Qualité de prévision et risque

Résumé

Définition et propriétés du risque ou erreur de prévision ou erreur de généralisation dans le cas de la régression et de la classification. Décomposition biais / variance du risque. Critères de pénalisation et méthodes ou algorithmes d'estimation du risque empirique. Estimation sur échantillons de validation ou de test, par critère pénalisé, par validation croisée, par bootstrap, courbe ROC en discrimination binaire.

Retour à l'introduction.

Tous les tutoriels sont disponibles sur le dépôt :
`github.com/wikistat`

1 Introduction

1.1 Objectif

La performance du modèle ou algorithme issu d'une méthode d'apprentissage s'évalue par un *risque* ou *erreur de prévision*, dite encore *capacité de généralisation* dans la communauté informatique. La mesure de cette performance est très importante puisque, d'une part, elle permet d'opérer une *sélection de modèle* dans une famille associée à la méthode d'apprentissage utilisée et, d'autre part, elle guide le *choix de la méthode* en comparant chacun des modèles optimisés à l'étape précédente. Enfin, elle fournit, tous choix faits, une mesure de la qualité ou encore de la *confiance* que l'on peut accorder à la prévision.

Une fois que la notion de modèle statistique ou *règle de prévision* est précisée, le *risque* est défini à partir d'une fonction *perte* associée. En pratique, ce risque nécessite d'être estimé et différentes stratégies sont proposées.

Le principal enjeu est de construire une estimation *sans biais* de ce risque en notant que le *risque* dit *empirique*, qui est aussi l'erreur d'ajustement du modèle, est estimé sur les mêmes données dites d'*apprentissage* du modèle ; par construction le risque empirique est *biaisé par optimisme* car toute erreur ou risque estimé sur d'autres données (capacité de généralisation) et qui n'ont pas servi à estimer le modèle ou apprendre l'algorithme, conduit, en moyenne, à des valeurs plus élevées et sans biais si ces données sont bien représentatives.

Trois stratégies sont décrites pour construire des *estimations sans biais du risque* :

1. une *pénalisation* de l'erreur d'ajustement ou risque empirique ;
2. un partage de l'échantillon : apprentissage, (validation,) test afin de distinguer l'estimation du modèle et celle du risque ;
3. par simulation : validation croisée, *bootstrap*.

Le choix dépend de plusieurs facteurs dont l'objectif recherché, la taille de l'échantillon initial, la complexité du modèle envisagé, la variance de l'erreur, la complexité des algorithmes.

1.2 Risques et choix de modèle

Une estimation du risque ou qualité de la prévision est donc un élément central de la mise en place de la stratégie de choix de modèle, choix de méthode, en science des données, telle que cette stratégie est décrite dans l'introduction.

La recherche d'un meilleur modèle, qui sera déclinée ensuite pour chaque méthode d'apprentissage, conduit à optimiser un ou des paramètres contrôlant la flexibilité ou complexité du modèle. C'est facile à illustrer dans le cas de la régression (variable à prévoir quantitative) polynomiale (figure 1). Dans ce cas élémentaire, la *complexité du modèle* est le degré du polynôme ou encore le nombre de paramètres. En augmentant ce degré, l'erreur d'ajustement décroît jusqu'à s'annuler lorsque le degré est tel que les observations sont exactement ajustées par interpolation.

Intuitivement, si les données sont bruitées, le graphique montre que des prévisions réalisées en dehors des observations peuvent conduire à des valeurs très excentrées et donc à des erreurs quadratiques moyennes très élevées.

Le point important à souligner, et qui sera constamment rappelé, est que le "meilleur" modèle en un sens prédictif n'est pas nécessairement celui qui ajuste le mieux les données, ni même le "vrai" modèle si la variance des estimations est importante. L'objectif est la recherche d'un modèle optimal *parcimonieux* dont on verra qu'en régression il réalise un meilleur *compromis biais / variance*.

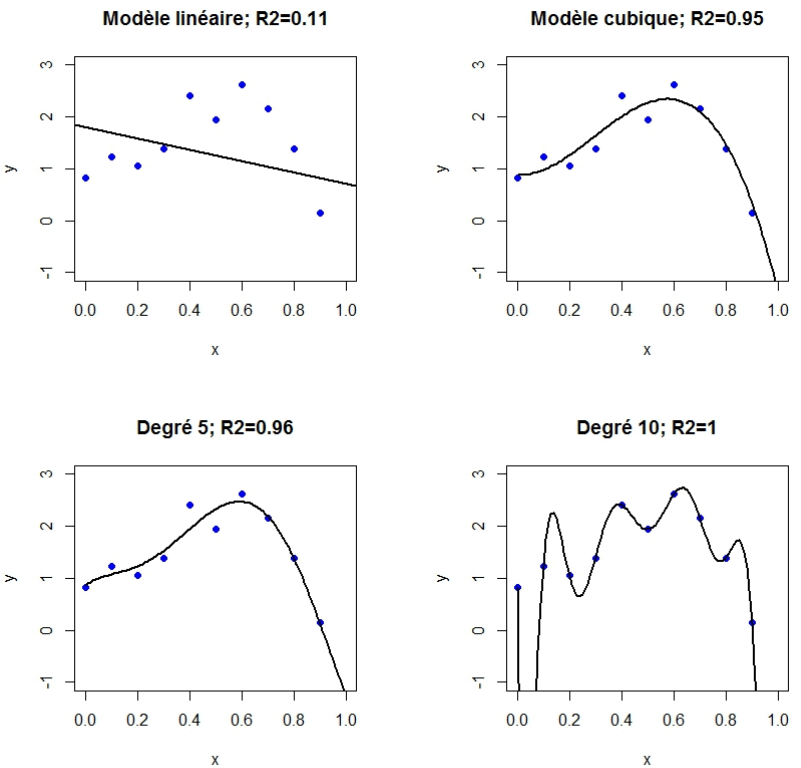

FIGURE 1 – *Régression polynomiale : ajustement par successivement des polynômes de degré 1, 2, 5 et 10*

2 Risque, risque empirique

2.1 Modèle statistique

On suppose que $\boldsymbol{D}_n$ est l'observation d'un n-échantillon

$$\boldsymbol{D}_n = \{(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)\}$$

d'une loi conjointe P sur $\mathcal{X} \times \mathcal{Y}$, *inconnue*, et que $\boldsymbol{x}$ est une observation de la variable $\boldsymbol{X}$, $(\boldsymbol{X}, Y)$ étant un couple aléatoire de loi conjointe P *indépendant* de $\boldsymbol{D}_n$.

