

# Estimation de l'erreur de prédiction

## Les techniques de ré-échantillonnage

Ricco RAKOTOMALALA

# Plan

1. Évaluation des performances
2. Erreur en resubstitution
3. Schéma apprentissage test
4. Validation croisée
5. Bootstrap
6. Influence du mode d'échantillonnage

# Évaluation des performances en apprentissage supervisé

## L'impossibilité de mesurer l'erreur de prédiction sur la population

Point de départ : On dispose d'un échantillon de taille « n » pour construire un modèle de prédiction  $M(n)$

$$\hat{Y} = M(X, n)$$

Erreur de prédiction : Comptabilisée en confrontant « vraie » valeur de  $Y$  et valeur prédite par  $M$  dans la population

$$\varepsilon = \frac{\sum_{\omega \in \Omega_{pop}} [Y(\omega) \neq \hat{Y}(\omega)]}{\text{card}(\Omega_{pop})}$$

Interprétation : Probabilité de mal classer avec le modèle de prédiction

Problème : (1) On ne dispose (presque) jamais de la population  
(2) L'accès à tous les individus serait trop coûteux



Comment s'en sortir en ne disposant en tout et pour tout que de l'échantillon « n » pour construire le modèle et en évaluer les performances...


## Évaluation des performances en apprentissage supervisé Illustration avec les « Ondes » de Breiman (1984)

### Description :

- Prédire la forme d'onde (3 valeurs possibles) à partir de 21 mesures continues
- Données simulées, donc virtuellement infinies
- $n = 500$  ind., utilisés pour construire les modèles (dataset)
- $n = 500.000$  ind., pour mesurer les performances sur la « population » (notre référence)
- 3 modèles avec des caractéristiques différentes (LDA, C4.5 et RNA-Perceptron)

Les « vrais » taux d'erreur : Mesurés sur la « population ».

	Erreur "théorique" (Calculé sur 500000 obs.)
<b>LDA</b>	0.185
<b>C4.5</b>	0.280
<b>RNA (10 CC)</b>	0.172



Comment obtenir (approcher) ces valeurs en disposant uniquement des  $n=500$  observations ?

## Erreur en resubstitution

Utiliser le « dataset » pour construire ET tester le modèle

Démarche :

- Construire le modèle sur le dataset (n= 500)
  - Ré appliquer le modèle sur ce même dataset
  - Construire la matrice de confusion et en extraire une estimation de l'erreur théorique
- On parle d'erreur en resubstitution

$$e_r = \frac{\sum_{\omega \in \Omega} [Y(\omega) \neq \hat{Y}(\omega)]}{n}$$

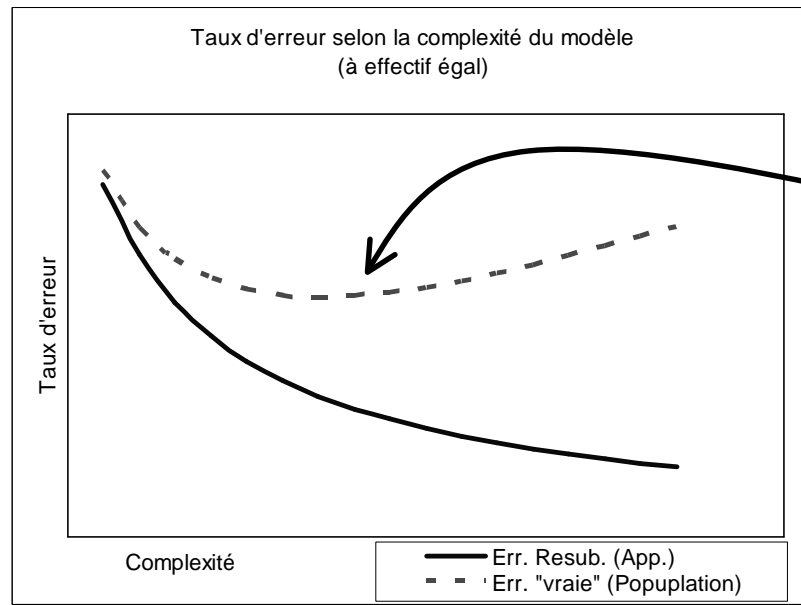
Commentaires :

