SEGMENTATION

Pierre-Louis GONZALEZ

I. Les méthodes de segmentation. Introduction

- Les méthodes de segmentation cherchent à résoudre les problèmes de discrimination et de régression en **divisant** de façon progressive l'échantillon en sous-groupes (segments), pour construire un arbre de décision, ou un arbre de régression.

- Lors de chaque dichotomie, les deux parties sont les plus contrastées vis-à-vis de la variable à expliquer.

- Les premières approches ont été proposées par Sonquist et Morgan (1964) avec la méthode dite AID (Automatic Interaction Detection).

I. Les méthodes de segmentation. Introduction

➢Les travaux de Breiman , Friedman, Olshen et Stone (1984) connus sous le nom de **méthode CART** (Classification And Regression Tree) ont donné un nouvel essor à la segmentation.

➢Notons que les méthodes ne sont pas toujours présentes dans les logiciels statistiques « classiques ». Par contre de nombreux produits spécifiques sont présents sur le marché et connaissent un succès croissant avec le **développement du data mining**.

I. Les méthodes de segmentation. Introduction

Expliquer une variable

qualitative :

- Ménage propriétaire de sa résidence principale
- Hôtel équipé de la climatisation
- Diagnostic médical

ARBRE DE DECISION

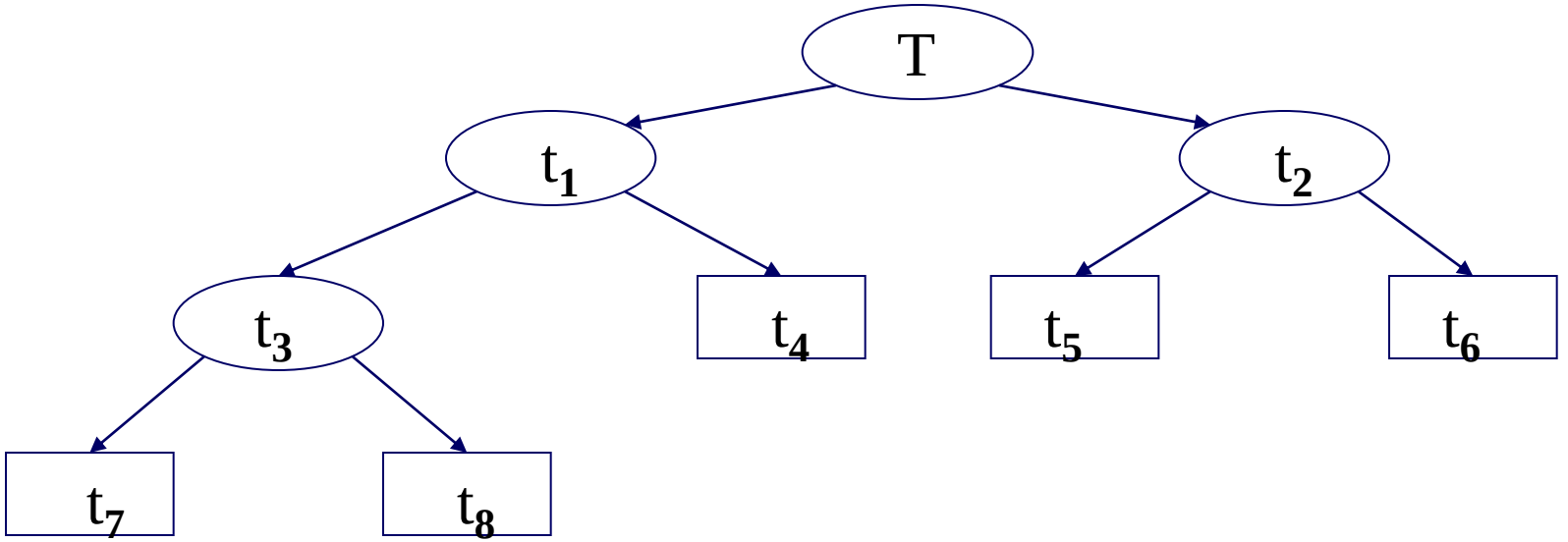
quantitative :

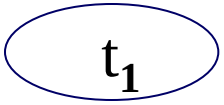
- C.A. d'entreprise/salarié
- Taux de mémorisation d'une annonce
- Salaire d'un cadre

REGRESSION PAR ARBRE

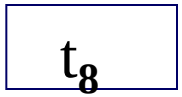
I. Les méthodes de segmentation. Introduction

