données mixtes (méthode qui sera abordée ultérieurement), soit sur les composantes d'une ACM après avoir discrétisé les variables quantitatives.

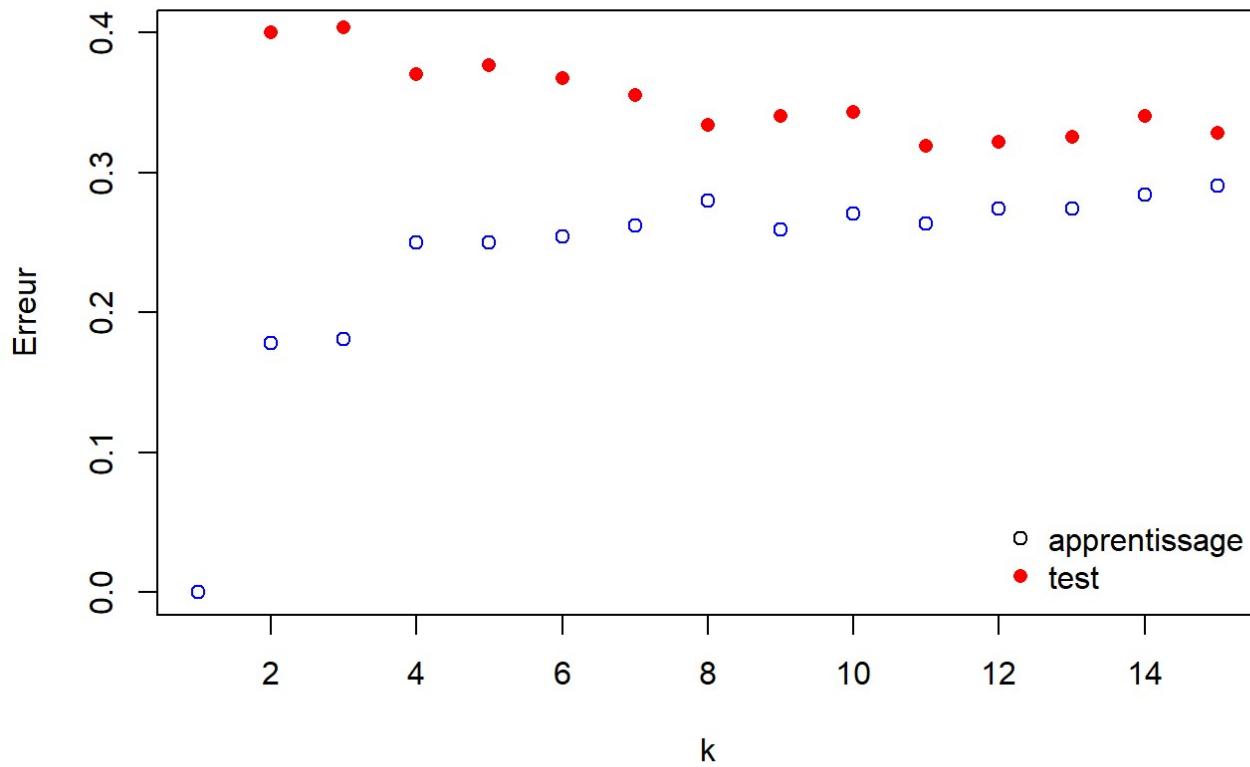


Figure 3.1: Taux de mauvais classement en fonction de la complexité du modèle

On constate que si le modèle est complexe ( $k$  petit), alors l'erreur d'apprentissage (i.e. l'erreur calculée sur l'échantillon d'apprentissage) est faible, mais l'erreur de test (i.e. l'erreur calculée 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  $k = 11$ . Au delà, le modèle est trop peu complexe et l'erreur sur l'échantillon test augmente de nouveau. Sur cet exemple, choisir  $k = 11$  correspond au modèle proposant le meilleur compromis en termes de complexité parmi les modèles considérés. Notons

que ce phénomène est également observé dans cadre de régression.