- L'erreur en resubstitution est (quasiment) toujours optimiste – « Biais d'optimisme »
  - L'optimisme dépend des caractéristiques du classifieur c.-à-d. de son aptitude à « coller » aux données
  - Plus un point influe sur sa propre affectation, plus le biais d'optimisme sera élevé
- (1) RNA et un point atypique, 1-ppv : 0% d'erreur en apprentissage, etc.
- (2) Modèles à forte complexité
- (3) Cas des faibles effectifs (n petit)
- (4) Dimensionnalité élevée (surtout par rapport aux effectifs) et variables bruitées

### Résultats

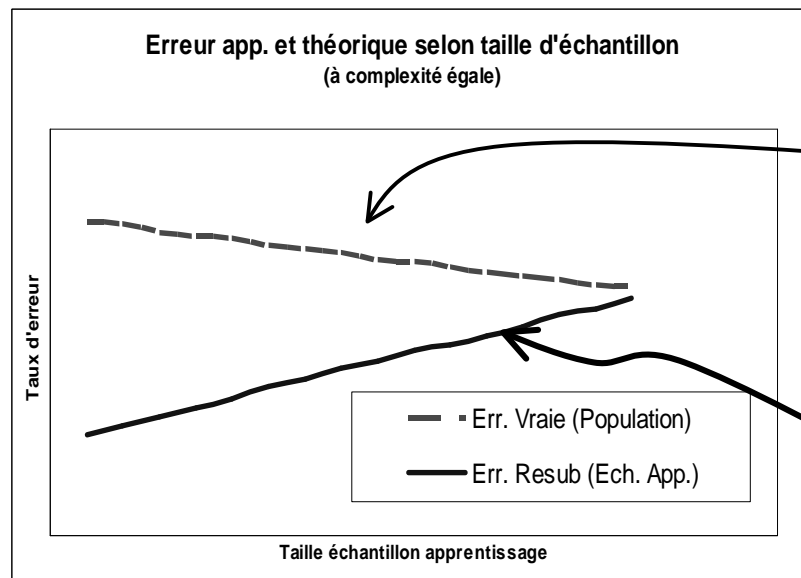
	Erreur "théorique"	Erreur Resubstitu
<b>LDA</b>	0.185	0.124
<b>C4.5</b>	0.280	0.084
<b>RNA (10 CC)</b>	0.172	0.064

## Erreur en resubstitution et « vraie » erreur Selon la complexité du modèle et selon les effectifs



On commence à « apprendre » les informations spécifiques (les scories) au fichier qui ne sont pas transposable à la population

(ex. trop de variables ; trop de neurones dans la c.c. ; arbre de décision trop profond...)