L'échantillon $\boldsymbol{D}_n$ est appelé *échantillon d'apprentissage*.

Une *règle de prévision, régression / discrimination* ou prédicteur est une fonction (mesurable) $f : \mathcal{X} \to \mathcal{Y}$ qui associe la sortie $f(\boldsymbol{x})$ à l'entrée $\boldsymbol{x} \in \mathcal{X}$.

Pour mesurer la qualité de prévision, on introduit une fonction de perte :

DÉFINITION 1. — *Une fonction (mesurable)* $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ *est une* fonction de perte *si* $l(y, y) = 0$ *et* $l(y, y') > 0$ *pour* $y \neq y'$.

2.2 Risque

Si f est une règle de prévision, $\boldsymbol{x}$ une entrée, y la sortie qui lui est réellement associée, alors $l(y, f(\boldsymbol{x}))$ mesure une perte encourue lorsque l'on associe à $\boldsymbol{x}$ la sortie $f(\boldsymbol{x})$.

En régression réelle : on définit les pertes $\mathbb{L}^p$ $(p \geq 1)$

$$l(y, y') = |y - y'|^p.$$

Si $p = 1$ on parle de perte absolue, si $p = 2$ de perte quadratique.

En discrimination binaire : $\mathcal{Y} = \{-1, 1\}$

$$l(y, y') = \mathbb{1}_{y \neq y'} = \frac{|y - y'|}{2} = \frac{(y - y')^2}{4}.$$

On va s'intéresser au comportement moyen de cette fonction de perte, il s'agit de la notion de *risque* :

DÉFINITION 2. — *Étant donnée une fonction de perte l, le* risque – *ou l'*erreur de généralisation – *d'une règle de prévision f est défini par*

$$R_P(f) = \mathbb{E}_{(\boldsymbol{X},Y)\sim P}[l(Y, f(\boldsymbol{X}))].$$

Une estimation $\hat{f}$ de f sur un échantillon d'apprentissage $\boldsymbol{D}_n$ est obtenue par un *algorithme de prévision*

DÉFINITION 3. — *Un* algorithme de prévision *est représenté par une application (mesurable)* $\hat{f} : (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{F}$ *qui à un ensemble d'apprentissage* $\boldsymbol{D}_n = \{(\boldsymbol{x}_i, y_i), 1 \leq i \leq n\}$ *associe une règle de prévision* $\hat{f}(\boldsymbol{D}_n)$*, ou par une suite* $(\hat{f}_n)_{n\geq 1}$ *d'applications (mesurables) telles que pour* $n \geq 1$*,* $\hat{f}_n : (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{F}$

DÉFINITION 4. — *Le* risque moyen *d'un algorithme de prévision $\hat{f}$ est défini par*

$$\mathbb{E}_{\boldsymbol{D}_n\sim P^{\otimes n}}[R_P(\hat{f}(\boldsymbol{D}_n))].$$

Les travaux de Vapnik en théorie de l'apprentissage ont conduit à focaliser l'attention sur la présence ou l'absence de propriétés théoriques basiques d'une technique d'apprentissage ou de modélisation :

consistance qui garantit la capacité de généralisation. Un processus d'apprentissage est dit *consistant* si le risque empirique et un risque estimé sur un échantillon test indépendant convergent en probabilité vers la même limite lorsque les tailles de l'échantillon augmentent.

vitesse de convergence. Une évaluation, quand elle est possible, de la vitesse de convergence de l'estimation du risque lorsque la taille augmente, est une indication sur la façon dont la généralisation s'améliore et informe sur la nature des paramètres, comme le nombre de variables explicatives, dont elle dépend.

contrôle Est-il possible, à partir d'un échantillon d'apprentissage de taille fini donc sans considération asymptotique, de contrôler la capacité de généralisation et donc de majorer le terme de risque ?

Plus précisément :

DÉFINITION 5. — *Un algorithme de prévision est dit* universellement consistant *si*

$$\forall P \lim_{n\to+\infty} \left\{\mathbb{E}_{\boldsymbol{D}_n\sim P^{\otimes n}}[R_P(\hat{f}_n(\boldsymbol{D}_n))]\right\} = \inf_{f\in\mathcal{F}} R_P(f).$$

On montre ainsi, par le théorème de Stone (1977)[9], que l'algorithme dit des k plus proches voisins, qui opère par moyennage locale, est universellement consistant. Il en est de même pour les arbres de décision (CART) (Breiman et al. 1984)[2]

2.3 Minimisation du risque empirique

Définitions

Le risque d'une règle de prévision f est défini par

$$R_P(f) = \mathbb{E}_{(\boldsymbol{X},Y)\sim P}[l(Y, f(\boldsymbol{X}))]$$

qui dépend de P inconnue.

En l'absence de toute information ou hypothèse sur la loi P (cadre non paramétrique), il est naturel de remplacer P par P_n, mesure empirique associée à l'échantillon $\boldsymbol{D}_n$, et de minimiser le risque empirique.

DÉFINITION 6. — *Le* risque empirique *(associé à* $\boldsymbol{D}_n = \{(\boldsymbol{X}_i, Y_i), 1 \leq i \leq n\}$*) d'une règle de prévision* $f \in \mathcal{F}$ *est défini par*

$$\widehat{R_n}(f, \boldsymbol{D}_n) = \frac{1}{n}\sum_{i=1}^{n} l(Y_i, f(\boldsymbol{X}_i)).$$

La minimisation du risque empirique, qui est une extension de la procédure d'estimation d'un modèle, par exemple par les moindres carrés, a été développée par Vapnik (1999)[11].

DÉFINITION 7. — *Étant donné un sous-ensemble F de $\mathcal{F}$ (un* modèle*), l'*algorithme de minimisation du risque empirique sur F *est défini par :*

$$\hat{f}_F(\boldsymbol{D}_n) \in argmin_{f\in F}\widehat{R_n}(f, \boldsymbol{D}_n).$$

Le choix d'un modèle $f \in F$ adéquat est crucial et relève des méthodes de *sélection de modèles*.

