# Classification supervisée & validation

## 1/ Introduction

| | $X_1 \dots X_p$ | Y |
|---|---|---|
| 1 | | |
| 2 | | |
| ⋮ | | |
| N | | |

variables explicatives — variable à expliquer

Objectifs de la segmentation.

Discrimination ⟷ Y ⟷ Régression

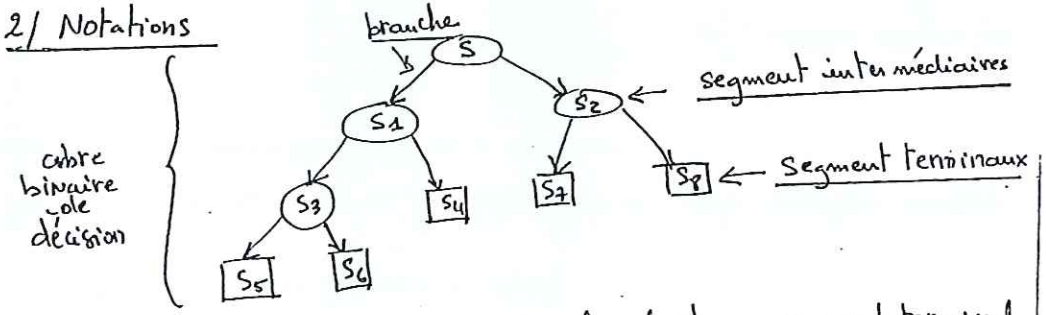
Y → variable qualitative à k modalités

Y → variable quantitative

- Sélectionner les variables explicatives les plus discriminantes
- Construire une règle de décision permettant d'affecter un nouvel individu à l'une des k classes

- Sélectionner les variables explicatives les plus liées au phénomène décrit par Y
- Construire une règle de décision permettant d'affecter à un nouvel individu une valeur $\hat{y}$.

## 2/ Notations




$A_{max}$ : arbre binaire complet pour lequel chaque segment terminal contient 1 seul individu.

A un sous arbre de $A_{max}$ obtenu par élagage d'une ou de plusieurs branches.

L'idée de base consiste à effectuer la division d'un noeud de telle sorte que les deux segments descendants soient :

- plus homogènes que le noeud parent
- les plus différents possible entre eux vis-à-vis de la variable y.

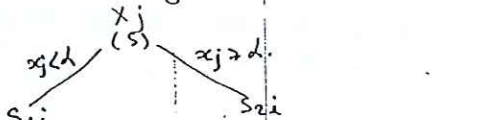
1 - Établir pour chaque noeud l'ensemble des divisions admises
2 - Définir un critère permettant de sélectionner la "meilleure" division d'un noeud.
3 - Définir une règle permettant de déclarer un noeud comme terminal ou intermédiaire.
4 - Affecter chaque noeud terminal à l'un des groupes (cas de la discrimination), ou affecter une valeur à y pour chaque noeud terminal (cas de la régression).
5 - Estimer le risque d'erreur de classement (cas de la discrimination) ou prévision (cas de la régression) associé à l'arbre

### 3.1 Définition de l'ensemble des divisions admissibles

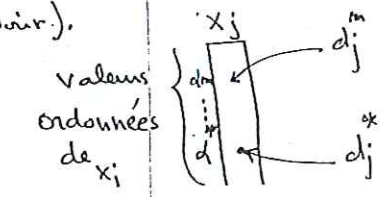
#### Cas 1 / Variable explicative continue

Soit S le premier segment contenant tous les individus

- pour une variable $X_j$, on examine toutes les divisions possibles de la forme $X_j < d$ ($d \in$ à l'étendue de



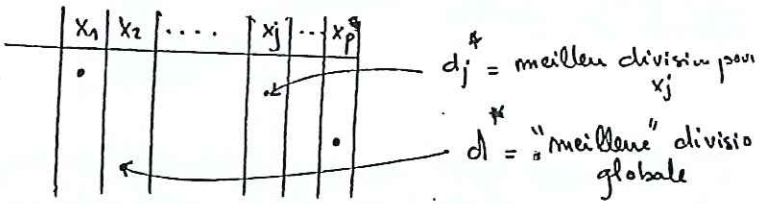

Soit $d_j^1, \dots d_j^m$ les m divisions possibles, et $d_j^*$ la meilleure division. (critère à voir en 3.2)



- Une variable à 2 modalités ne fournit qu'1 seul divisi
- " à k modalités ordonnées fournit k-1 divisi
- " " " non-ordonnées " $2^{k-1}-1$ divisi

Toutes les divisions sont examinées et l'on retient la meilleu au sens d'1 critère (à voir en 3.2).