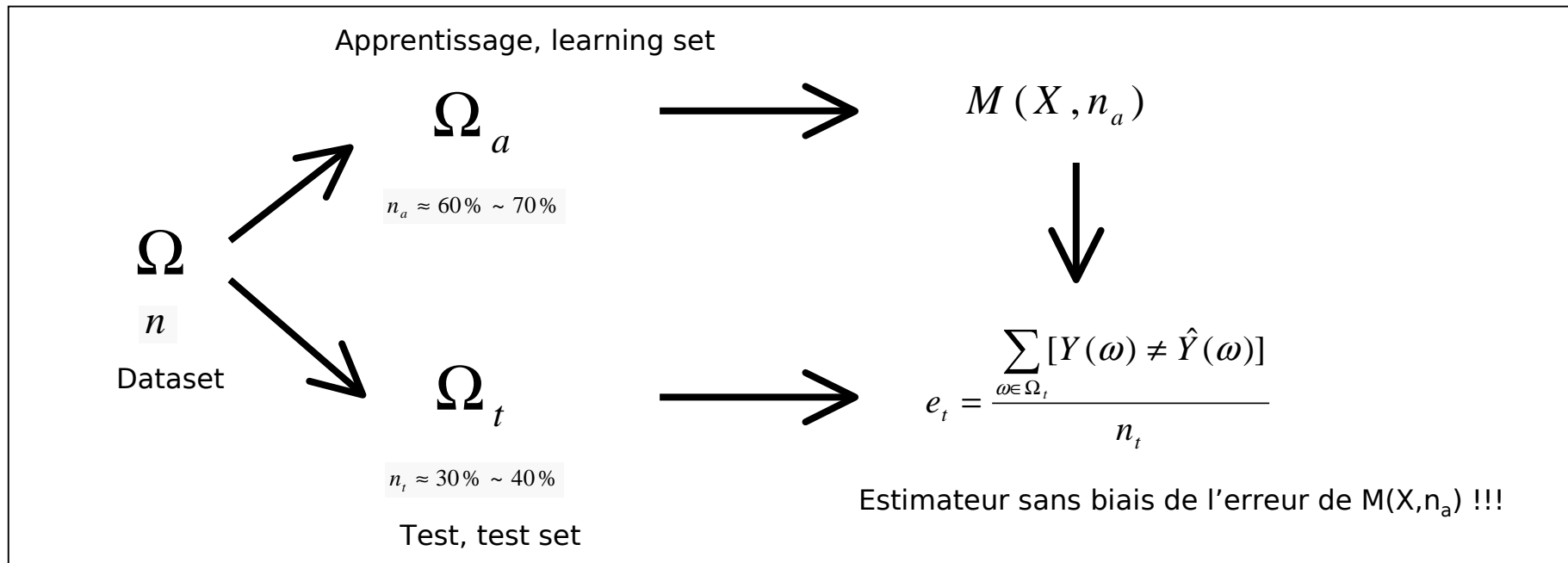


Plus la taille d'échantillon augmente, plus on apprend efficacement la « relation sous-jacente » entre Y et les X dans la population

Plus la taille d'échantillon augmente, moins on « sur apprend » les spécificités c.-à-d. (souvent) les probabilités conditionnelles  $P(Y/X)$  sont mieux estimées

# Schéma Apprentissage – Test

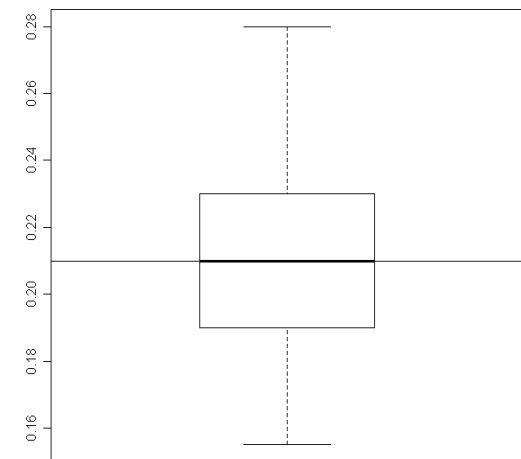
Dissocier les données pour construire et pour évaluer le modèle



Modèle : LDA(X,300)  $\rightarrow \varepsilon = 0.2099$  Calculé sur les 500.000 obs.

Test set : 200 obs.  $\rightarrow \varepsilon = 0.1950$

Expérimentation  
Répéter 100 fois le schéma  
300 ind. App., 200 ind. Test  $\rightarrow$



$e_i$  Est un estimateur sans biais de l'erreur de  $M(X, n_a)$  LDA(X,300)

**Mais** C'est un estimateur **biaisé** de l'erreur de  $M(X, n)$  LDA(X,500)

Une partie des données seulement (300 obs.) sert à construire le modèle → l'apprentissage est de moins bonne qualité (que si on utilise les 500 obs.)

**Biais**

Pourquoi ?  
Deux phénomènes s'opposent

Moins on met d'obs. en apprentissage, plus l'estimation de l'erreur sera biaisée

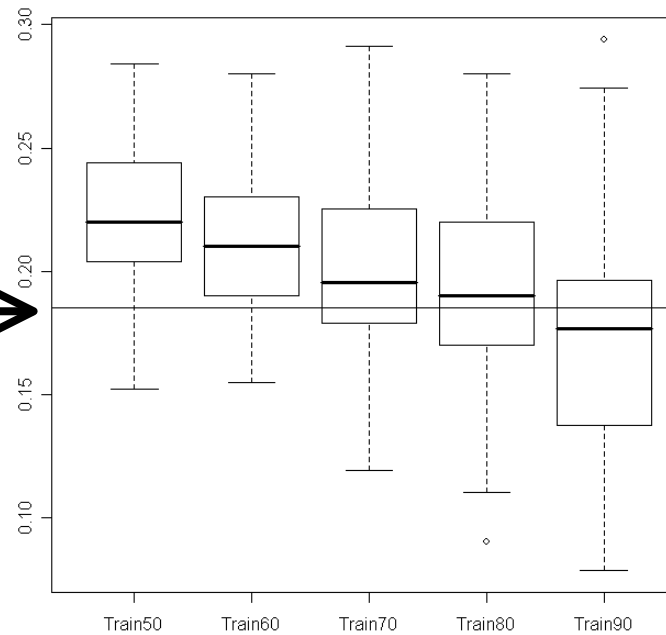
Plus on met des observations en test, plus l'estimation de l'erreur sera précise