Décomposition approximation/estimation (ou biais/variance)

Soit f^* telle que $R_P(f^*) = \inf_f R_P(f)$, f^* est appelé "*oracle*". L'objectif est de déterminer un modèle F pour lequel le risque de l'estimateur $\hat{f}_F(\boldsymbol{D}_n)$ est proche de celui de l'oracle.

$$R_P(\hat{f}_F(\boldsymbol{D}_n)) - R_P(f^*) = \underbrace{\left\{ R_P(\hat{f}_F(\boldsymbol{D}_n)) - \inf_{f \in F} R_P(f) \right\}}_{\text{Erreur d'estimation}} + \underbrace{\left\{ \inf_{f \in F} R_P(f) - R_P(f^*) \right\}}_{\text{Erreur d'approximation}}$$

Erreur d'estimation (Variance) ↗ — **Erreur d'approximation** (Biais) ↘ (taille de F)

Ces deux termes sont de natures différentes. Pour les évaluer, nous aurons recours à des considérations issues respectivement de la statistique et de la théorie de l'approximation.

La sélection d'un modèle $\hat{F}$ parmi une collection de modèles $\mathcal{C}$ pour lequel le risque de $\hat{f}_{\hat{F}}(\boldsymbol{D}_n)$ est proche de celui de l'oracle va s'obtenir par la minimisation d'un critère pénalisé du type :

$$\hat{F} = \text{argmin}_{F \in \mathcal{C}} \{ \hat{R}_n(\hat{f}_F(\boldsymbol{D}_n), \boldsymbol{D}_n) + pen(F) \}.$$

La pénalité permet de pénaliser les modèles de "grande" taille, afin d'éviter le sur-ajustement. Le choix optimal de la pénalité (selon les modèles statistiques considérés) est un sujet de recherche très actif en statistique.

Très généralement, plus un modèle (la famille des fonctions admissibles) est complexe, plus il est flexible et peut s'ajuster aux données observées et donc plus le biais est réduit. En revanche, la partie variance augmente avec le nombre de paramètres à estimer et donc avec cette complexité. L'enjeu, pour minimiser le risque quadratique ainsi défini, est donc de rechercher un meilleur compromis entre biais et variance : accepter de biaiser l'estimation comme par exemple en régression *ridge* pour réduire plus favorablement la variance.

3 Estimations d'un risque

Le *risque empirique* défini ci-dessus exprime la qualité d'ajustement du modèle sur l'échantillon observé. C'est ce critère, moindres carrés en régression, taux d'erreur en discrimination, qui est généralement minimisé lors de l'estimation des paramètres d'un modèle. Il constitue également une mesure de la qualité, mais biaisée, car associée à une estimation par principe trop *optimiste*, de l'erreur de prévision. Celle-ci est liée aux données qui ont servi à l'ajustement du modèle et est d'autant plus faible que le modèle est complexe ; sélectionner la complexité d'un modèle en minimisant le risque empirique conduit au sur-apprentissage ou sur-ajustement. Cette estimation ne dépend que de la partie "biais" de l'erreur de prévision et ne prend pas en compte la partie "variance" de la décomposition.

L'estimation du risque empirique est obtenue par :

$$\widehat{R_n}(\widehat{f}(D_n), D_n) = \frac{1}{n} \sum_{i=1}^{n} l(y_i, \widehat{f}(D_n)(\boldsymbol{x}_i)).$$

où l désigne une fonction perte adaptée au problème et D_n un échantillon de taille n.

Comme le risque empirique, le R^2 (coefficient de détermination de la régression) ne peut-être un "bon" critère de sélection de modèles ; il ne peut servir qu'à comparer des modèles de même dimension ou complexité car sinon conduit à sélectionner le modèle le plus complexe, c'est-à-dire celui correspond au plus grand espace de projection, ce qui entraîne un sur-ajustement. C'est simplement illustré dans le cas de la régression polynomiale. La figures 1 représente un jeu de données $Y_i = f(x_i) + \varepsilon_i$, $i = 1, \ldots, n$ et $x_i \in [0, 1]$ ajusté par des polynômes de degrés croissants. Le critère R^2 augmente pour atteindre la valeur 1 pour le polynôme qui interpole toutes les observations. L'ajustement du modèle mesuré par la R^2 croît logiquement avec le nombre de paramètres.

L'estimation d'une erreur de prévision fait appel à différentes stratégies pour contrôler l'optimisme ou réduire le biais.

3.1 Estimation avec échantillons indépendants

La façon la plus simple d'estimer *sans biais* l'erreur de prévision ou un risque consiste à utiliser un échantillon indépendant n'ayant pas participé à

l'estimation du modèle. Ceci nécessite donc d'éclater aléatoirement l'échantillon en trois parties respectivement appelées *apprentissage*, *validation* et *test* :

$$D_n = D_{n_1}^{\text{Appr}} \cup D_{n_2}^{\text{Valid}} \cup D_{n_3}^{\text{Test}} \quad avec \quad avecn_1 + n_2 + n_3 = n.$$

1. $\widehat{R_n}(\widehat{f}(D_{n_1}^{\text{Appr}}), D_{n_1}^{\text{Appr}})$ est minimisé pour déterminer l'estimateur ou l'algorithme de prévision : $\widehat{f}(D_{n_1}^{\text{Appr}})$, pour modèle fixé ; par exemple un modèle de régression polynomiale de degré fixé ;
2. $\widehat{R_n}(\widehat{f}(D_{n_1}^{\text{Appr}}), D_{n_2}^{\text{Valid}})$ sert à la comparaison des modèles au sein d'une même famille afin de sélectionner celui qui minimise cette erreur ; par exemple une famille de modèles polynomiaux de degrés variés.
3. $\widehat{R_n}(\widehat{f}, D_{n_3}^{\text{Test}})$ est utilisée pour comparer entre eux les meilleurs modèles de chacune des méthodes considérées ; par exemple le meilleur modèle polynomial au meilleur réseau de neurones.

Cette solution n'est acceptable que si la taille de l'échantillon initiale est importante sinon la qualité d'ajustement est dégradée car n_1 est trop faible et la variance de l'estimation de l'erreur peut être importante (n_2, n_3 petits).