Ce compromis porte le nom de *compromis biais-variance*. Afin de justifier 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  $f$ ). Un bon modèle doit minimiser l'erreur de généralisation, en particulier, pour l'entrée  $x_0$ ,  $E \left[ (y_0 - \hat{y}_0)^2 \right]$  doit être petite. Cette erreur peut se réécrire comme suit :

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

L'erreur quadratique moyenne de prédiction peut donc être décomposée en trois termes. Le premier terme est  $\text{Bais}^2(\hat{f}(x_0))$ , 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  $\text{Var} \left[ \hat{f}(x_0) \right]$ , 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  $\sigma^2$  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 figure 3.2 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.

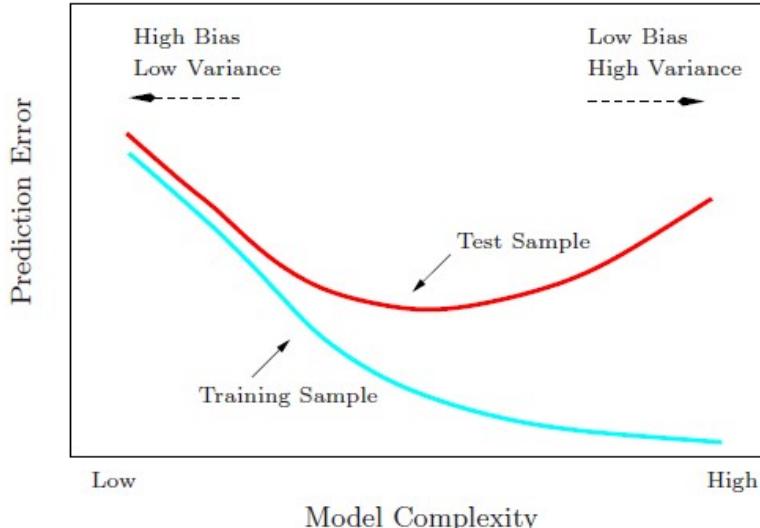


Figure 3.2: Erreur de test et d'apprentissage en fonction de la complexité du modèle (Source : Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc.)

Nous avons supposé que la variable réponse  $Y$  était quantitative, mais la recherche d'un bon compromis biais-variance est 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 un critère d'erreur basé sur l'échantillon d'apprentissage. Comme ce critère d'erreur 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 (Akaike 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  $\theta$ . Par exemple, en régression logistique,  $\theta = (\beta_0, \dots, \beta_p)$  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  $\theta$  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. On utilise en fait plutôt la log-vraisemblance, plus commode pour les calculs.

## 4.1.1 AIC

L'AIC est défini selon

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

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

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

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 il existe 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 modèle 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 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. Au contraire, 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) ou le cours ([https://youtu.be/nLpJd\\_iKmrE](https://youtu.be/nLpJd_iKmrE)) issu du MOOC de T.Hastie et R. Tibshirani.