Arbre de décision





: Segments **intermédiaires**



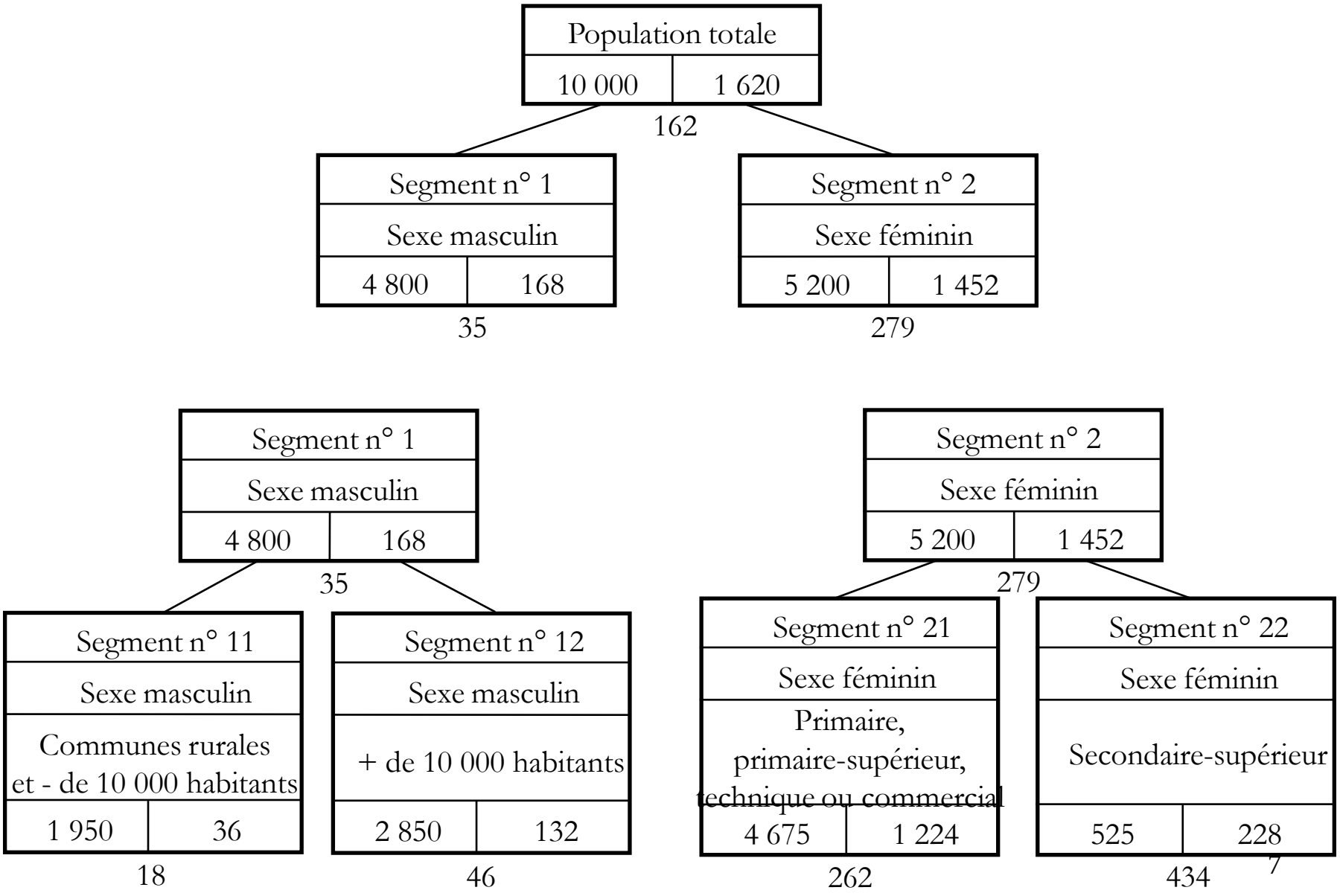
: Segments **terminaux**

II. Les méthodes de segmentation. Exemple

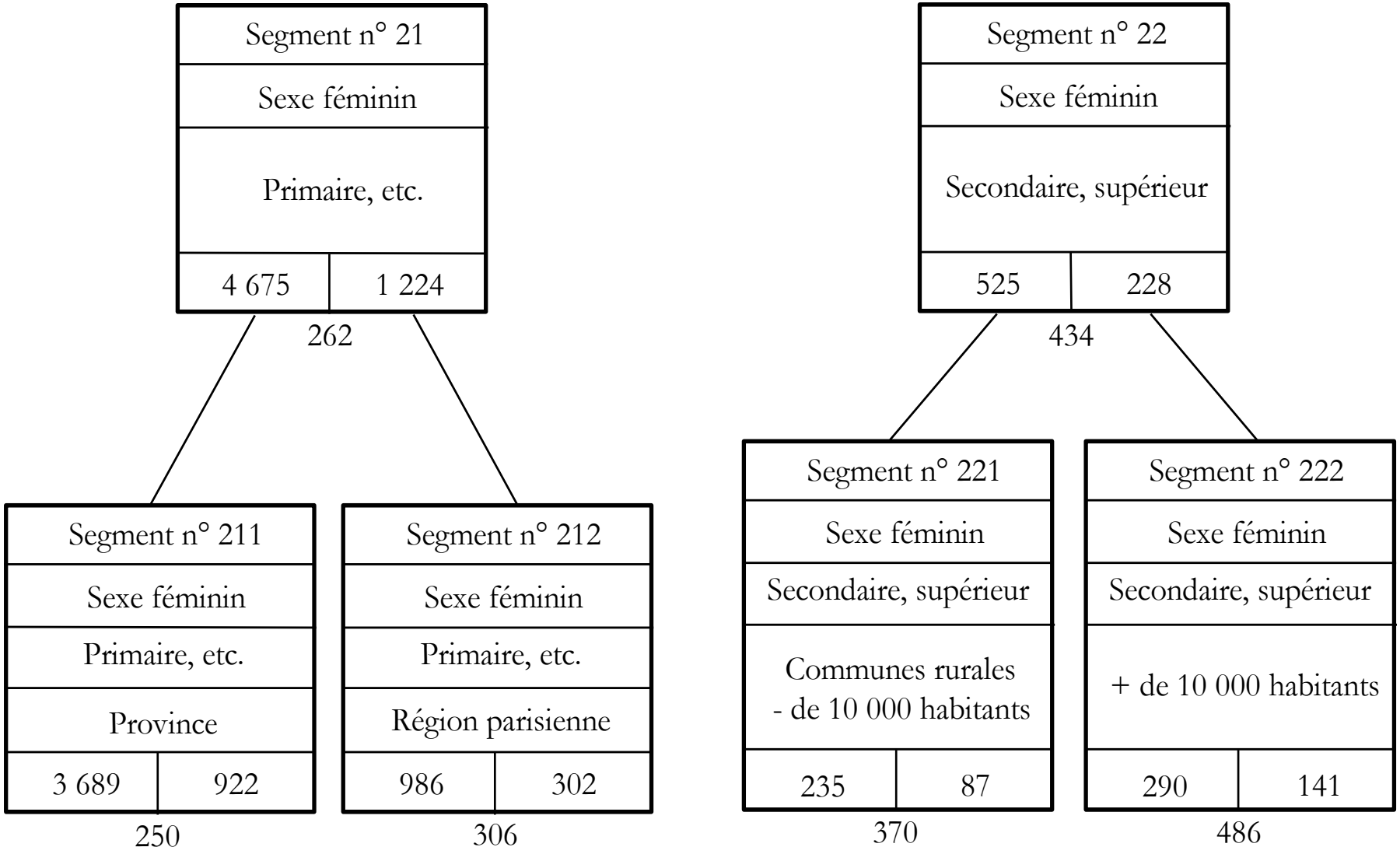
Intéressons-nous à **l'audience d'une revue mensuelle** auprès d'un échantillon de 10000 personnes. La variable à expliquer Y est le fait de lire ou de ne pas lire la revue. Les variables explicatives sont au nombre de 6 :