3.2 Estimation par pénalisation

La notion de sélection de modèle avec l'objectif d'une prévision est ancienne en Statistique et a donné lieu à de nombreux développements et critères bien adaptés à de petits échantillons. Sommairement, ceux-ci se présentent comme l'ajout d'une pénalisation de la fonction objectif du problème d'optimisation : moindres carrés ou risque empirique, vraisemblance.

C_p de Mallows

Le C_p de Mallows (1973)[6] fut, historiquement, le premier critère visant à une meilleure estimation de l'erreur de prévision que la seule considération de l'erreur d'ajustement (ou le R^2) dans le modèle linéaire. Il repose sur une mesure de la qualité sur la base d'un risque quadratique. L'erreur de prévision se décompose en :

$$\widehat{R_P}(\widehat{f}(D_n)) = \widehat{R_n}(\widehat{f}(D_n), D_n) + \text{Optim}$$

qui est le risque empirique plus une estimation du biais par abus d'optimisme. Il s'agit donc d'estimer cet optimisme pour apporter une correction et ainsi une meilleure estimation de l'erreur recherchée. Cette correction peut prendre plusieurs formes. Elle est liée à l'estimation de la variance dans la décomposition en biais et variance du risque ou c'est encore une pénalisation associée à la complexité du modèle.

Son expression est détaillée dans le cas de la régression linéaire avec une fonction perte quadratique. On montre (cf. Hastie et al. 2009)[5], à des fins de comparaison qu'il peut aussi se mettre sous une forme équivalente :

$$C_p = \widehat{R_n}(\widehat{f}(D_n), D_n) + 2\frac{d}{n}\widehat{\sigma}^2$$

où d est le nombre de paramètres du modèles (nombre de variables plus un), n le nombre d'observations, $\widehat{\sigma}^2$ une estimation de la variance de l'erreur par un modèle de faible biais. Ce dernier point est fondamental pour la qualité du critère, il revient à supposer que le modèle complet (avec toutes les variables) est le "vrai" modèle ou tout du moins un modèle peu biaisé afin de conduire à une bonne estimation de σ^2.

La figure 2 montre le comportement du C_p dans l'exemple trivial de la régression polynomiale. Ce critère décroît avec le biais jusqu'à un choix optimal de dimension 3 (figure 2) avant d'augmenter à nouveau approximativement linéairement avec la variance.

AIC, AIC_c, BIC

Contrairement au C_p associé à une fonction perte quadratique, le critère d'information d'Akaïke (1974)[1] (AIC) découle d'une expression de la qualité du modèle basée sur la dissemblance de Kullback. Il se présente sous une forme similaire mais plus générale que le C_p de Mallows. Il s'applique en effet à tout modèle estimé par maximisation d'une log-vraisemblance L et suppose que la famille de densités considérée pour modéliser la loi de Y contient la "vraie" densité de Y.

Après quelques développements incluant de nombreuses approximations (estimation de paramètres par maximum de vraisemblance, propriétés asymptotiques, formule de Taylor), le critère d'Akaïke se met sous la forme :

$$\text{AIC} = -2L + 2\frac{d}{n}.$$

Dans le cas gaussien en supposant la variance connue, moindres carrés et dé-

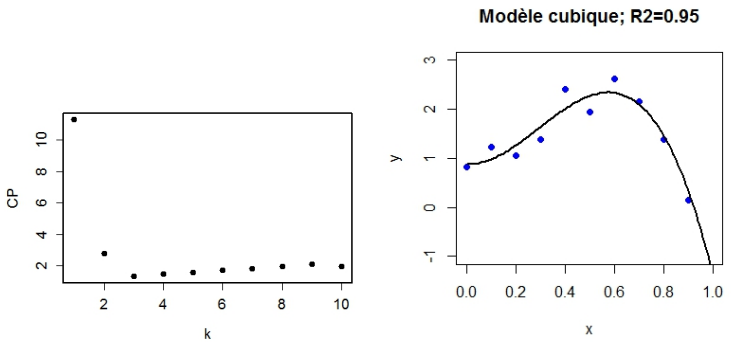

FIGURE 2 – *Régression polynomiale : C_p de Mallows en fonction du degré du polynôme et modèle de degré 3 sélectionné.*

viance coïncident, AIC est équivalent au C_p. Ce critère possède une version plus raffinée (AIC_c) dans le cas gaussien et plus particulièrement adaptée aux petits échantillons et asymptotiquement équivalente lorsque n est grand.

$$\text{AIC} = -2\mathcal{L} + \frac{n+d}{n-d-2}.$$

Une argumentation de type bayésien conduit à un autre critère BIC (*Bayesian Information Criterion* (Schwartz ; 1978)[8] qui cherche, approximativement (asymptotiquement), le modèle associé à la plus grande probabilité *a posteriori*. Dans le cas d'un modèle issu de la maximisation d'une log-vraisemblance, il se met sous la forme :

$$\text{BIC} = -2L + \log(n)\frac{d}{n}.$$

On montre, dans le cas gaussien et en supposant la variance connue que BIC est proportionnel à AIC avec le facteur 2 remplacé par $\log n$. Ainsi, dès que $n > e^2 \approx 7.4$, BIC tend à pénaliser plus lourdement les modèles complexes. Asymptotiquement, on montre que la probabilité pour BIC de choisir le bon modèle tend vers 1 lorsque n tend vers l'infini. Ce n'est pas le cas d'AIC ni du C_p qui tendent alors à choisir des modèles trop complexes. Néanmoins à taille finie, petite, BIC risque de se limiter à des modèles trop simples.

Quelque soit le critère adopté, il suffit de choisir le modèle présentant le plus faible AIC, AIC_c ou BIC parmi ceux considérés.

3.3 Estimation par simulation

Plusieurs stratégies sont proposées pour estimer des erreurs de prévision sans biais ou de biais réduit en évitant d'utiliser les mêmes données pour estimer un modèle et une erreur. Les plus utilisées sont deux versions de validation croisée et le *bootstrap*. Ces stratégies diffèrent essentiellement par le choix de procédure permettant de séparer itérativement l'échantillon initial en parties *apprentissage* et *validation*. L'estimation du risque ou de l'erreur de prévision est itérée puis toutes les estimations moyennées avant d'en calculer la moyenne pour réduire la variance et améliorer la précision lorsque la taille de l'échantillon initial est trop réduite pour en extraire des échantillons de validation et test indépendants de taille suffisante.