**Variance**

## Schéma apprentissage-test Biais-variance (expérimentation)

La partie apprentissage augmente →

« Vrai » taux d'erreur de  
LDA(X,500) = 0.185



Biais fort  
Variance faible

Biais faible  
Variance forte

### Conclusion :

- L'erreur en test est un estimateur non biaisé du modèle construit sur la partie apprentissage
- C'est un mauvais estimateur de l'erreur commise par le modèle construit sur l'ensemble des données
- La subdivision « apprentissage-test » n'est intéressante que sur les bases de taille importante
- Sinon, on est face à un dilemme : pénaliser le modèle pour mieux estimer ses performances, ou favoriser la construction du modèle sans savoir ce qu'il vaut dans la population

# Comment estimer l'erreur de $M(X,n)$ ?

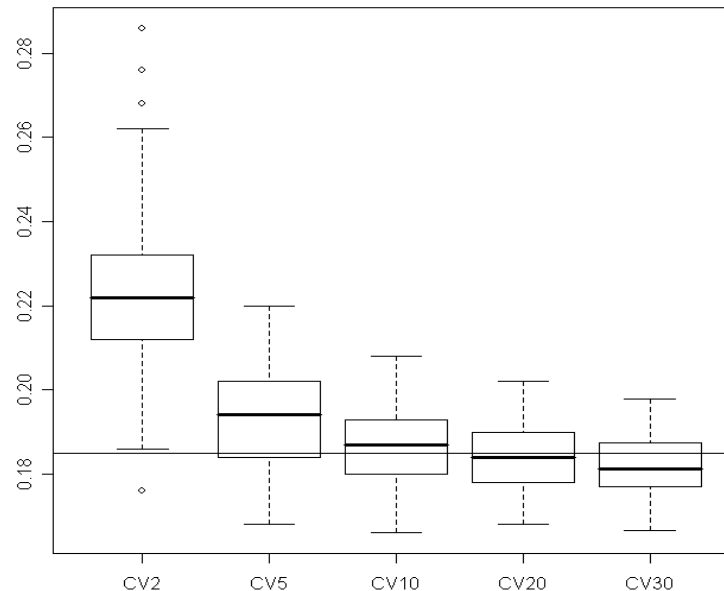
## Les méthodes de ré-échantillonnage

Validation croisée  
Leave-one-out  
Bootstrap

### Algorithme

- Subdiviser l'échantillon en K blocs
- Pour chaque k :
  - Construire sur le modèle  $M(X, n-n_k)$
  - Calculer l'erreur en test sur  $n_k \rightarrow e_k$
- Calculer la moyenne  $e_{cv}$  des erreurs en test

« Vrai » taux d'erreur  
de LDA(X,500) = 0.185



- K=10 assure un bon compromis entre « biais » et « variance » pour la majorité des cas (données et méthodes)
- Utiliser la validation répétée (B x K-Fold Cross validation) en améliore les caractéristiques

- Sur les cas de fort sur apprentissage (certaines méthodes mal paramétrées, ratio élevé de variables vs. individus, beaucoup de variables non pertinentes, etc.), la validation croisée (avec K élevé) a tendance à sous-estimer l'erreur !!!

Leave-one-out  
Validation croisée où  $K = n$

### Algorithme

- Subdiviser l'échantillon en  $K=n$  blocs
- Pour chaque individu  $k$  :
  - Construire sur le modèle  $M(X, n-1)$
  - Calculer l'erreur en test sur le  $k$ -ième ind.  $\rightarrow e_k$
- Calculer la moyenne  $e_{100}$  des erreurs en test

- Nettement plus coûteux en calcul que la  $K$  validation croisée sans être meilleur
- Sous-estimation (dramatique) de l'erreur en cas de sur apprentissage (ex. un classifieur 1-PPV dans un espace sur dimensionné)

Simulation  
(sur l'échantillon de taille 500)