- Sexe (2) : M F
- Âge (2) : <20, 20-34, 35-49, 50-64, >=65 ans
- Niveau d'études (5) : Primaire, Primaire-Sup, Technique, Secondaire, Supérieur
- CSP (6)
- Catégorie commune (7)
- Région OJD (12)

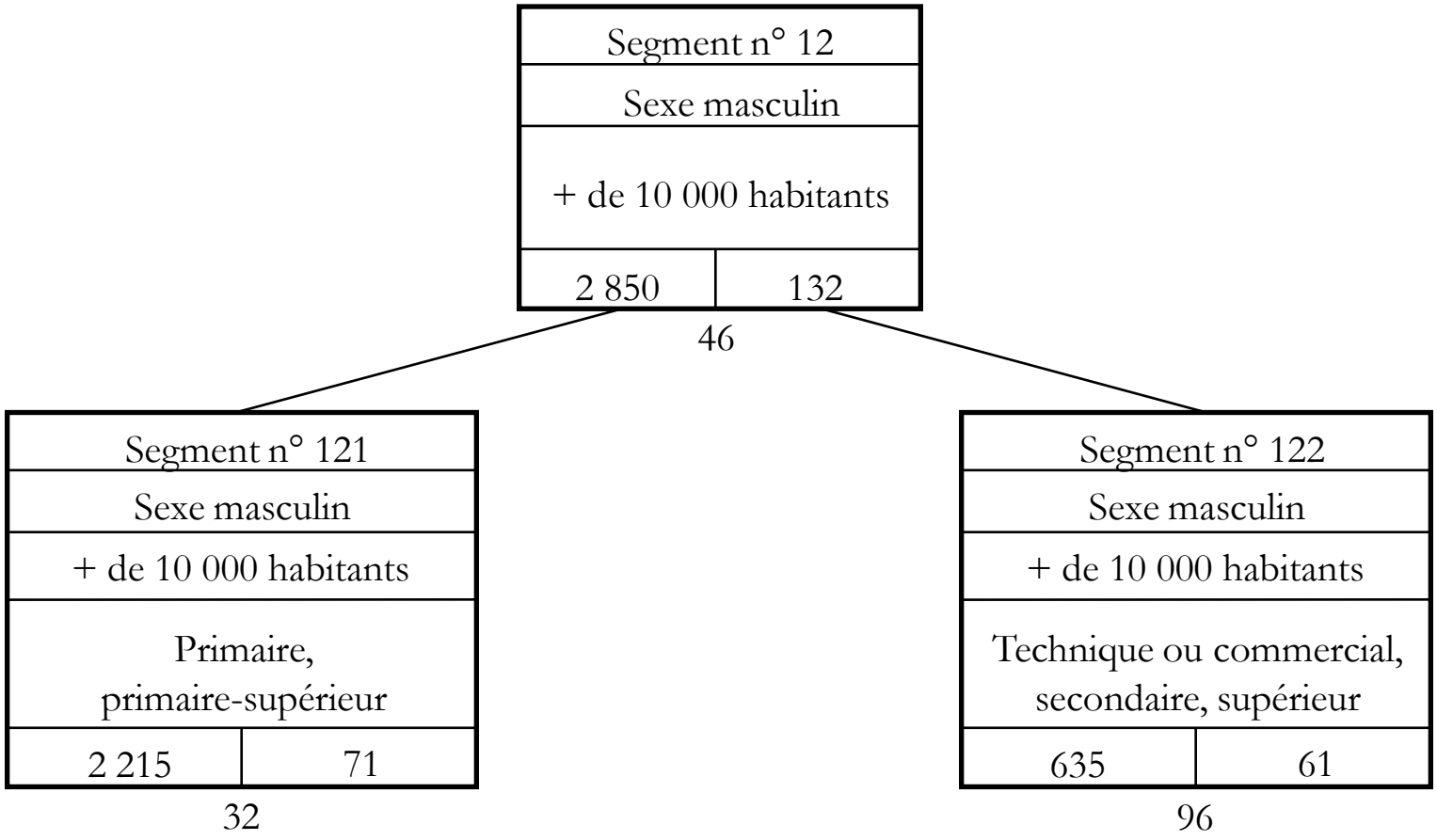
II. Les méthodes de segmentation. Exemple



II. Les méthodes de segmentation. Exemple



II. Les méthodes de segmentation. Exemple



III Principe des méthodes

III.1 Les dichotomies

Le nombre de dichotomies possibles d'une **variable qualitative à m modalités est égal à : $2^{m-1}-1$.**

Exemple: 4 catégories A, B, C, D.