Validation croisée en V segments

La *V-fold cross validation* ou validation croisée en V segments est décrite dans l'algorithme ci-dessous. Il consiste à partager aléatoirement l'échantillon en V segments puis, itérativement à faire jouer à chacun de ces segments le rôle d'échantillon de validation tandis que les $V-1$ autres constituent l'échantillon d'apprentissage servant à estimer le modèle.

Algorithm 1 Validation croisée en V segments

Découper aléatoirement l'échantillon en V segments (V-fold) de tailles approximativement égales selon une loi uniforme ;
for k=1 à V **do**
  Mettre de côté le segment k,
  Estimer le modèle sur les $V-1$ segments restants,
  Calculer la moyennes des erreurs sur chacune des observations du segment k
end for
Moyenner les k erreurs pour aboutir à l'estimation par validation croisée.

Plus précisément, soit $\tau : \{1, \ldots, n\} \mapsto \{1, \ldots, V\}$ la fonction d'indexation qui, pour chaque observation, donne l'attribution uniformément aléatoire de sa classe. L'estimation par *validation croisée* de l'erreur de prévision est :

$$\widehat{R_{\text{CV}}} = \frac{1}{n} \sum_{i=1}^{n} l(y_i, \widehat{f}^{(-\tau(i))}(\boldsymbol{x}_i))$$

où $\widehat{f}^{(-v)}$ désigne l'estimation de f sans prendre en compte la k-ième partie de l'échantillon.

Le choix de V entre 5 et 15, est couramment $V = 10$ par défaut dans les logiciels. Historiquement, la validation croisée a été introduite avec $V = n$ (*leave-one-out or "loo" cross validation* ou PRESS de Allen (1974)) en régression linéaire. Ce dernier choix n'est possible que pour n relativement petit à cause du volume des calculs nécessaires. D'autre part, l'estimation de l'erreur présente alors une variance importante car comme chaque couple de modèle partagent $(n-2)$ observations, ceux-ci peuvent être très similaires donc très dépendants ; cette dépendance limite la réduction de variance issue du calcul de la moyenne. En revanche, si V est petit (*i.e.* $V = 5$), la variance sera plus faible mais le biais (pessimiste) devient un problème dépendant de la façon dont la qualité de l'estimation se dégrade avec la taille de l'échantillon. L'optimisation de V qui correspond donc encore à un meilleur équilibre entre biais et variance, nécessite généralement trop d'observations pour être pratiquée ; d'où le choix par défaut.

Minimiser l'erreur estimée par validation croisée est une approche largement utilisée pour optimiser le choix d'un modèle au sein d'une famille paramétrée. $\widehat{f}$ est défini par

$$\widehat{\theta} = \arg\min_{\theta} \widehat{R_{\text{CV}}}(\theta)$$

.

Validation croisée Monte Carlo

Le deuxième principe de validation croisée consiste à itérer plusieurs fois la division de l'échantillon initial en une partie validation ou plutôt *test* et une partie *apprentissage*.

Cette stratégie de validation croisée est généralement couplée avec celle en V segments dans l'algorithme ci-dessous.

Algorithm 2 Validation croisée Monte Carlo

for k=1 à B **do**
Séparer aléatoirement l'échantillon en deux parties : *test* et *apprentissage* selon une proportion à déterminer,
for méthode *in* liste de méthodes **do**
Estimer le modèle de la méthode en cours sur l'échantillon d'apprentissage
Optimiser les paramètres du modèle (complexité) par *Validation Croisée* à V segments.
Calculer l'erreur sur la partie test de l'échantillon pour la méthode courante
end for
end for
Calculer pour chaque méthode la moyenne des B erreurs et tracer les graphes des distributions de ces erreurs (diagramme boîte).

La proportion entre échantillons : apprentissage et test ; dépend de la taille initial de l'échantillon afin de préserver une part suffisante à la qualité de l'apprentissage ; le nombre d'itérations dépend des moyens de calcul. Plus l'échantillon initial est réduit et moins les évaluations des erreurs sont "indépendantes" et donc la réduction de variance obtenue à l'issue de la moyenne.

Bootstrap

L'idée, d'approcher par simulation (*Monte Carlo*) la distribution d'un estimateur lorsque l'on ne connaît pas la loi de l'échantillon ou, plus souvent, lorsque l'on ne peut pas supposer qu'elle est gaussienne, est l'objectif même du *bootstrap* (Efron, 1982)[3].

Le principe fondamental de cette technique de ré-échantillonnage est de substituer, à la distribution de probabilité inconnue F, dont est issu l'échantillon d'apprentissage, la distribution empirique F_n qui donne un poids $1/n$ à chaque réalisation. Ainsi on obtient un échantillon de taille n dit *échantillon bootstrap* selon la distribution empirique F_n par n tirages aléatoires avec remise parmi les n observations initiales.

Il est facile de construire un grand nombre d'échantillons bootstrap (*e.g.*

$B = 100$) sur lesquels calculer l'estimateur concerné. La loi simulée de cet estimateur est une approximation asymptotiquement convergente sous des hypothèses raisonnables [1] de la loi de l'estimateur. Cette approximation fournit ainsi des estimations du biais, de la variance, donc d'un risque quadratique, et même des intervalles de confiance (avec B beaucoup plus grand) de l'estimateur sans hypothèse (normalité) sur la vraie loi. Les grands principes de cette approche sont rappelés dans l'annexe sur le bootstrap.

