Les arbres de décision (decision trees)

Christine Decaestecker, ULB

Marco Saerens, UCL

Arbres de Décision
(ou Méthode de Segmentation)

- **Origines:**
  Ces méthodes ont pris essentiellement leur essor dans le cadre des approches d'**apprentissage automatique** (***machine learning***) en **Intelligence Artificielle.**

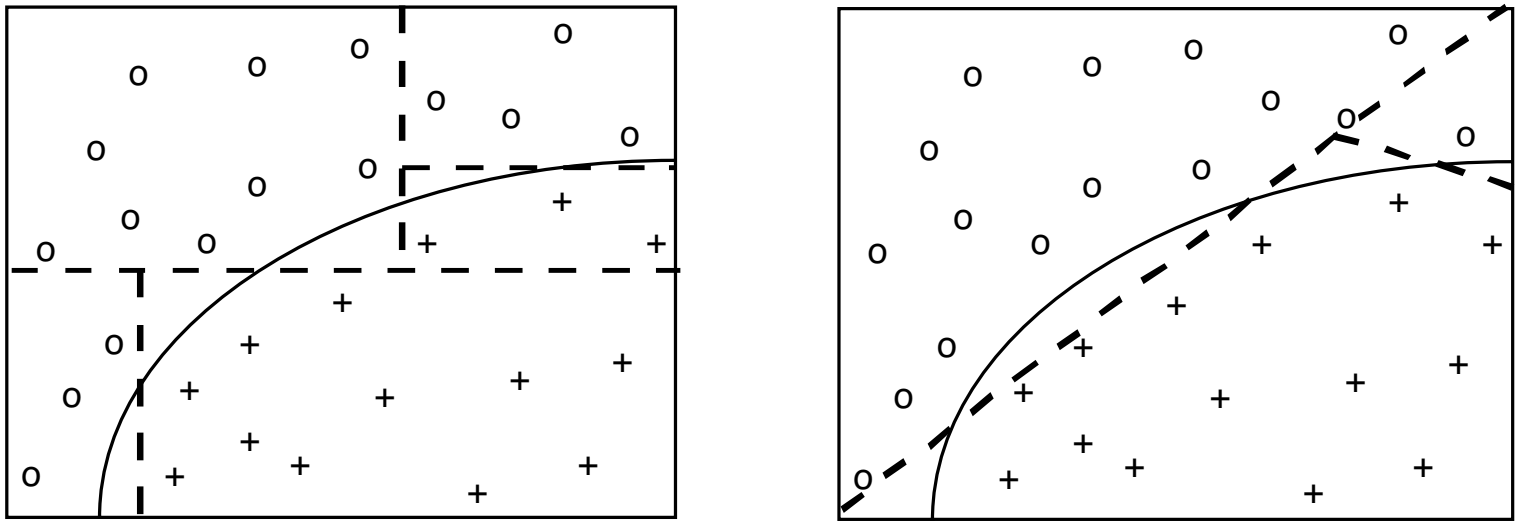
- **Particularités (de l'I.A. en général):**
  met l'accent sur sur la convivialité et l'intelligibilité (ou la lisibilité) des résultats
  => en classification supervisée: sortie de résultats sous la forme de règles logiques de classification:

  "SI tel ensemble de conditions sur telles variables est satisfait ALORS le cas appartient à telle classe".

  => résultats plus facilement interprétables et donc exploitables
  => communication plus aisée avec les spécialistes du domaine traité.

- **Ex d'algorithme:** ID3 (*Inductive Decision Tree*) et son successeur C4.5, CART (*Classification and Regression Tree*), CHAID (*Chi-Square Automatic Interaction Detection*), QUEST (*Quick, Unbiased, Efficient Statistical Trees*), cf. TP.