A,BCD AB,CD B,ACD AC,DB
C,ABD AD,BC D,ABC 7 dichotomies

m	2	3	4	5	6	7	8	9	10	11	12
$2^{m-1}-1$	1	3	7	15	31	63	127	255	511	1023	2047

III Principe des méthodes

Si la variable est à **modalités ordonnées**, et que l'on souhaite respecter cet ordre lors des dichotomies, m-1 dichotomies sont possibles.

A,BCD AB,CD ABC,D

III.2 Les étapes de l'algorithme

A chaque étape de la segmentation, pour chaque variable:

- **il faut chercher sa meilleure dichotomie**
- **et retenir la meilleure variable.**

III Principe des méthodes

III. 3 Les critères

III.3.1. Variable à expliquer quantitative Y

Méthode AID (Morgan Sonquist 1963)

L'objectif est d'avoir des moyennes $\overline{y_1}$ $\overline{y_2}$ des deux groupes, d'effectifs n_1 n_2 , les plus différentes possible à chaque coupure.

Maximiser la variance interclasse:

$$\text{Max}\left[n_1\left(\overline{y_1} - \overline{y}\right)^2 / n + n_2\left(\overline{y_2} - \overline{y}\right)^2 / n\right] / s^2$$

III Principe des méthodes

$$\rightarrow \quad \max \frac{n_1 n_2}{n s^2} \left(\overline{y_1} - \overline{y_2} \right)^2$$

Remarque: On peut aussi minimiser la variance intra-classe

$$\rightarrow \quad \min \frac{n_1}{n} \; s_1^2 + \frac{n_2}{n} \; s_2^2$$

Simplification :

Il est inutile avec ce critère d'étudier les $2^{m-1} - 1$ dichotomies de chaque variable, $m - 1$ suffisent.
En effet, la meilleure dichotomie doit respecter l'ordre des moyennes.
On ordonne $\overline{y_1}, \overline{y_2}, \cdots \overline{y_m}$

Exemple : si n = 12, alors on étudie 11 dichotomies au lieu de 2 047.

III Principe des méthodes

III.3.2. Variable à expliquer Y qualitative à deux modalités

On calcule le Khi-deux associé à chaque dichotomie possible et on choisit la dichotomie qui maximise ce critère.

$$\chi^2 = \sum_{i,j} \left(n_{ij} - n_{i.} n_{.j} / n \right)^2 / \left(n_{i.} n_{.j} / n \right)$$

III Principe des méthodes

- **Tableau de contingence**

	M	F	
Lit	168	1452	1620
Ne lit pas	4632	3748	8380
	4800	5200	10000

n_{ij}

- **Effectifs espérés sous hypothèse d'indépendance**

777,6	842,2
4022,4	4357,6

$n_{i.}n_{.j}/n$

III Principe des méthodes

On choisit la dichotomie qui rend le χ^2 max.

Soient f_1 et f_2 les pourcentages de la première catégorie de Y dans les deux groupes.

$$\chi^2 = (f_1 - f_2)^2 \times n_1 n_2 / nf(1-f)$$

où $f = (n_1 f_1 + n_2 f_2) / n$

Il s'agit d'un cas particulier de la méthode AID

$\rightarrow$ AID avec Y = 1 / 0

III Principe des méthodes

III.4. Les règles d'arrêt.

Elles prennent en compte :

- Effectif minimum par segment.
- **Y quantitative**: **test de Student** comparant les moyennes dans les deux segments .
- **Y qualitative:** **Critère du Khi-deux** significatif pour un risque α (en général 5%) défini par l'utilisateur.

III Principe des méthodes

III.5. Précautions

- Méthode **très dépendante de l'échantillon**
- **Nécessité de grands échantillons** (environ 1000 observations au minimum)
- **Validation sur échantillons tests.**

IV. La méthode CART

Méthode proposée par Breiman, Friedman, Ohlsen, Stone (1984)

- La méthode CART permet de **construire un arbre de décision binaire** par divisions successives de l ’échantillon en **deux** sous-ensembles.

- Contrairement aux autres méthodes de segmentation, **elle n’impose aucune règle d’arrêt de division des segments basée sur une approche statistique.** Elle fournit, à partir de l’arbre binaire complet, la séquence des sous-arbres obtenue en utilisant une procédure d’élagage.
- Celle-ci est basée sur la suppression successive des branches les moins informatives.

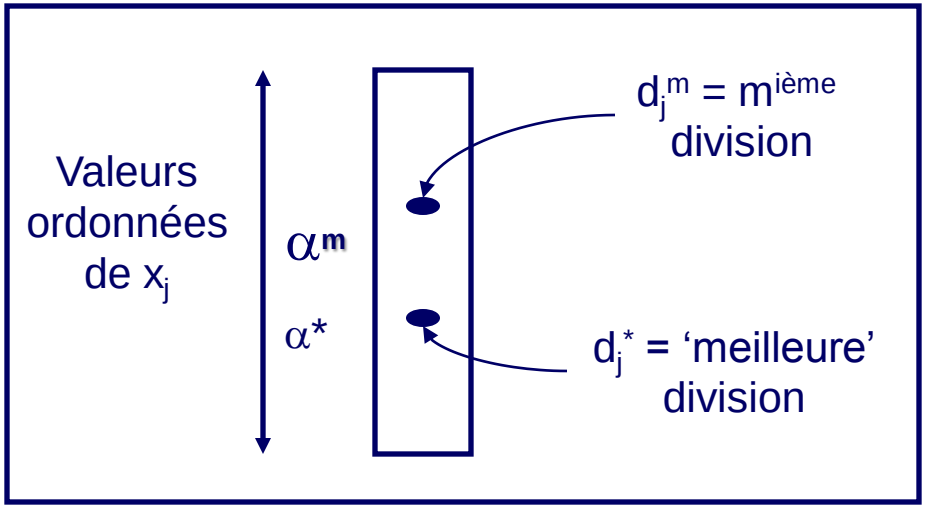
IV. La méthode CART

- Au cours de la phase d'élagage, la méthode sélectionne un sous-arbre optimal, en se fondant sur l'estimation de l'erreur théorique d'affectation ou de prévision **(réduction du critère d'impureté)** à l'aide soit **d'un échantillon-test**, quand on dispose de suffisamment d'observations, soit par **techniques de validation croisée.**

- La meilleure division d* d'un nœud est celle qui assure la plus grande réduction de l'impureté en passant du nœud à ses segments descendants. Cette notion de maximum absolu est très stricte. Il peut exister en effet des divisions presque aussi bonnes, pouvant jouer un rôle important au niveau des interprétations et qui conduiraient à la construction d'un arbre différent.