Estimateur naïf

Soit $\boldsymbol{z}^*$ un échantillon bootstrap (n tirages avec remise) des données tiré selon la loi empirique $\hat{F}$ associée à l'échantillon d'apprentissage $\boldsymbol{D}_n$

$$\boldsymbol{z}^* = \{(\boldsymbol{x}_1^*, y_1^*), \ldots, (\boldsymbol{x}_n^*, y_n^*)\}.$$

L'estimateur *plug-in* de l'erreur de prévision $R_P(\hat{f}(\boldsymbol{D}_n))$ est donné par :

$$\widehat{R}_n(\hat{f}_{\boldsymbol{z}^*}, \boldsymbol{D}_n) = \frac{1}{n}\sum_{i=1}^{n} l(y_i, \hat{f}_{\boldsymbol{z}^*}(\boldsymbol{x}_i))$$

où $\hat{f}_{\boldsymbol{z}^*}$ désigne l'estimation de f à partir de l'échantillon bootstrap. Il conduit à l'estimation bootstrap de l'erreur moyenne de prévision $\mathbb{E}_{\boldsymbol{D}_n \sim P^{\otimes n}}[R_P(\hat{f}(\boldsymbol{D}_n)]$ par

$$R_{\text{Boot}} = E_{\boldsymbol{Z}^* \sim \widehat{F}}[\widehat{R}_n(\hat{f}_{\boldsymbol{Z}^*}, \boldsymbol{D}_n)] = E_{\boldsymbol{Z}^* \sim \widehat{F}}\left[\frac{1}{n}\sum_{i=1}^{n} l(y_i, f_{\boldsymbol{Z}^*}(\boldsymbol{x}_i))\right].$$

Cette estimation est approchée par simulation :

$$\widehat{R_{\text{Boot}}} = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{n}\sum_{i=1}^{n} l(y_i, f_{\boldsymbol{z}^{*b}}(\boldsymbol{x}_i)).$$

L'estimation ainsi construite de l'erreur de prévision est généralement biaisée par optimisme car, au gré des simulations, les mêmes observations $(\boldsymbol{x}_i, y_i)$ apparaissent à la fois dans l'estimation du modèle et dans celle de l'erreur. D'autres approches visent à corriger ce biais.

1. Échantillon indépendant de même loi et estimateur indépendant de l'ordre des observations.

Estimateur *out-of-bag*

La première s'inspire simplement de la validation croisée. Elle considère d'une part les observations tirées dans l'échantillon bootstrap et, d'autre part, celles qui sont laissées de côté pour l'estimation du modèle mais retenue pour l'estimation de l'erreur.

$$\widehat{R_{\text{oob}}} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{B_i}\sum_{b \in K_i} l(y_i, f_{\boldsymbol{z}^{*b}}(\boldsymbol{x}_i))$$

où K_i est l'ensemble des indices b des échantillons bootstrap ne contenant pas la $i$ème observation à l'issue des B simulations et $B_i = |K_i|$ le nombre de ces échantillons ; B doit être suffisamment grand pour que toute observation n'ait pas été tirée au moins une fois ou bien les termes avec $K_i = 0$ sont supprimés.

L'estimation $\widehat{R_{\text{oob}}}$ résout le problème d'un biais optimiste auquel est confrontée $\widehat{R_{\text{Boot}}}$ mais n'échappe pas au biais introduit pas la réduction tel qu'il est signalé pour l'estimation pas validation croisée $\widehat{R_{\text{CV}}}$. C'est ce qui a conduit Efron et Tibshirani (1997)[4] a proposer des correctifs.

Estimateur .632-bootstrap

La probabilité qu'une observation soit tirée dans un échantillon bootstrap est

$$P[\boldsymbol{x}_i \in \boldsymbol{x}^{*b}] = 1 - (1 - \frac{1}{n})^n \approx 1 - \frac{1}{e} \approx 0,632.$$

Très approximativement, la dégradation de l'estimation provoquée par le bootstrap et donc la surévaluation de l'erreur sont analogues à celle de la validation croisée avec $K = 2$. À la suite d'un raisonnement trop long pour être reproduit ici, Efron et Tibshirani (1997) proposent de compenser : excès d'optimisme du taux apparent d'erreur et excès de pessimisme du bootstrap *out-of-bag*, par une combinaison :

$$\widehat{R}_{.632} = 0,368 \times \widehat{R}_n(\hat{f}(\boldsymbol{D}_n), \boldsymbol{D}_n) + 0,632 \times \widehat{R}_{\text{oob}}.$$

Remarques

- Toutes les estimations du risque empirique considérées (pénalisation, validation croisée, bootstrap) sont asymptotiquement équivalentes et il

n'est pas possible de savoir laquelle concrètement sera, à n fini, la plus précise. Une large part d'arbitraire ou d'"expérience" préside donc le choix d'une estimation plutôt qu'une autre.
- Conceptuellement, le bootstrap est plus compliqué et pratiquement encore peu utilisé. Néanmoins, cet outil joue un rôle central dans les algorithmes récents de combinaison de modèles en association avec une estimation *out-of-bag* de l'erreur. Il ne peut être négligé.
- L'estimateur .632-bootstrap pose des problèmes en situation de surajustement aussi les mêmes auteurs ont proposé un rectificatif complémentaire noté *.632+bootstrap*.

En conclusion, l'estimation d'une erreur de prévision est une opération délicate aux conséquences importantes. Il est donc nécessaire
- d'utiliser le *même estimateur* pour comparer l'efficacité de deux méthodes,
- de se montrer très prudent, en dehors de tout système d'hypothèses probabilistes, sur le caractère absolu d'une estimation des erreurs.

3.4 Discrimination et courbe ROC

Dans une situation de discrimination le seul critère de risque ,comme le taux d'erreur de classement, n'est pas toujours bien adapté surtout, par exemple, dans le cadre de classes déséquilibrées : un modèle trivial qui ne prédit jamais une classe peu représentée ne commet pas un taux d'erreur supérieur au pourcentage de cette classe. Cette situation est souvent délicate à gérer et nécessite une pondérations des observations ou encore l'introduction de coûts de mauvais classement dissymétriques afin de forcer le modèle à prendre en compte une petite classe.

Discrimination à deux classes

Dans le cas fréquent de la discrimination de deux classes, plusieurs critères sont proposés afin d'évaluer précisément une qualité de discrimination. La plupart des méthodes (*e.g* régression logistique) estiment, pour chaque individu i, un *score* ou une probabilité $\widehat{\pi}_i$ que cette individu prenne la modalité $Y = 1$ (ou succès, ou possession d'un actif, ou présence d'une maladie...). Cette probabilité ou ce score compris entre 0 et 1 est comparé avec une valeur seuil s fixée *a priori* (en général $0,5$) :

$$\text{Si} \quad \widehat{\pi}_i > s, \ \widehat{y}_i = 1 \quad \text{sinon} \quad \widehat{y}_i = 0.$$

Matrice de confusion

Pour un échantillon de taille n dont l'observation de Y est connue ainsi que les scores $\widehat{\pi}_i$ fournis par un modèle, il est alors facile de construire la matrice dite de *confusion* croisant les modalités de la variable prédite pour une valeur de seuil s avec celles de la variable observée dans une table de contingence :