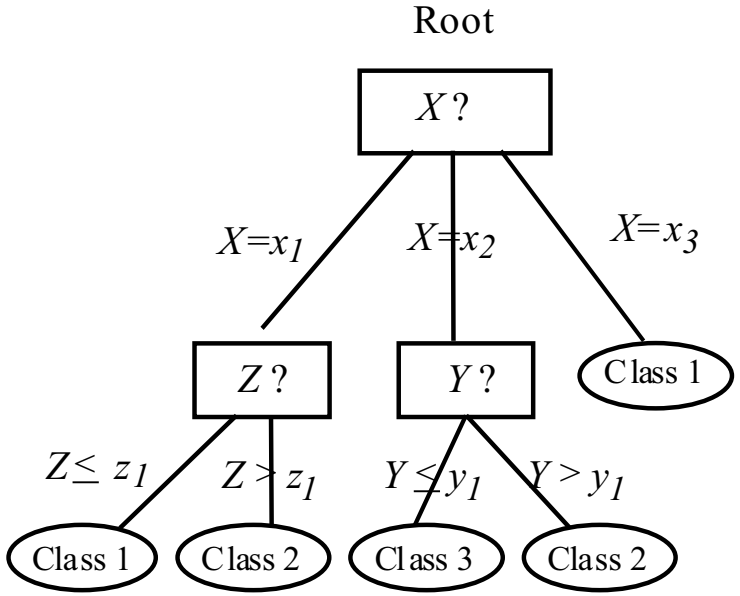
- **Principes: 2 phases**
  - ***Phase 1: construction***: sur base d'un ensemble d'apprentissage, processus récursif de division (souvent binaire) de l'espace des données en sous-régions de + en + pures en terme de classes (estimé sur base d'un critère). Dans le cas de ***données numériques*** 2 approches possibles: séparations **parallèles** aux axes versus **obliques:**



=> décomposition d'un problème de classification en une suite de tests (imbriqués) portant sur une variable (parallèle aux axes) ou une combinaison linéaire de plusieurs variables (oblique).

=> règles de classifications sous forme d'arbres dont chaque extrémité (encore appelée "feuille") indique l'appartenance à une classe.

Ex: arbre sur variables mixtes (*X* catégorielle, *Y* et *Z* numériques)



Chaque "nœud" intermédiaire réalise un test portant sur une variable dont le résultat indique la branche à suivre dans l'arbre. Pour ***classer un nouveau cas***: suivre le chemin partant de la racine (nœud initial) à une feuille de l'arbre en effectuant les différents tests à chaque nœud.

=> La classification est réalisée en posant une suite de questions relatives à certaines propriétés du cas considéré.

Classe allouée à une feuille: déterminée sur base de la classification de l'ens. d'apprentissage: classe majoritairement représentée parmi les exemples qui "tombent" dans cette feuille.

– ***Phase 2: élagage ("pruning")***: supprimer les branches (parties terminales) peu représentatives pour garder de bonnes performances prédictives (généralisation) => nécessité d'un critère pour désigner les branches à élaguer.
Après élagage, les nouvelles feuilles sont labelisées sur base de la distribution des exemples d'apprentissage (classe majoritaire).

– ***Arbres parallèles aux axes versus obliques (données numériques)?***

Avantages des arbres parallèles:

- suite de tests monovariés;
- pas de problème de combinaison de var, d'échelle ... ;
- sélection de variables intégrée;
- rapidité;
- génération de règles logiques simples de classification.

***Désavantages - limitations*:**

- approximation par "escaliers" des surfaces de séparation;
- les critères usuels de sélection de tests ne tiennent pas compte des densités de points dans l'espace des données (pour sélectionner une variable et sa valeur seuil, cf. après).