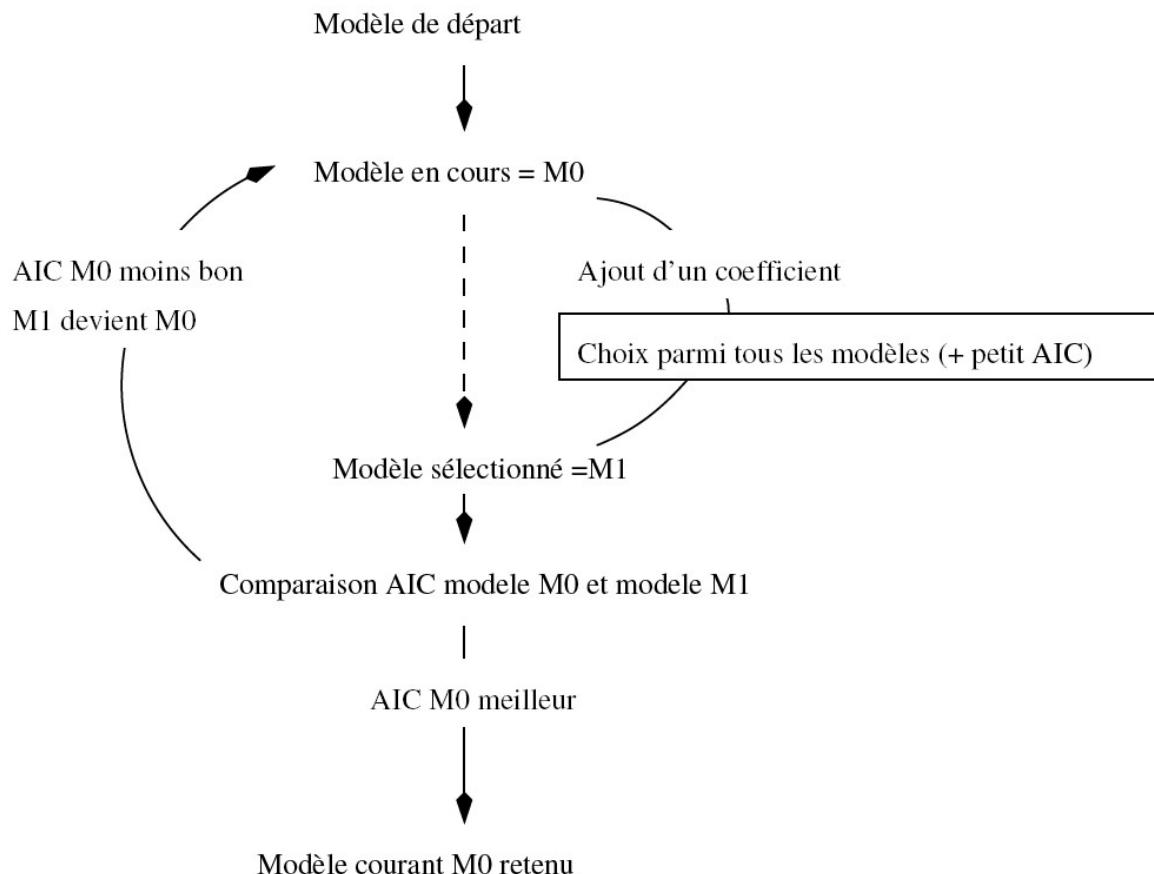


Figure 4.1: Algorithme de sélection pas à pas ascendante. Source : Cornillon et Matzner-Lober 2011

## 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 (sur laquelle nous reviendrons ultérieurement) 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 pas d'en apprécier les capacités prédictives.

## 4.2 Approches empiriques

Leur principe est de constituer un échantillon d'apprentissage pour estimer les différents modèles et un échantillon test pour n'en sélectionner qu'un seul : celui qui minimise l'erreur de généralisation.

### 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 étaient  $k$  plus proches voisins avec  $k = 1$  ou  $k = 2$ , etc. Le principe est de constituer un échantillon test et un échantillon d'apprentissage et pour chaque  $k$  on estime alors le modèle (illustré dans les graphiques de la Figure 2.1) ce qui nous amène à  $M$  modèles candidats. Plus généralement, il est possible de considérer des modèles issus de famille différentes (e.g.  $k$  plus proches voisins, régression logistique, analyse discriminante, etc). Pour choisir entre ceux-ci, on évalue l'erreur de généralisation à partir de l'échantillon test. On retient alors le modèle pour lequel l'erreur est la plus faible (comme évoqué en Section 3). Une fois ce modèle choisi on le ré-estimera en considérant l'intégralité des données à disposition.

Néanmoins, l'erreur calculée sur l'échantillon test reste une estimation, elle est donc 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.