Prévision	Observation		Total
	$Y=1$	$Y=0$	
$\widehat{y}_i = 1$	$n_{11}(s)$	$n_{10}(s)$	$n_{1+}(s)$
$\widehat{y}_i = 0$	$n_{01}(s)$	$n_{00}(s)$	$n_{0+}(s)$
Total	n_{+1}	n_{+0}	n

Dans les situations classiques de diagnostic médical, marketing, reconnaissance de forme, détection de signal... les principales quantités suivantes sont considérées :
- Nombre de conditions positives $P = n_{+1}$
- Nombre de conditions négatives $N = n_{0+}$
- Vrais positifs $TP = n_{11}(s)$ observations bien classées ($\widehat{y}_i = 1$ et $Y = 1$),
- Vrais négatifs $TN = n_{00}(s)$ observations bien classées ($\widehat{y}_i = 0$ et $Y = 0$),
- Faux négatifs $FN = n_{01}(s)$ observations mal classées ($\widehat{y}_i = 0$ et $Y = 1$),
- Faux positifs $FP = n_{10}(s)$ observations mal classées ($\widehat{y}_i = 1$ et $Y = 0$),
- *Accuracy* et Taux d'erreur : $ACC = \frac{TN+TP}{N+P} = 1 - \frac{FN+FP}{N+P}$,
- Taux de vrais positifs ou *sensitivity, recall* $TPR = \frac{TP}{P} = 1 - FNR$ ou taux de positifs pour les individus qui le sont effectivement,
- Taux de vrais négatifs ou *spécificity, selectivity* $TNR = \frac{TN}{N} = 1 - FPR$ ou taux de négatifs pour les individus qui le sont effectivement,
- Précision ou *positive predictive value* $PPV = \frac{TP}{TP+FP} = 1 - FDR$
- Taux de faux positifs $FPR = \frac{FP}{N} = 1 - TNR$ ou un moins la spécificité,

- Taux de faux négatifs $FNR = \frac{FN}{P} = 1 - TPR$ ou un moins la sensibilité,
- Taux de fausses découvertes $FDR = \frac{FP}{FN+TN}$,
- F_1 score ou moyenne harmonique de la précision et de la sensibilité, $F_1 = 2 \times \frac{PPV \times TPR}{PPV+TPR} = \frac{2 \times TP}{2 \times TP+FP+FN}$,
- $F_\beta (\beta \in \mathbb{R}^+)$ score, $F_\beta = (1+\beta^2) \frac{PPV \times TPR}{\beta^2 PPV+TPR}$.

Les notions de *spécificité* et de *sensibilité* proviennent de la théorie du signal ; leurs valeurs dépendent directement de celle du seuil s. En augmentant s, la sensibilité diminue tandis que la spécificité augmente car la règle de décision devient plus exigeante ; un bon modèle associe grande sensibilité et grande spécificité pour la détection d'un signal.

Le dernier critère F_β permet de pondérer entre spécificité et sensibilité en prenant en compte l'importance ou le coût des faux positifs. Plus β est petit, plus les faux positifs sont coûteux au regard des faux négatifs.

Courbe ROC et AUC

Le lien entre spécificité et sensibilité est représenté graphiquement par la courbe ROC (*Receiver Operating Caracteristic*) de la sensibilité (probabilité de détecter un vrai signal) en fonction de un moins la spécificité (probabilité de détecter un signal à tort) pour chaque valeur s du seuil. C'est donc, en bleu ci-dessus, le taux de vrais positifs (TPR) en fonction du taux de faux positifs (FPR). Notons que la courbe ROC est une fonction monotone croissante :

$$1 - \frac{n_{00}(s)}{n_{+0}} < 1 - \frac{n_{00}(s')}{n_{+0}} \Rightarrow s < s' \Rightarrow \frac{n_{11}(s)}{n_{+1}} < \frac{n_{11}(s')}{n_{+1}}.$$

Plus la courbe (figure 3) se rapproche du carré, meilleure est la discrimination, correspondant à la fois à une forte sensibilité et une grande spécificité. L'aire sous la courbe : AUC (*area under curve*) mesure la qualité globale de discrimination du modèle tandis qu'une analyse de la courbe aide au choix de la valeur du seuil s en fonction du TPR recherché ou du FPR acceptable.

L'aire sous la courbe est calculée en considérant toutes les paires (i, i') formées d'un premier individu avec $y_i = 1$ et d'un second avec $y_{i'} = 0$. Une paire est dite concordante si $\widehat{\pi}_i > \widehat{\pi}_{i'}$; discordante sinon. Le nombre d'*ex æquo* est $n_{+0}n_{+1} - n_c - n_d$ où n_c est le nombre de paires concordantes et n_d le nombre de paires discordantes. Alors,

$$\text{AUC} = \frac{n_c + 0,5(n_{+0}n_{+1} - n_c - n_d)}{n_{+0}n_{+1}}.$$

On montre par ailleurs (voir par exemple Tennenhaus (2007)[10]) que le numérateur de cette expression est encore la Statistique de test de Mann-Whitney tandis que le coefficient de Gini, qui est le double de la surface entre la diagonale et la courbe, vaut $2 \times \text{AUC} - 1$.

Pour comparer des modèles ou méthodes de complexités différentes, ces courbes doivent être estimées sur un échantillon test. Elles sont bien évidemment optimistes sur l'échantillon d'apprentissage. De plus, l'AUC ne définit pas un ordre total entre modèles car les courbes ROC peuvent se croiser.

Détection du dépassement de seuil d'ozone, optimisation par sélection

Deux modèles de régression logistique sont optimisés pour estimer la probabilité de dépasser le seuil. Le premier est linéaire tandis que le 2ème, faisant intervenir les interactions, comme dans le cas de l'ANCOVA, est quadratique. Le même procédure ascendante est utilisée par minimisation de l'AIC. Sur un échantillon test de 209 observations, avec un seuil fixé à 0.5, les taux d'erreur sont respectivement de 12.4, 8.6 et 25% pour les régressions logistiques linéaire et quadratique et enfin pour le modèle MOCAGE. Les résultats sont précisés avec les courbes ROC de la figure 3 également calculées sur l'échantillon test.