- **Phase de construction d'un arbre (parallèle aux axes):**
  2 étapes à chaque nœud d'un arbre en construction (processus récursif):
  1. Production d'une série de tests relatifs aux différentes variables (qualitatives ou quantitatives):
     - pour les variables quantitatives (discrète ou continue): tests (généralement) binaires: $X \leq$ ou $>$ seuil (=> seuil à déterminer);
     - pour les variables qualitatives (ou déclarées comme telles): chacune des valeurs possibles est considérée comme une alternative (des sous-ensembles de valeurs peuvent aussi être considérés).
  2. Sélection du meilleur test (au niveau considéré) d'après un certain critère dont l'objectif est de diminuer le plus possible le mélange des classes au sein de chaque sous-ensemble créé par les différentes alternatives du test.

  Conditions d'arrêt: différentes possibilités (dépendent des algorithmes):
  Ex: pourcentage de cas appartenant à la classe majoritaire > seuil, ou
  nbre de cas dans une feuille < seuil, ou combinaison des 2, ...
  (cf. T.P.).

Objectif général: générer une séquence hiérarchique de tests, aussi courte que possible, qui divise successivement l'ensemble des données d'apprentissage en sous-ensembles disjoints, tels que des sous-groupes de cas appartenant à la même classe soient rapidement détectés.
=> *stratégie de "diviser-pour-régner".*

=> le critère de sélection (étape 2) est souvent basé sur la ***théorie de l'information***, et notamment sur la notion d'***entropie*** (mesure de l'hétérogénéité d'un mélange).

Ex: critère du "***Gain d'entropie***" ou "***Information mutuelle***" (ID3, C4.5)

- Soit un ensemble E d'exemples divisés en classes ω_1, ..., ω_k, ..., ω_q.
  L'entropie de la distribution des classes = quantité moyenne d'information (ici en bits => $\log_2$) nécessaire pour identifier la classe d'un exemple de E:

$$H(E) = -\sum_k P(\omega_k)\log_2\left(P(\omega_k)\right)$$

  où $P(\omega_k)$ est la probabilité *a priori* de la classe ω_k
- Soit un test T (portant sur une variable X) ayant m alternatives possibles qui divisent E en m sous-ensembles E_j, caractérisé par une entropie $H(E_j)$.

– ***L'entropie de la partition*** résultante, c'est-à-dire ***l'entropie conditionnelle de E étant donné T***, est définie comme l'entropie moyenne des sous-ensembles:

$$H(E|T) = \sum_j P(E_j)H(E_j)$$

– Le ***gain d'information*** apporté par le test T est donc:

$$Gain(E, T) = H(E) - H(E \mid T)$$

En pratique, les probabilités *a priori* sont estimées par les fréquences relatives calculées sur les données d'apprentissage.

Propriétés:

– $Gain(E, T)$ est maximum $\Leftrightarrow$ $H(E \mid T)$ est minimum
$\Leftrightarrow$ T minimise (en moyenne) le mélange des classes au sein des E_j.

– $Gain(E, T) \approx 0$ si T apporte peu d'information sur la classe (ex: une variable qualitative indépendante de la classe).

– $Gain(E, T)$ est ***biaisé*** en faveur des tests ayant un grand nombre m d'alternatives.
=> Biais à rectifier => critère du ***Rapport des Gains***

***Rapport des Gains**:*

$R\_Gain(T) = Gain(E, T) / H(T)$ avec $H(T) = -\sum_j P(E_j) \log_2 \left(P(E_j)\right)$

=> $H(T)$ = information potentielle générée en divisant un ensemble E en m sous-ensembles E_j.
=> R_Gain = proportion d'information générée par T et utile pour la classification.

Autre biais: Diviser par $H(T)$ favorise les partitions de E ayant un fort déséquilibre de la répartition des cas entre les E_j: $H(T)$ est maximum si la distribution est équirépartie ($P(E_j)$ égaux) et diminue en fonction du déséquilibre.

=> rectification par une contrainte supplémentaire sur le *Gain*:
à chaque nœud, choisir le test T qui maximise R_Gain parmi ceux dont le *Gain* est suffisamment grand, c-à-d $\geq$ *Gain* moyen de tous les tests examinés.

Tous ces problèmes sont évités si on se limite aux tests binaires !!