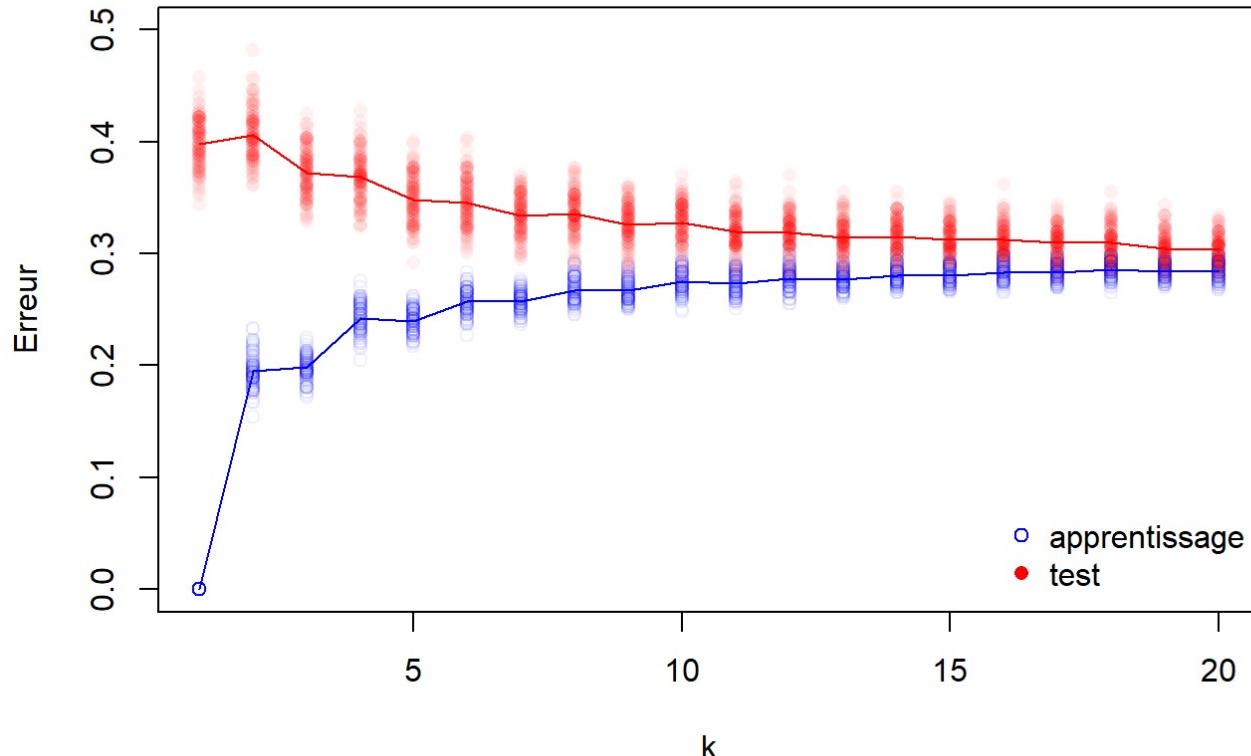


Figure 4.2: Taux de mauvais classement en fonction de la complexité du modèle pour 100 découpages de l'échantillon. Les erreurs moyennes sont représentées par une ligne.

On peut alors préférer retenir le modèle qui minimise l'erreur moyenne obtenues sur les différents échantillons tests 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.

L'ensemble des  $K - 1$  premiers sert alors d'échantillon d'apprentissage et le  $K$ ème sert d'échantillon test. On obtient ainsi un premier taux d'erreur. 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 de l'erreur de test 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 l'erreur de généralisation

Le découpage en échantillon en deux parties (apprentissage - test) permet de choisir un modèle mais il ne permet pas d'avoir une mesure objective de l'erreur de généralisation commise par le 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 l'erreur de généralisation sur un échantillon indépendant, d'où la nécessité de disposer d'un 3eme échantillon. Ainsi, on considérera généralement : un échantillon d'*apprentissage*, un échantillon de *validation* et enfin un échantillon *test*. 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 :

- on ajuste les différents modèles sur l'échantillon d'apprentissage
- on choisit le "meilleur" modèle à partir de l'erreur commise sur l'échantillon de validation
- on ajuste le meilleur modèle sur les données rassemblant échantillon d'apprentissage et de validation
- on évalue l'erreur de généralisation sur l'échantillon test
- on ré-ajuste ce modèle sur l'ensemble des 3 échantillons

## 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 recherche 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

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://web.stanford.edu/~hastie/ElemStatLearn/download.html> (<https://web.stanford.edu/~hastie/ElemStatLearn/download.html>).

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>).