Ainsi on obtient après la détermination de la meilleu division pour chaque variable (tout type confondu) :

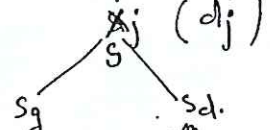



$d^*$ est la meilleure division, au sens du critère, toutes variables confondues.

## 3.2 Définition du critère de sélection de la meilleure divisio

### Cas 1 / Y est une variable continue (Régression)

Le critère est fondé sur la variance de y au sein des segment descendants. Cette variance doit être plus faible que la variance de y dans le noeud parent.




On mesure pour chaque division $d_j^m$ sa variance résiduelle comme la moyenne pondérée des variances de y à l'intérieu de chacun de ses segments descendants :

$$Var_Y\left(d_j^m, S\right) = \left(\frac{n_g}{n_S} Var_Y(S_g) + \frac{n_d}{n_S} Var_Y(S_d)\right)$$

variance résiduelle :

$$\boxed{Var\left(d_j^*, S\right) = \min_{m \in d_j} \left\{ var_Y\left(d_j^m, S\right) \right\}}$$

La meilleure division globale $d^*$ est la division $d_j^*$ minimi la variance résiduelle :

$$\boxed{Var\left(d^*, S\right) = \min_{i=1..p} \left\{ var\left(d_j^*, S\right) \right\}}$$

### Cas 2 / Y est une variable qualitative (Discrimination)

Soit y une variable qualitative à k modalités.
La sélection d'1 division doit être telle que les segments descendants soient plus "purs" que le noeud parent.
Autrement dit il faudrait que le mélange des modalités dans les nœuds descendants soit moins important que dans le noeud parent.

A chaque segment S on associe une mesure de l'impureté i(

$$i(s) = \sum_{r=1}^{k} \sum_{t=1}^{k} P(r/s) \cdot P(t/s) \quad (r \neq t)$$

$P(r/s), P(t/s)$ : proportions d'individus prenant la modal r, t dans S.

Un segment est pur si il ne contient que des individus de la même classe (même modalité pour Y) ; avec $i(S) = 0$.

Chaque division $d_j^m$ de s entraîne une diminution de l'im pureté :

$$\boxed{\Delta_j^m = i(s) - p_g i(s_g) - p_d \cdot i(s_d)}$$

$p_g = \frac{n_g}{n}$ $\quad p_d = \frac{n_d}{n}$

maximise la réduction-ole l'impureté :

$$\Delta_j^* = \max_{m \in d_j} \{ \Delta_j^m \}$$

Sur l'ensemble de toutes les variable la division est effectuée selon la variable qui assure :

$$\Delta^* = \max_{j=1..p} \{ \Delta_j^* \}$$

- Une fois le segment s segmenté on reapplique 1/et 2/ aux noeuds descendants.
- On arrête la construction-ole l'arbre lorsque tous les segments sont déclarés terminaux :
  - ne nécessitent plus de divisions (exp : $i(s)=0$, vars
  - leur taille est < à un effectif fixé a priori.

## 3.3/ Règle d'affectation d'1 nouvel individu

### Cas 1/ Y variable continue (Prévision)

soit i un nouvel individu, on le fait déscendre dans l'arbre. soit $S_i$ son segment terminal d'appartenance

La valeur $y_i$ de l'individu i est estimée par la moyenne de y dans le segment $S_i$.

### Cas 2/ Y est une variable qualitative (classement)

Soit S un noeud terminal. On note $n_1(s), \dots n_k(s)$ les nb d'individu dans S prenant la modalité 1, ... , k de Y.

est affecté à la classe la mieux représentée :

$$\text{Classe de } i = \underset{j=1..k}{\text{ARGMax}} \; (n_j(s))$$

## 3.4 Estimation du risque d'erreur-ole classement (cas discrimin ou de prévision (cas Régression)

### Cas 1/ Y variable continue (Prévision)

On associe à chaque segment terminal $S_t$ -ole l'arbre A l'erreur de prévision $R_{S_t}$ :

$$R(S_t) = \frac{n_{S_t}}{n} \text{Var}_Y(S_t)$$

$n_{S_t}$ : taille du segment $S_t$

$n$ : nb total d'individu

$$\text{Var}_Y(S_t) = \frac{1}{n_{S_t}} \sum_{i \in S_t} (y_i - \bar{y})$$

L'erreur apparente de Prévision. $EAP(A)$ -ole l'abre A vaut :

$$EAP(A) = \sum_{S_t \in A} R(S_t)$$

$\frac{EAP(A)}{\text{Var}(s)} \simeq (1 - R^2)$ : représente la proportion de la variance totale non expliquée p l'arbre.