Il existe de ***nombreux critères*** basés sur différentes mesures caractérisant l'***efficacité d'un test T***, dont notamment les mesures statistique d'association entre 2 variables qualitatives (ex: Gini index, mesure du χ^2). On a montré qu'il n'existait pas de différences significatives quant à la qualité des arbres utilisant différents critères basés sur ces notions (ils ont tous leur points forts et leurs points faibles).

Cas des variables quantitatives:
Pour chaque variable quantitative X, toute valeur observée x_i donne lieu à un test binaire: $X \leq xi$?

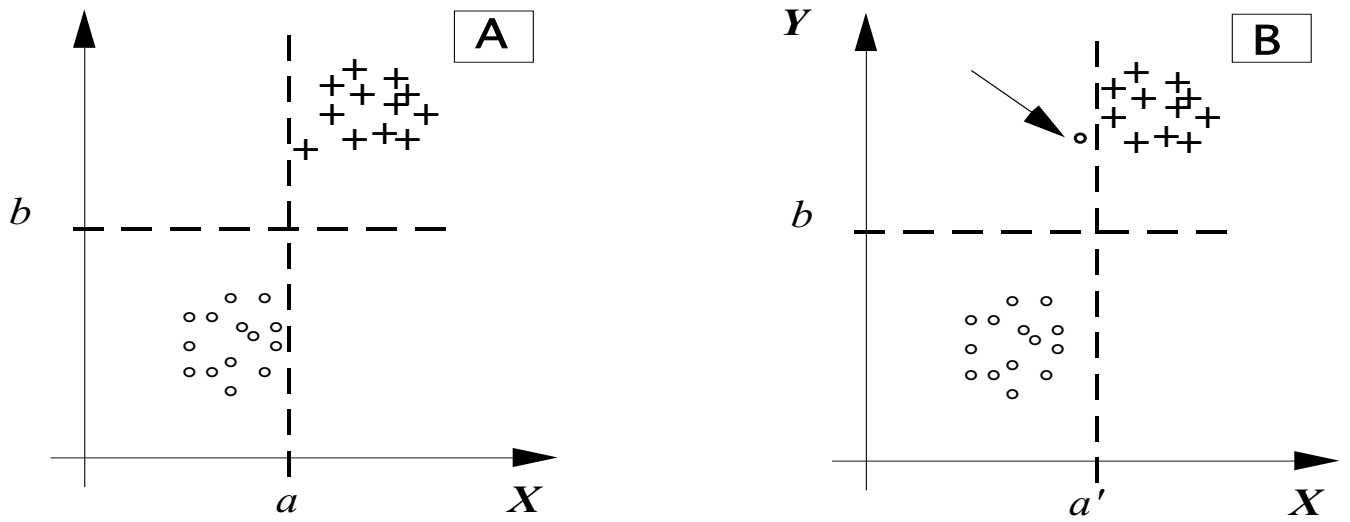
Propriété:
Les critères convexes (tels que le gain d'entropie) sélectionnent toujours des valeurs "seuil" situées à la frontière entre deux classes, c'est-à-dire entre deux valeurs consécutives observées par des exemples de classes différentes.

=> diminution du nombre de seuils à tester pour une variable continue,
MAIS ce type de critère considère la frontière inter-classe au sens strict.
=> forte sensibilité à tout ce qui peut altérer les frontières apparentes entre les classes (variation de l'ensemble d'apprentissage, bruit dans les données, et erreurs sur l'attribution des classes).

Cause essentielle: Les critères (classiques) de sélection des tests opèrent par comptage des éléments de chaque classe au sein des différentes branches générées par un test => ne tiennent pas compte de la distance (ou proximité) à une valeur seuil, ou de la densité des observations.

Autre possibilité: utiliser des tests statistiques (cf. TP).

Illustration de la problématique:



A: 2 tests ($X \leq a$) et ($Y \leq b$) jugés équivalents par un critère d'information pour séparer les 2 classes en présence (séparation parfaite).