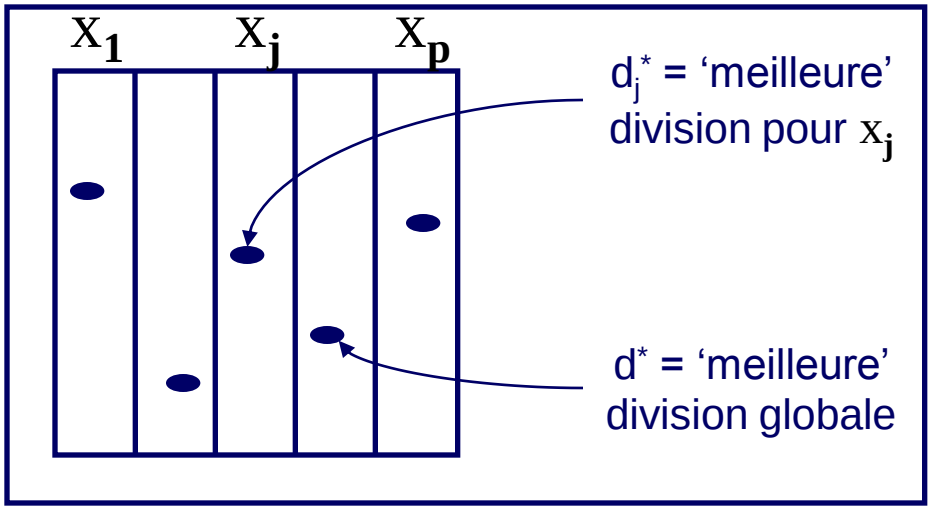
IV. La méthode CART

IV.1 Choix des meilleures divisions: Cas des variables explicatives quantitatives



La meilleure division pour la variable xj

IV. La méthode CART



Meilleures divisions pour l 'ensemble des variables

IV. La méthode CART

IV.2 Choix des meilleures divisions:

Cas des variables explicatives qualitatives

Selon le nombre de modalités et la présence ou non d ’un ordre logique entre elles, on obtient les possibilités de division :

- Une variable **binaire** fournit une division
- Une variable **nominale** N (k modalités non ordonnées) fournit $2^{k-1} - 1$ divisions
- Une variable **ordinale** O (k modalités ordonnées) fournit k-1 divisions possibles

IV. La méthode CART

IV.3 Discrimination : critère de division

- **Indice d'impureté associé au segment t :**

$$i(t) = \sum_{r}^{k} \sum_{s}^{k} P(r/t)P(s/t)$$

Avec r≠ s et où P(r/t) et P(s/t) sont les proportions d 'individus dans les classes c_r et c_s dans le segment t
(i(t) est l 'indice de diversité de Gini)

Exemple:

Segment t
Taille 313
Oui 135
Non 178

i(t)=2[(135/313) *(178/313)

= 0,49056

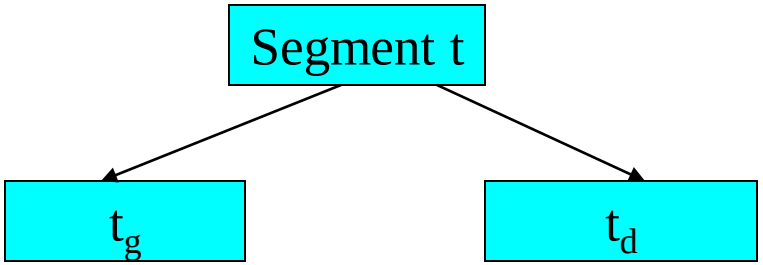
Segment pur : ne contient que des individus d 'une classe, **i(t) = 0**
Segment mélangé : **i(t) ≠ 0**. Plus le mélange des classes dans le segment est important, plus l'impureté i(t) est élevée.
Remarque: i(t) représente la probabilité de mauvais classement pour un individu tiré au hasard parmi les individus du segment t

IV. La méthode CART

- **Réduction de l'impureté**

Chaque division d'un segment entraîne une **réduction de l'impureté**

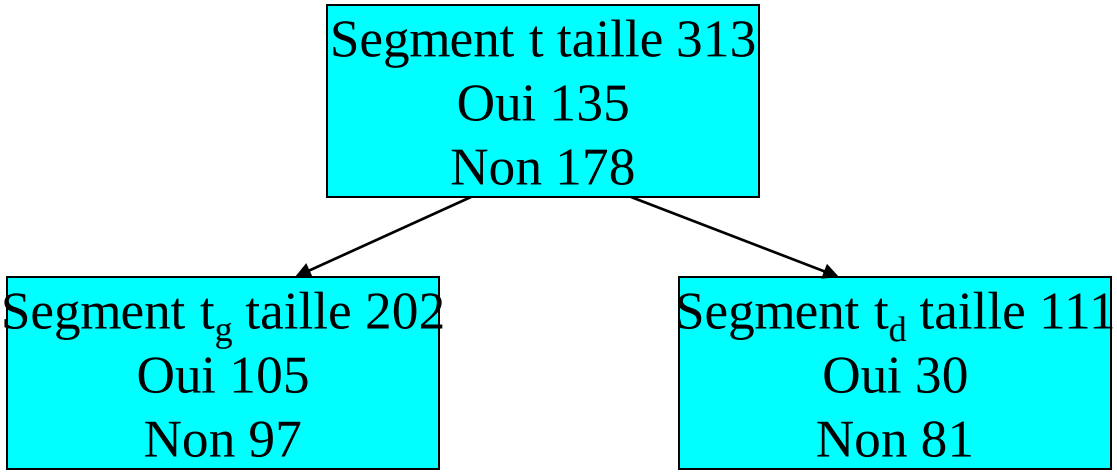
$$\Delta i(s,t) = i(t) - p_g i(t_g) - p_d i(t_d)$$