Remarque : $EAP(A_{max}) = 0$

### Cas 2/ Y variable qualitative (classement)

à tout segment terminal $S_t$ -ole l'arbre A associé à une classe j (modalité j) correspond une erreur apparente de classe

$$R(j/S_t) = \sum_{r=1}^{k} P(r/S_t)$$

affectés à la classe j' dans le segment $s_t$.

$$P(r/s_t) = \frac{n_r}{n_{st}}$$

L'erreur apparente de Classement (EAC) associé à A:

$$EAC(A) = \sum_{s_t \in A} \frac{n_{st}}{n} R(j/s_t) = \sum_{s_t \in A} \sum_{\substack{r=1 \\ (r \neq s)}}^{k} \frac{n_{rt}}{n}$$

$EAC(A)$ : représente la proportion d'individus mal classés dans l'ensemble des segments terminaux.

Bibliographie :


* - Gilbert Saporta. "Probabilités, Analyse des données et Statistique". TECHNIP

* - Edwin Diday, J. lemaire, J Pouget, F. Testu. "Eléments d'analyse des données". DUNOD

* - Gilles Celeux." Analyse discriminante sur variables continues INRIA. collection DIDACTIQUE.

- Gilles Celeux & J.P NAKACHE. " Analyse discriminante sur variable Qualitatives".

- P Bertier & J.M. BOUROCHE. " Analyse des données multidimensionnelles". Presses Universitaires de France

** - Trevor Hastie, Robert Tibshirani & Jerome Friedman. "The elements of statistical Learning. Data Mining, Inf and Prediction". Springer

** - Anil. K. JAIN & Richard C. Dubes. "Algorithms for clus Prentice Hall, Advanced Reference series. Data."

** - Breiman, Friedman, olshen & Stone. "Classification And Regression Trees".

## Validation et portée des résultats en classification.

- Observe-t-on vraiment quelque chose ?
- Les données ont-elles une structure
- A-t-on découvert des classes préexistantes ou au contraire a-t-on découpé une réalité continue en classes ?
- Est ce que les configurations obtenues sont stables, compte tenu de ce que l'on sait sur la précision des données ?
- Quelle est l'influence sur les résultats d'une modification du tableau de données ?

## Étude de la stabilité des classes, des résultats par rééchantillonnage.

Les méthodes par rééchantillonnage sont privilégiées pour étudier la stabilité des résultats vis-à-vis de fluctuations d'échantillonnage

- se sont des méthodes de calculs intensifs qui reposent sur des techniques de simulations d'échantillons à partir d'1 seul échantillon.
- Elle consiste à répéter des analyses sur les différents échantillons simulés puis à étudier les fluctuations des résultats obtenus.

Il existe trois principales approches permettant de générer les échantillons artificiels : Jackknife, BootStrap, et la Validation croisée.

## 1. Technique de Jackknife. (Quenouille (1949); Tukey (1958), Miller (19

Objectif : estimation du biais et de la variance d'1 estimateur

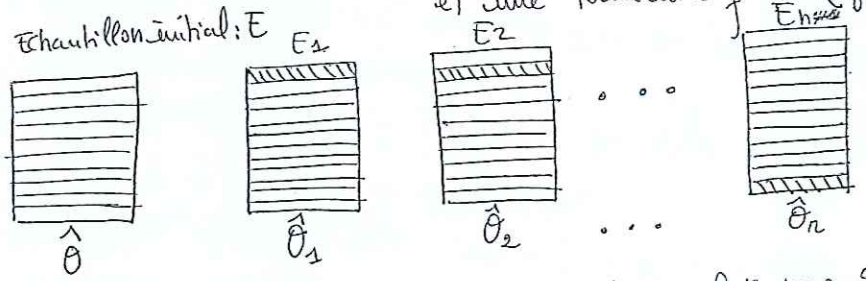
Le biais : mesure la précision de l'estimateur

La variance : mesure la stabilité de l'estimateur

Le biais permet de dire si mon estimateur, estime correctement, de manière precise la valeur esomptée

La variance permet de dire si mon estimateur (bon ou mauvais) fluctue beaucoup ou pas. en fonction des données d'origine.

La situation idéale : c'est un estimateur précis (biais: et une variance faible (grande stabilité




- On génère n échantillons en supprimant à chaque fois une observati de l'échantillon initial.

$\theta$ : paramètre à Estimer ; $\hat{\theta} = s(E)$ l'estimation de $\theta$ sur la base de E

$\hat{\theta}_i = s(E_i)$ " " $\theta$ sur la base de $E_i$

On note $\hat{\theta}_{\cdot} = \frac{1}{n} \sum_i \hat{\theta}_i$ : la moyenne des estimations.