B: Si altérations de la frontière inter-classes (par un cas "hors norme" ou *outliers,* ou erreur sur la classe) => le test sur X en a' sera avantagé par une mesure d'information.

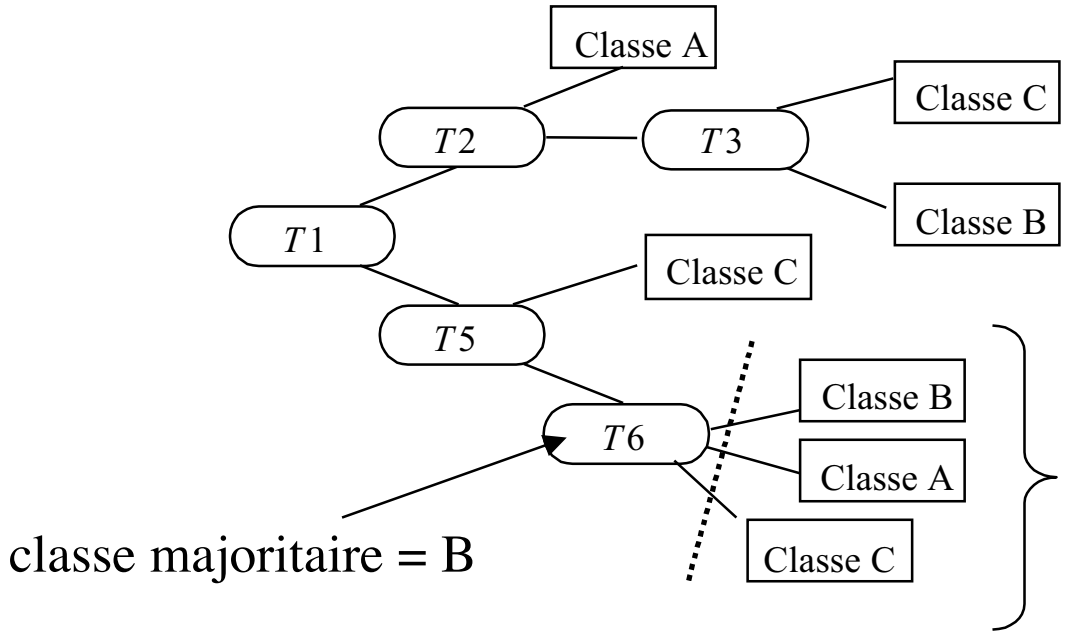
En **A et B**: le ***test sur Y*** apparaît ***préférable*** (vu la distance séparant les deux classes) => plus grande robustesse vis-à-vis des altérations de la frontière apparente inter-classes (cf B).

MAIS: pas d'influence si 1 ou 2 exceptions au sein d'une zone occupée par des cas d'une même classe (grâce à l'élagage).

=> sensibilité au "bruit" uniquement aux frontières inter-classes.

- **Phase d'élagage d'un arbre:**
  Objectif: supprimer les parties de l'arbre qui ne ***semblent*** pas performantes pour prédire la classe de nouveaux cas
  => remplacées par un nœud terminal (associé à la classe majoritaire).

  Processus: généralement de type "*bottom-up*" (du bas vers le haut: des extrémités vers la racine), basé sur une estimation du taux d'erreur de classification: un arbre est élagué à un certain nœud si le taux d'erreur estimé à ce nœud (en y allouant la classe majoritaire) est inférieur au taux d'erreur obtenu en considérant les sous-arbres terminaux.



3 branches élaguées:
taux d'erreur (estimé) en $T6 \leq$ taux d'erreur (estimé) obtenus en considérant les 3 feuilles.

=> élagages successifs (au départ des extrémités) jusqu'à ce que tous les sous-arbres restants satisfassent la condition sur les taux d'erreur de classification.