les p_g sont les proportions d 'individus du nœud t respectivement dans les segments descendants t_g et t_d (la fonction i(t) étant concave, l 'impureté moyenne ne peut que décroître par division d 'un nœud).

IV. La méthode CART

- **Exemple de calcul de l'impureté**



$$\Delta = 2[(135/313)*(178*313)] - (202/313)*2*[(105/202)*(97/202)$$

$$- (111/313)*2*[(30/111)*(81/111)]$$

$$=0,49056-0,645*0,4992-0,355*0,3944$$

$$=0,0285$$

IV. La méthode CART

- **Réduction maximale pour chaque variable**

$$\Delta i(s^*, t) = \max\{\Delta i(s, t)\}$$

- **Réduction maximale pour l'ensemble des p variables**

$$\Delta^* = \max_{j=1...p}\{\Delta i(s^*, t)\}$$

IV. La méthode CART

IV.4 Discrimination : Arrêt des divisions, affectation, taux d'erreur apparent (T.E.A.)

- **Nœud terminal :**
  - s'il est pur (il contient des observations toutes identiques)
  - s'il contient trop peu d'observations

- Un segment terminal est **affecté à la classe qui est la plus représentée**

- **Taux d 'erreur apparent de classement (T.E.A) associé à un segment terminal de l 'arbre T affecté à la classe s**:

$$R(s/t) = \sum_{r=1}^{k} p(r/t)$$

p(r/t) = est la proportion d'individus du segment t affecté à la classe c_s et qui appartiennent à la classe c_r

IV. La méthode CART

- **Taux d'erreur associé à l'arbre**

$$TEA(T) = \sum_{t \in T} \frac{n(t)}{n} R(s/t)$$

où n(t) = effectif du segment t

TEA(T) représente la proportion d'individus mal classés dans l'ensemble des segments terminaux.

IV. La méthode CART

IV.5 Discrimination : Sélection du meilleur sous-arbre.

- **Construction de l 'arbre complet T_{max},**

  sur l'échantillon d'apprentissage

- **Elagage sur l'échantillon d'élagage**

A partir de l'arbre complet, on détermine la séquence optimale de sous-arbres emboîtés: $\{T_{max-1}, \ldots T_h, \ldots T_1\}$ avec $1 \leq h < max$

Le taux d 'erreur apparent (TEA) de T_h vérifie :

$$TEA(T_h) = \min_{T \in S_h} \{TEA(T)\}$$

où S_h est l 'ensemble des sous-arbres de T_{max} ayant h segments terminaux

➢ **Evaluation des performances sur Échantillon test:**

Choix de T* tel que l ’erreur théorique de classement (ETC) vérifie :

$$ETC(T^*) = \min_{1 \le h \le max} \{ETC(T_h)\}$$

Si l’on dispose de suffisamment d’observations (environ un millier), on peut choisir la répartition suivante:

échantillon d’apprentissage: 60%
échantillon d ’élagage: 20%
échantillon test : 20%

Dans le cas de petits échantillons on peut procéder par validation croisée.

IV. La méthode CART

IV.6 Sélection du meilleur sous-arbre par validation croisée

Dans le cas d'un échantillon de petite taille il est impossible de répartir les individus en trois échantillons. On procède alors par validation croisée.

Procédure:

- La première étape consiste à déterminer la séquence de sous-arbres emboîtés,
à partir du grand arbre T_{max} construit à l'aide de l'échantillon total L

- La deuxième étape consiste à diviser l'échantillon L en V sous-ensembles, qui vont permettre de déterminer l'arbre optimal et d'évaluer ses performances

Sélection du meilleur sous-arbre par validation croisée. Exemple

Présentons cette approche à l'aide d'un exemple développé dans l'ouvrage de Jean-Pierre Nakache et Josiane Confais: « Statistique explicative appliquée » TECHNIP 2003

Les tableaux qui suivent présentent:

- 1. La séquence de sous-arbres associée à l'arbre T_{max} construit sur l'ensemble des données.

- 2. Les séquences d'élagage construits sur les sous-ensembles L_1 L_2 ….. L_{10}

- 3. Le choix du meilleur sous-arbre et l'évaluation de ses performances.

Remarque:

La validation présentée utilise la notion de coût-complexité $C_\alpha(T)$ d'un arbre, compromis prenant en compte le coût des erreurs de classement et la complexité de l'arbre: nombre de segments terminaux.

$$C_\alpha(T) = TEA(T) + \alpha \left|\tilde{T}\right|$$

TEA(T) est le taux d'erreur apparent de T par resubstitution

$\left|\tilde{T}\right|$ est le nombre de segments terminaux de T

Sélection du meilleur sous-arbre par validation croisée. Exemple

Exemple d'application : données Nodule

Il s'agit du diagnostic de malignité chez 382 patients ayant un nodule du foie et divisés en deux groupes : 259 cas de tumeur maligne et 123 cas de tumeur bénigne. Les variables mesurées sur chaque patient sont les variables binaires – jaunisse, hépatomégalie, hémachromatose, cirrhose, cancer primitif et antécédents hépatiques – codées 1-non et 2-oui, et les variables ordinales – taux de sédimentation, phosphatases alcalines et gamma gt – codées 1-normal, 2-augmenté (+) et 3-augmenté (++).