	Erreur "théorique" (Calculé sur 500000 obs.)	10-CV	Leave one out
LDA	0.185	0.170	0.174
C4.5	0.280	0.298	0.264
RNA (10 CC)	0.172	0.174	0.198

On peut « stabiliser » en  
répétant la procédure

Un seul essai possible sur  
un échantillon de taille  $n$

### Algorithme

- Répéter B fois (on parle de réplifications)
  - Tirage avec remise d'un échantillon de taille n  $\rightarrow \Omega_b$
  - Distinguer les individus non échantillonnés  $\rightarrow \Omega_{(b)}$
  - Apprentissage du modèle sur  $\Omega_b$
  - Erreur en resubstitution sur  $\Omega_b$
  - Erreur en test sur  $\Omega_{(b)}$
  - Calcul de l'optimisme  $o_b$

Sur l'échantillon complet, calculer l'erreur en resubstitution

(1) 
$$e_B = e_r + \frac{\sum o_b}{B}$$
 C'est l'optimisme qui est estimé  
Il est utilisé pour corriger l'erreur en resubstitution

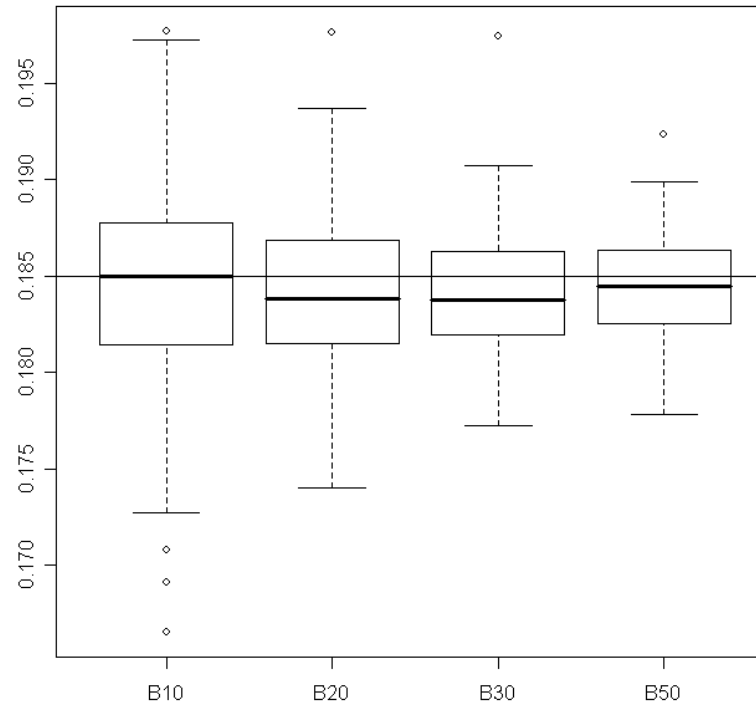
(2) 
$$e_{0.632 B} = 0.368 \times e_r + 0.632 \times \frac{\sum o_b}{B}$$
 0.632 bootstrap  
 $\rightarrow$  Pondérer par la probabilité qu'un individu fasse partie de  $\Omega_b$  sur une réplification (#0.632)

(3) Il existe une troisième formule pour corriger le biais induit par la méthode d'apprentissage : 0.632+ bootstrap

Augmentation du nombre de réplifications



« Vrai » taux d'erreur  
LDA = 0.185



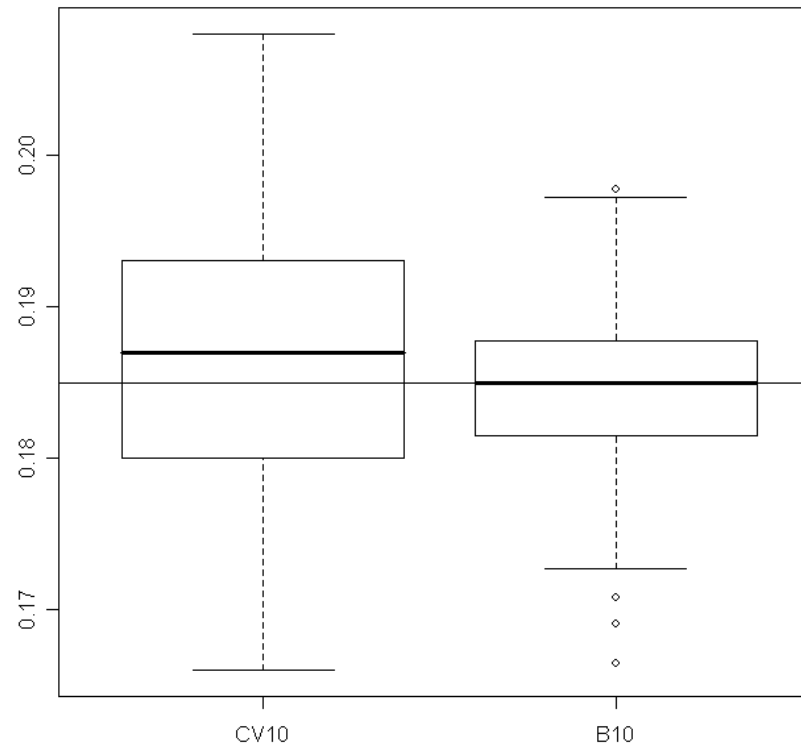
Réduction de la variance – B # 25 suffit généralement  
Peu d'effet sur le biais