Différentes façons d'estimer l'erreur (dépendent des algorithmes):

- sur base de nouveaux exemples disponibles;
- via une validation croisée (cf. précédemment);
- sur base d'une estimation statistique, ex: borne supérieure d'un intervalle de confiance construit sur un modèle binomial (cf. rappels de proba.);
- ...

- **Production de règles de classification et autre processus d'élagage**
  Règle (aspect général) :
  "SI ... (partie conditionnelle) ... ALORS ... (partie conclusive) ...".

  Production d'un système de règles de classification à partir d'un arbre:
  via l'ensemble des chemins partant de la racine de l'arbre à chacune des feuilles.
  Chaque chemin = une règle:
  - partie conditionnelle = conjonction ("ET" logique) des tests rencontrés,
  - partie conclusive = classe associée à la feuille de l'arbre.

  Propriétés du système initial: (comme l'arbre) ***exhaustivité*** (couvre toutes les possibilités) et ***exclusivité*** mutuelle des règles (=> assure une partition de l'espace).

  Phase de simplification (élagage): (basée sur le gain en taux d'erreur)
  - élimination de certains tests de la partie conditionnelle d'une règle,
  - élimination d'une règle entière.

  => Autre technique d'***élagage plus souple***: n'importe quel test peut être supprimé directement (libéré du principe "*bottom up*").

Conséquences de la simplification:

Perte possible des propriétés d'exhaustivité et d'exclusivité.

=> Ordonnancement des règles finales suivant un ordre de priorité (défini suivant le taux d'erreur estimé):

=> Système final ordonné où la première règle qui couvre un cas (partie conditionnelle satisfaite) est alors choisie comme opérationnelle:

SI "règle **1**"
SINON "règle **2**"
SINON "règle **3**"
...
SINON ***classe par défaut*** (la plus fréquemment observée parmi les cas d'apprentissage non couverts par les règles précédentes).

- **Avantages et désavantages des arbres de décision (parallèles)**

  Avantages: (cf. précédemment)

  1. prise en compte simultanée de variables qualitatives et quantitatives (discrètes ou continues);
  2. pas d'hypothèse au sujet des données (modèle non-paramétrique);
  3. non affecté par les problèmes d'échelles de mesure des variables quantitatives (pas de combinaison arithmétique des variables) et détermine des seuils discriminants pour ces dernières;
  4. sélection des variables les plus informatives (en tenant compte des interactions);
  5. peu d'influence des données erronées, SAUF aux frontières inter-classes;
  6. algorithmes très rapides en phase de construction des arbres et lors de la classification de nouveaux cas (1 seul chemin est parcouru);
  7. règles logiques de classification aisément interprétables
     => extraction de connaissances explicites d'un ens. de données
     (= "*data mining*")
     => meilleure compréhension du problème étudié.

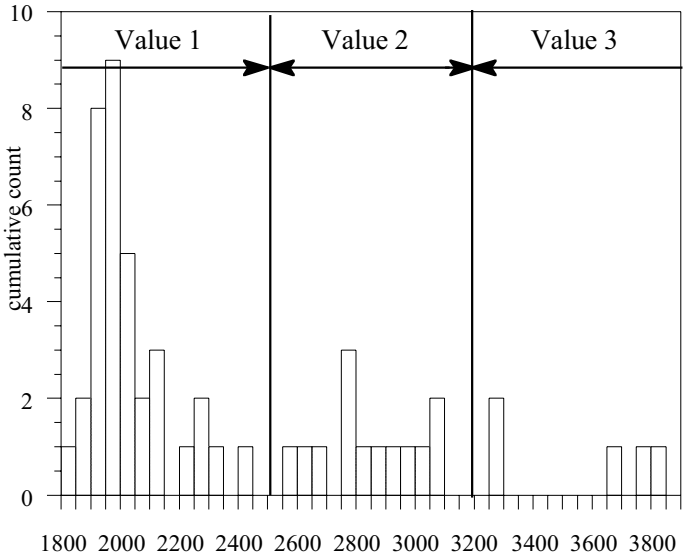
Limitations:

1. Traitement des ***variables numériques*** (cf. précédemment): génération des tests (choix des seuils) ne tient pas compte des propriétés de densité (proximité) des valeurs.
   => nouveaux développements avec de nouveaux critères de sélection pour les variables numériques.
2. ***Algorithmes séquentiels*** sans remise en cause des étapes précédentes (d'où rapidité), un peu assoupli dans la production de systèmes de règles.
3. ***Instabilité***: sensibles aux variations (même assez faibles) de l'ensemble d'apprentissage en termes d'exemples, des variables considérées;
   => variations dans les arbres produits (variables sélectionnées, seuils des variables numériques, structure de l'arbre, ...) et de leurs performances.

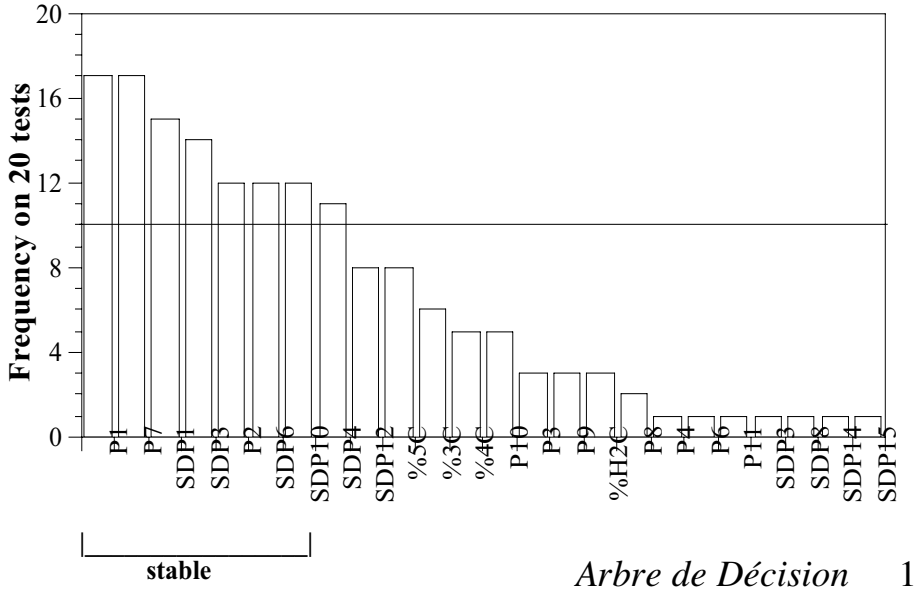
2. & 3. = Limitations similaires aux algorithmes de sélection de variables *stepwise*: algorithmes rapides mais n'investiguant qu'un nombre restreint de possibilités à chaque étape, sans remise en cause des choix précédents!
=> Une petite variation dans les données peut entraîner un choix différent à un certain niveau => sous-arbres différents => Quel est l'arbre optimal ?
(peut remettre en cause l'aspect "extraction des connaissances" !)

- **Problème d'instabilité et nouveaux développements:**
  - ***Discrétisation des variables numériques: "***prédécoupage" des intervalles de variation des variables numériques => variable ordinale.
  - ***Présélection de variables stables:*** analyse des différents arbres produits lors d'une validation croisée => identification d'un sous-ensemble de variables reprises régulièrement dans les arbres (> 50%) => construction d'un arbre sur cette base (+ stable, généralement + performant, et plus sûr du point de vue "*data mining*").

Discrétisation d'une variable continue par une méthode de clustering



Analyse de la stabilité de la sélection des variables. lors d'une validation croisée en 20 blocs



– Utilisation de la nature instable des arbres pour ***combiner des modèles différents*** => développement en plein essor (combinaison de classificateurs).
Principes:

- produire de petites variations:
  – dans l'ensemble d'apprentissage (via validation croisée ou *bootstrapping*),
  – dans les variables soumises à l'algorithme (sélection aléatoire d'une certaine proportion);
- produire un grand nombre de classificateurs (arbres) sur les différents ensemble de données ainsi produits (visions partielles et différentes des données);
- combiner les décisions des différents classificateurs (vote ou méthode plus sophistiquée).