Pour ces données, la méthode de validation croisée est nécessaire du fait de la faible taille de l'échantillon total (n = 382). La séquence S de sous-arbres associée à l'arbre T_{max} obtenue en utilisant la procédure d'élagage est la suivante :

S	T_1	T_2	T_3	T_4	T_5
$\lvert\tilde{T}_i\rvert$	10	7	5	3	1
$\alpha_i (\times 10^3)$	0	1.82	16.81	46.43	102.09

où $\lvert\tilde{T}_i\rvert$ est le nombre de segments terminaux de l'arbre T_i et α_i la valeur du paramètre de complexité de l'arbre T_i.

Sélection du meilleur sous-arbre par validation croisée. Exemple

Tableau 8.9 Séquences $S_1, S_2, \ldots, S_{10}$, construites à partir des échantillons respectifs $L^1, L^2, \ldots, L^{10}$

S_1	T_1^1	T_1^2	T_1^3	T_1^4	T_1^5	T_1^6
$\alpha_i^1 \times 10^3$	0.0	2.01	4.81	9.43	44.00	100.58
$mc(L_1)$	1	1	2	2	4	13
S_2	T_2^1	T_2^2	T_2^3	T_2^4		
$\alpha_i^2 \times 10^3$	0.0	2.07	52.00	97.67		
$mc(L_2)$	3	2	2	13		
S_3	T_3^1	T_3^2	T_3^3			
$\alpha_i^3 \times 10^3$	0.0	36.39	110.47			
$mc(L_3)$	6	10	12			
S_4	T_4^1	T_4^2	T_4^3	T_4^4	T_4^5	
$\alpha_i^4 \times 10^3$	0.0	2.00	9.52	44.72	109.01	
$mc(L_4)$	7	7	7	9	12	
S_5	T_5^1	T_5^2	T_5^3	T_5^4	T_5^5	T_5^6
$\alpha_i^5 \times 10^3$	0.0	2.00	6.54	18.69	43.65	109.01
$mc(L_5)$	4	3	1	0	2	12

Sélection du meilleur sous-arbre par validation croisée. Exemple

S_6	T_6^1	T_6^2	T_6^3	T_6^4	
$\alpha_i^6 \times 10^3$	0.0	13.89	50.78	104.65	
$mc(L_6)$	5	6	6	12	
S_7	T_7^1	T_7^2	T_7^3	T_7^4	T_7^5
$\alpha_i^7 \times 10^3$	0.0	9.26	27.52	47.24	100.29
$mc(L_7)$	4	4	2	3	12
S_8	T_8^1	T_8^2	T_8^3	T_8^4	T_8^5
$\alpha_i^8 \times 10^3$	0.0	9.43	18.52	50.39	100.29
$mc(L_8)$	3	3	3	3	12
S_9	T_9^1	T_9^2	T_9^3	T_9^4	T_9^5
$\alpha_i^9 \times 10^3$	0.0	1.98	18.69	44.00	100.29
$mc(L_9)$	1	1	1	3	12
S_{10}	T_{10}^1	T_{10}^2	T_{10}^3		
$\alpha_i^{10} \times 10^3$	0.0	46.61	107.56		
$mc(L_{10})$	4	6	13		

$mc(L_v)$ est le nombre de mal classés de l'échantillon L_v ($v = 1, \ldots, 10$).

Sélection du meilleur sous-arbre par validation croisée. Exemple

Tableau 8.10 Détermination de l'estimation du coût par validation croisée.

S	T_1	n_i^1	T_2	n_i^2	T_3	n_i^3	T_4	n_i^4	T_5	n_i^5
$\lvert\tilde{T}_i\rvert$	10		7		5		3		1	
$\alpha_i \times 10^3$	0		1.82		16.81		46.43		102.09	
$\alpha_i' \times 10^3$	0		5.53		27.94		68.85		>68.85	
S_1	T_1^1	1	T_1^3	2	T_1^4	2	T_1^5	4	T_1^6	13
S_2	T_2^1	3	T_2^2	2	T_2^2	2	T_2^3	2	T_2^4	13
S_3	T_3^1	6	T_3^1	6	T_3^1	6	T_3^2	10	T_3^3	12
S_4	T_4^1	7	T_4^2	7	T_4^3	7	T_4^4	9	T_4^5	12
S_5	T_5^1	4	T_5^2	3	T_5^4	0	T_5^5	2	T_5^6	12
S_6	T_6^1	5	T_6^1	5	T_6^2	6	T_6^3	6	T_6^4	12
S_7	T_7^1	4	T_7^1	4	T_7^3	2	T_7^4	3	T_7^5	12
S_8	T_8^1	3	T_8^1	3	T_8^3	3	T_8^4	3	T_8^5	12
S_9	T_9^1	1	T_9^2	1	T_9^3	1	T_9^4	3	T_9^5	12
S_{10}	T_{10}^1	4	T_{10}^1	4	T_{10}^1	4	T_{10}^2	6	T_{10}^3	13
Nb total de mal classés $C^{vc}(T_i)$		(38) 9.9 %		(37) 9.7 %		(33) **8.6 %**		(48) 12.6 %		(123) 32.2 %

IV. La méthode CART

IV.7 Compléments concernant la méthode CART

IV.7.1. Divisions concurrentes et divisions suppléantes

- La meilleure division d* d'un nœud est celle qui assure la plus grande réduction de l'impureté en passant du nœud à ses segments descendants. Cette notion de maximum absolu est **très stricte**. Il peut exister en effet des divisions presque aussi bonnes, pouvant jouer un rôle important au niveau des interprétations et qui conduiraient à la construction d'un arbre différent.
- De façon à apporter des solutions à ce problème, deux options supplémentaires sont proposées dans l'approche CART :
- **les divisions concurrentes (équi-réductrices dans SPAD)** qui assurent, après d* les plus fortes réductions de l'impureté. Elles permettent d'intervenir sur le choix de la meilleure variable explicative. On peut définir la première division concurrente, puis la deuxième et ainsi de suite.