Le biais vient du fait qu'à chaque réplification, on utilise bien n obs. pour construire le modèle, mais comme certains individus se répètent, l'information est redondante, la construction du modèle est pénalisée... nous n'avons pas pris dessus...

## Validation croisée ou bootstrap ?

A effort de calcul égal (ici 10 séquences apprentissage – test)

- Bootstrap a une variance plus faible (c'est flagrant ici)
- Mais la validation croisée est moins biaisée (nous avons un contre-exemple ici, ce n'est qu'une simulation à 100 essais)



# *Schéma d'échantillonnage*

Et estimation de l'erreur à l'aide de la validation croisée

Stratification, effet grappe, etc.

Principe général : le mode de constitution des blocs doit respecter le mode de constitution de l'échantillon de données

→ Si l'échantillon est stratifié, les blocs doivent être stratifiés de la même manière (notamment, on essaie de respecter la distribution de la variable à prédire dans chaque bloc)

Objectif : réduire la variance.  
L'amélioration est faible dans la pratique...

→ Si l'échantillon est formé par un tirage par grappes, les blocs doivent être formés à partir de tirages sur les grappes

Attention : Dans le cas contraire, on sous-estime fortement l'erreur, la V.C. est biaisée...

## Conclusion

- Erreur en resubstitution est (presque) toujours trop optimiste
- Son optimisme dépend des caractéristiques des données et du classifieur → de l'aptitude de ce dernier à « coller aux données »
  
- Le découpage apprentissage-test n'est vraiment intéressant que sur les « grandes » bases
- L'erreur en test estime l'erreur du modèle construit sur la partie apprentissage
- Il n'indique « rien » (ou très mal) sur les performances du modèle construit sur la totalité des données
  
- K-validation croisée et leave-one-out proposent des performances équivalentes en général
- K = 10 semble un bon compromis pour la validation croisée
- La 10-validation croisée répétée améliore la précision
  
- 0.632 bootstrap a en général une variance plus faible que la validation croisée, mais un biais plus fort (et on ne peut pas réellement agir dessus)
- Bootstrap et validation croisée se valent, surtout lorsque la taille de l'échantillon augmente
  
- En situation de fort sur apprentissage, validation croisée et bootstrap donnent de mauvaises indications

## Références

D. Zighed, R. Rakotomalala, « Graphes d'induction », Hermès, 2000.  
Chapitre 11, pages 237-261, « Évaluation empirique des classifieurs »  
Ou « Graphes d'Induction », R. Rakotomalala, 1997, Chapitre 2, pages 23-41  
[http://eric.univ-lyon2.fr/~ricco/doc/Graphes\\_Induction\\_These\\_Rakotomalala\\_1997.pdf](http://eric.univ-lyon2.fr/~ricco/doc/Graphes_Induction_These_Rakotomalala_1997.pdf)

J-C. Turlot, « Évaluation de la qualité d'une règle de décision et sélection de variables »,  
in G. Celeux, J-P. Nakache, Eds, « Analyse discriminante sur variables qualitatives »,  
Polytechnica, Chapitre 6, pages 147-180, 1994.

A. Molinaro, R. Simon, R. Pfeiffer, « Prediction error estimation: a comparison of  
resampling methods », in Bioinformatics, 21(15), pages 3301-3307, 2005  
<http://bioinformatics.oxfordjournals.org/cgi/content/full/21/15/3301>