L'estimateur de Jackknife $\tilde{\theta}$ est défini :

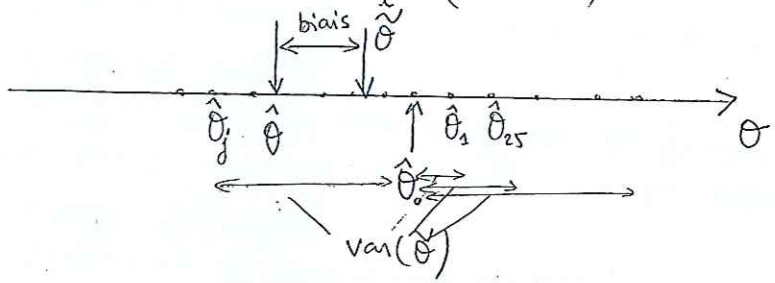
$$\tilde{\theta} = n\hat{\theta} - (n-1)\hat{\theta}_{\cdot} = n(\hat{\theta} - \hat{\theta}_{\cdot}) + \hat{\theta}_{\cdot}$$

Le biais de Jackknife vaut:

$$b = \hat{\theta} - \tilde{\theta} = (n-1)\left(\hat{\theta}_{\cdot} - \hat{\theta}\right)$$

L'estimation de Jackknife de la variance :

$$\text{Var}(\tilde{\theta}) = \frac{n-1}{n} \sum \left(\hat{\theta}_{i\cdot} - \hat{\theta}_{\cdot}\right)^2$$

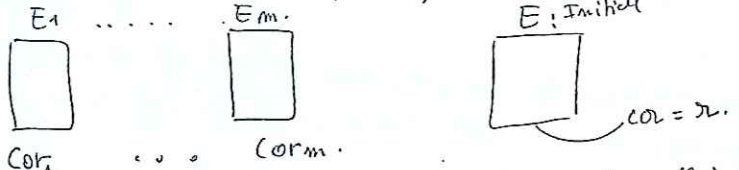



## 2. Technique du Bootstrap: (Efron (1979))

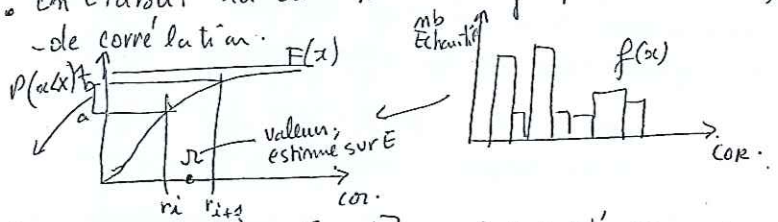
objectif: estimer la variabilité d'1 paramètre en fournissant un intervalle de confiance.

- Il s'agit de simuler $m > 30$ échantillons de même taille $n$ que l'échantillon initial. Chaque échantillon est obtenu par tirage aléatoire avec remise parmi les $n$ individus.

Exemple : Estimation du $\text{cor}(X_1, X_2)$ :




- On Établit la distribution de fréquence du coefficient de corrélation.




$P(\text{cor} \in [r_i, r_{i+1}]) = [a, b]$ ← Probabilité pour que l'intervalle de confiance de $r \in [r_i, r_{i+1}]$

- On obtient ainsi une estimation de l'intervalle de confiance de $r$ sans avoir recours à des hypothèses distributionnelle sur les données.

## 3. La validation croisée

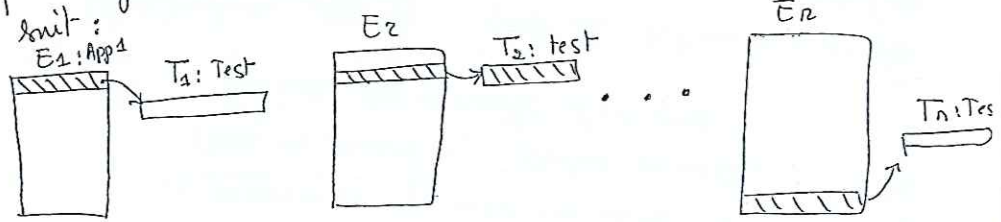
objectif: Sert à évaluer des erreurs de prédictions. Elle est utilisée dans le cas où l'échantillon initiale est de petite taille (cas la regression / Analyse discriminante

- A l'origine il s'agit de séparer l'échantillon de base en deux blocs de tailles # généralement: L'échantillon d'apprentissage, et l'échantillon-test.

L'échantillon d'apprentissage : sert à formulé le modèles, à établir les règles de décision ou d'affectation.

L'échantillon test : Sert à appliquer les règles et à estimer les performances du modèle.

- pour agrandir la taille de l'échantillon on procède comme suit :




- On extrait à chaque fois (k observations ici $k=1$) pour testes le modèle et le construir sur $(n-k)$ observations.
- La moyenne des taux de prédiction des $n$ Echantillons Tests constitue la prédiction à retenir pour le modèle définit sur E.