Résultats: performances supérieures et stables.
MAIS: perte de l'intelligibilité des règles de classification générées!

- **Arbre de régression (méthode CART)**
  => Variable à expliquer (Y) numérique (continue ou discrète).
  - ***Principe:*** déterminer (par segmentation récursive) dans l'espace des variables descriptives (indépendantes) X_j des régions où la valeur de Y est homogène (variance faible).
    => Dans chacune de ces régions R_i, une bonne valeur prédictive de Y est sa moyenne (calculée sur l'ensemble d'apprentissage), et une évaluation de l'erreur est donnée par la variance (calculée sur l'ensemble d'apprentissage) (cf. notions de régression):

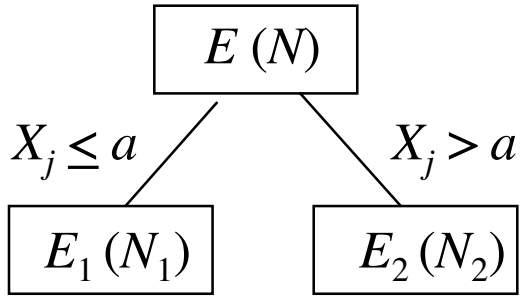
$$\hat{y} = E(Y \mid \mathbf{x}_j \in R_i) = \frac{1}{\#R_i} \sum_{j \mid \mathbf{x}_j \in R_i} y_j \quad \text{et}$$

$$\text{erreur quadratique} = s^2_{Y \mid \mathbf{x}_j \in R_i} = \frac{1}{\#R_i} \sum_{j \mid \mathbf{x}_j \in R_i} (y_j - \hat{y})^2$$

    Dans le cas de données numériques, possibilité (comme en classification) de construire des arbres parallèles ou obliques (par combinaisons linéaires des X_j) aux axes.

– ***Construction d'un arbre de régression:***
Même processus que pour la classification, mais avec un critère de sélection basé sur la variance résiduelle de Y dans les segments descendants (qui doit être plus faible que dans le nœud précédant).

Soit un ens. de données E (d'effectif N) séparé en 2 sous-ensembles E_1 et E_2 par un test sur une variable X_j.



Variance de Y dans E = $s^2_{Y|E}$

Variance résiduelle après division de E:

$$\frac{N_1}{N} s^2_{Y|E_1} + \frac{N_2}{N} s^2_{Y|E_2}$$

=> choix du test qui produit le minimum de variance résiduelle !

Arrêt de la construction lorsque plus (ou peu) de diminution de variance.

Valeur de Y associée à un nœud = valeur moyenne de Y dans ce nœud (sur base des données d'apprentissage) => $E(Y|E)$.

– ***Mesure d'erreur:***
Chaque feuille de l'arbre concerne sous-ens. de données F_i (d'effectif N_i)
=> erreur quadratique par feuille = $s^2_{Y|F_i}$

Erreur totale associée à un arbre: $C = \frac{1}{N} \sum_i N_i \; s^2_{Y|F_i}$ $\left(\text{avec } N = \sum_i N_i\right)$

Calculée ***sur l'ensemble d'apprentissage***: C = mesure d'adéquation ("*fitting*") du modèle aux données.
=> $C \big/ s^2_Y$ = ***% de variance totale non-expliquée par le modèle***
=> équivalent à l'expression $(1 - R^2)$ de la régression linéaire multiple (cf. précédemment).

– ***Elagage de l'arbre:*** comme en classification sur base d'une estimation de l'erreur en prédiction (sur base d'un ensemble test indépendant, ou via validation croisée, ...).