Ces résultats montrent encore plus clairement l'intérêt de l'adaptation statistique de la prévision MOCAGE mais aussi la difficulté de la décision qui découle de la courbe ROC. Le choix du seuil, et donc de la méthode à utiliser si les courbes se croisent, dépend d'un choix dans ce cas politique : quel est le taux de faux positifs acceptable d'un point de vue économique ou le taux de vrais positifs à atteindre pour des raisons de santé publique ? Le problème majeur est de quantifier les coûts afférents, par la définition d'une matrice dissymétrique de ces coûts de mauvais classement.

Autre critère pour la discrimination à deux classes

Cette situation illustre bien que le taux d'erreur de classement n'est pas toujours adapté à une situation surtout dans le cas de classes très déséquili-

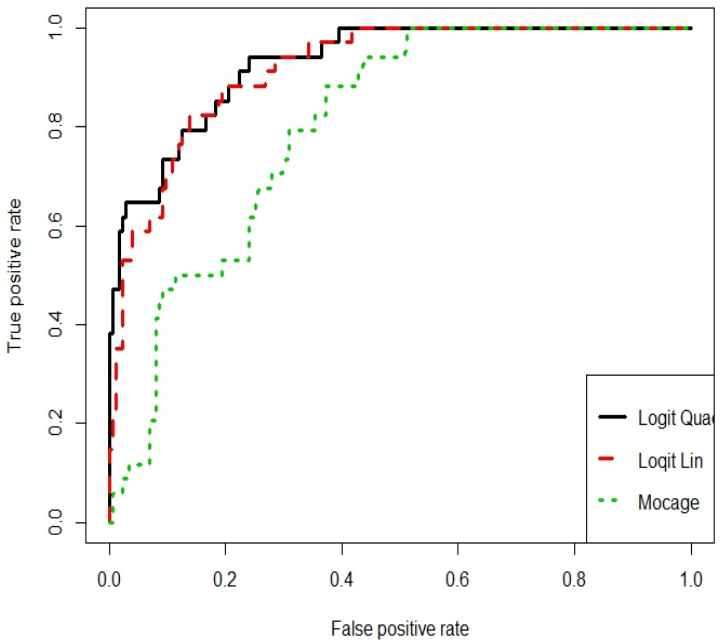

FIGURE 3 – *Ozone : Courbe ROC dans le cas de l'ANCOVA quadratique, de la régression logistique quadratique et du modèle MOCAGE.*

brées. Un modèle trivial qui ne prédit jamais une classe peu représentée ne commet pas un taux d'erreur supérieur au pourcentage de cette classe.

D'autres critères ont été proposés pour intégrer cette difficulté dont le *Score de Pierce* basés sur le taux de bonnes prévisions : $H = \frac{n_{11}(s)}{n_{+1}(s)}$ et le taux de fausses alertes : $F = \frac{n_{10}(s)}{n_{+0}}$. Le score de Pierce est alors défini par PSS$=$ $H - F$ et est compris entre -1 et 1. Il évalue la qualité de la prévision. Si ce score est supérieur à 0, le taux de bonnes prévisions est supérieur à celui des fausses alertes et plus il est proche de 1, meilleur est le modèle.

Le score de Pierce a été conçu pour la prévision d'événements climatiques rares afin de pénaliser les modèles ne prévoyant jamais ces événements ($H = 0$) ou encore générant trop de fausses alertes ($F = 1$). Le modèle idéal prévoyant tous les événements critiques ($H = 1$) sans fausse alerte ($F = 0$). Des coûts de mauvais classement peuvent être introduits pour pondérer ce score.

Log loss

Ce critère est mentionné ici car souvent utilisé comme score de référence dans les concours de prévision pour départager les concurrents. Il est défini par

$$\text{Ll} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{L} y_{ik}\log(\hat{\pi_{ik}})$$

où y_{ik} est l'indicatrice qui vaut 1 si le $i$ème individu prend la modalité k, 0 sinon, et $\hat{\pi_{ik}}$ est la probabilité estimé par le modèle que l'individu i prenne la $k$ème modalité de Y. Comme pour la définition de l'entropie, la convention $0\log(0) = 0$ est adoptée. Pour chaque observation, seule le terme de la bonne classe de Y est comptabilisé dans la somme. Ce critère est minorée par 0 si toutes les observations sont bien classés avec une probabilité 1 tandis qu'un avis neutre conduit à une valeur de $\log(L)$ alors qu'une confiance exagérée dans une mauvaise classe à pour conséquence de faire exploser ce critère qui n'est pas borné. Cette propriété incite les participants d'un concours à la prudence en pondérant les estimations des probabilités par lissage dit de Laplace ou additif (cf. Manning et al. 2008)[7]) p60).

Contrairement à la courbe ROC, le *Log loss* est un critère qui s'adapte à la mesure d'un risque lorsque la variables à modéliser est qualitative avec plus de

deux modalités. Un autre critère utilisable dans ce cas est le risque bayésien ; il est présenté dans la vignette sur l'analyse discriminante.

Références

[1] H. Akaïke, *A new look at the statistical model identification*, IEEE Transactions on Automatic Control **19** (1974).

[2] L. Breiman, J. Friedman, R. Olshen et C. Stone, *Classification and regression trees*, Wadsworth & Brooks, 1984.

[3] B. Efron, *The Jackknife, the Bootstrap and other Resampling Methods*, SIAM, 1982.

[4] B. Efron et R. Tibshirani, *Improvements on Cross-Validation : The .632+ Bootstrap Method*, Journal of the American Statistical Association **92** (1997), n° 438, 548–560.

[5] T. Hastie, R. Tibshirani et J Friedman, *The elements of statistical learning : data mining, inference, and prediction*, Springer, 2009, Second edition.

[6] C.L. Mallows, *Some Comments on Cp*, Technometrics **15** (1973), 661–675.

[7] C. Manning, P. Raghavan et H. Schütze, *Introduction to information retrieval*, Cambridge University Press, 2008.

[8] G. Schwarz, *Estimating the dimension of a model*, Annals of Statistics **6** (1978), 461–464.

[9] M. Stone, *An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion*, Journal of The Royal Statistical Society B **39** (1977), 44–47.

[10] M. Tenenhaus, *Statistique : méthodes pour décrire, expliquer et prévoir*, Dunod, 2007.

[11] V.N. Vapnik, *Statistical learning theory*, Wiley Inter science, 1999.