IV. La méthode CART

▪ **les divisions suppléantes (équi-divisantes dans SPAD)** qui fournissent les répartitions les plus proches de la division d*. Elles permettent de gérer l'existence de données manquantes dans l'affectation d'un nouvel individu à une classe.

Quand pour un individu une donnée est manquante pour une variable divisant un segment, on cherche la variable la remplaçant au mieux. Comme pour les divisions équi-réductrices, il est possible de définir la première, la deuxième,…, meilleure division suppléante.

IV.7.2. Coût de classement

Il est possible d'associer des coûts aux erreurs de classement.

V. Comparaison des méthodes de segmentation avec les autres méthodes de discrimination et de classement

- La segmentation n'est pas vraiment multidimensionnelle au sens géométrique du terme (pas de calcul de distance comme en analyse factorielle discriminante).

- Par contre on utilise les variables explicatives conditionnellement les unes par rapport aux autres. On peut donc parfois atteindre des **effets d'interaction** assez difficiles à saisir par d'autres méthodes. Ceci ne signifie d'ailleurs pas qu'ils sont tous étudiés.

V.1 Avantages des méthodes de segmentation

- **l'ergonomie des résultats** : l'arbre de décision binaire est lisible par tout utilisateur et constitue à ce titre un moyen de communication des résultats très apprécié.
- **la mixité des variables** qu'accepte la procédure : variables nominales, ordinales, continues peuvent être mélangées au niveau des variables explicatives .
- **la validation par une méthode de rééchantillonnage** est une des techniques de validation les plus transparentes pour l'utilisateur.
- **la robustesse de la méthode** vis à vis des valeurs extrêmes et des données erronées.

V.2 Inconvénients des méthodes de segmentation

On doit également reconnaître aux méthodes de segmentation un certain nombre de points faibles, qui rendent son utilisation exclusive insuffisante.

- **l'aspect séquentiel est redoutable** : les covariations qui servent à sélectionner les variables ne mesurent pas un lien causal et une variable peut en cacher une autre beaucoup plus fondamentale, qui n'apparaîtra pas dans la suite du processus.

  Les divisions de réserve (concurrentes et suppléantes) sont là pour pallier partiellement cet inconvénient. L'approche par segmentation perd alors en simplicité. En effet l'utilisateur est alors souvent déconcerté par **l'instabilité des arbres obtenus**.
- De nouvelles pratiques apparaissent pour tenter d'apporter des solutions à ce problème: **Bagging , Boosting.**

- **La sélection des variables présente un léger biais** du fait que les variables explicatives ayant un plus grand nombre de modalités, en offrant plus de divisions possibles, sont plus souvent sélectionnées.

- Il se peut que la nature du phénomène étudié fasse que des combinaisons linéaires (après éventuel recodage) soient optimales pour prévoir la variable étudiée (ou son logit ou toute autre fonction).

- **Les méthodes de segmentation nécessitent des échantillons de grande taille.**

V.3 NOUVELLES PRATIQUES

BAGGING

Un arbre T_t est construit à partir de l'échantillon « boostrapé » EB_t qui fournit une règle de décision R_t conduisant à une estimation c_t du coût associé à l'arbre T_t. La règle de décision finale R* est fournie par l'agrégation des règles R_t (t= 1,2…n). Ainsi pour classer un individu x, un vote pour la classe k est enregistré pour chaque règle R_t, pour laquelle $R_t(x)$= k et **R*(x) est alors la classe avec le plus de votes (classe majoritaire)**

Breiman rapporte dans deux articles les résultats de cette méthode obtenus sur 7 fichiers de taille moyenne. En utilisant 50 échantillons « boostrapés », le coût moyen de la règle R* est égal au **coût de la règle, obtenu en utilisant uniquement l'échantillon de base, multiplié par une valeur a comprise dans l'intervalle [0,57 ; 0,94].**

V.3 NOUVELLES PRATIQUES

BAGGING

Un cas particulier de bagging est celui appliqué à des arbres de décision avec introduction **d'un tirage aléatoire parmi les variables explicatives.** On a dans ce cas une double randomisation et on parle de **forêts aléatoires (Breiman 2001)**

BOOSTING (algorithme Adaboost 1999)

L'idée est de générer plusieurs règles d'affectation d'un individu. Dans la construction de la **deuxième règle, une attention particulière est portée aux individus mal classés par la première règle** pour essayer de les classer correctement, et ainsi de suite.

En général dix règles permettent de réduire le nombre d'erreurs de l'échantillon test d'environ 25%.

VI .Conclusions.

- La **validation** (validation croisée ou utilisation d'un échantillon test) est un aspect les plus importants lors de la mise en œuvre des méthodes de segmentation et de discrimination.
- **Utiliser plusieurs techniques.**

- **Nouvelles méthodes:**
  - Analyse discriminante PLS et régression logistique PLS.
  - Forêts aléatoires.

Quelques références

- NAKACHE J-P., CONFAIS J. **“ Statistique explicative appliquée ”** Technip (2003)
- ZIGHED D.A., RAKOTOMALALA R. **“ Graphes d’induction ”** Hermès (2000)
- Logiciels:
  - SPAD 7.4
  - SIPINA

EXEMPLE ASSURANCE