Les arbres de décision

Plan

Les données du Titanic

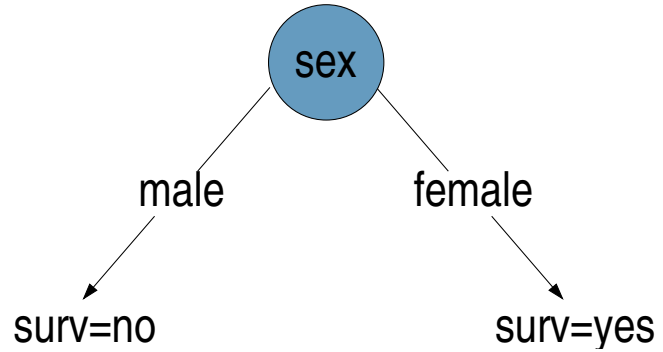
- Données historiques sur 2201 passagers du Titanic
- Tâche : prédire la survie d'un passager sur la base de 4 variables
- Var cible : survie {yes, no}
- Vars prédictives
  - classe {1st,2nd,3rd,crew}
  - age {adult, child}
  - sexe {male, female}

class	age	sex	surv
1st	adult	m	yes
crew	adult	m	no
3rd	child	m	no
2nd	adult	f	yes
...	...	...	...

Le principe du partitionnement

Etant donné un ensemble de données S ayant d variables prédictives
Trouver un test permettant de prédire la valeur de la variable cible

1. Choisir une variable de test x suivant un critère défini
2. Partitionner les exemples suivant les valeurs de x
3. Pour chaque partition, prédire la valeur de la variable cible



Un arbre à un noeud

Pour obtenir des arbres de complexité arbitraire, on applique cet algorithme de manière récursive.

L'algorithme de partitionnement récursif

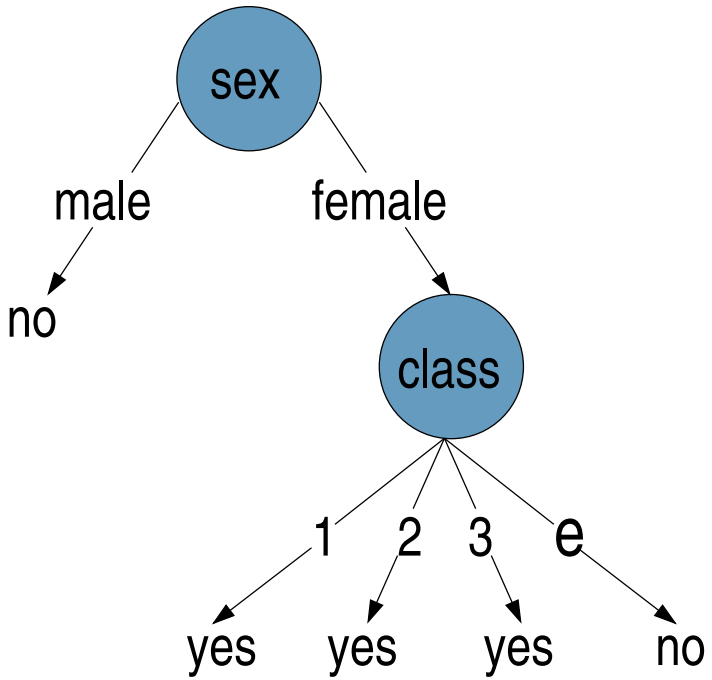
DT(S: données, X: vars prédictives, Y: var cible)

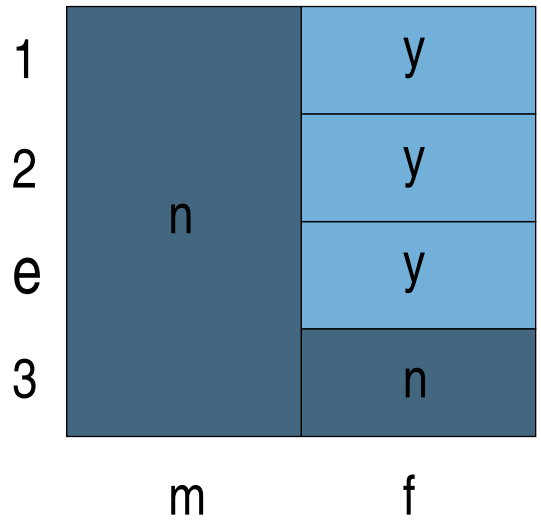
1. Créer noeud T
2. Si $X = \emptyset$ ou si Y a la même valeur $\forall s \in S$ alors retourner T avec prédiction :
   - en classification : classe majoritaire dans S
   - en régression : moyenne des y_i dans S
3. Choisir une variable de test $x \in X$ suivant un critère défini*
4. Partitionner S en m sous-ensembles S' suivant les valeurs de x*
5. Si x discrète, $X \leftarrow X - \{x\}$
6. Pour chaque partition S'
   DT(S', X, Y)

* Voir page 7

Remarques sur l'algorithme

- algorithme glouton pour éviter la recherche combinatoire : ni regard en avant, ni backtrack
- découpe des hyperrectangles dans l'espace des instances : frontières perpendiculaires aux axes





Variations sur l'algorithme de base

- Critère de choix de l'attribut de test (ligne 3)
  CART : index Gini
  C4.5 : gain d'information et rapport de gain
- Facteur de branchement b pour l'attribut de test X (ligne 4)
  CART : $b = 2$ (arbre binaire)
  C4.5 : par défaut, $b = |\mathcal{X}|$ si x nominal, $b = 2$ continu
- Même stratégie pour restreindre la complexité de l'arbre :
  Construire l'arbre jusqu'au bout, puis élaguer.

Plan

Entropie d'une variable

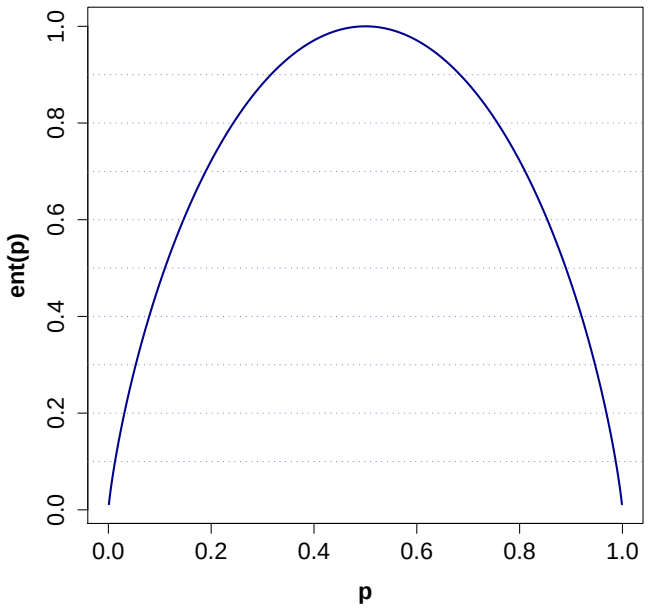
X variable aléatoire prenant ses valeurs x dans l'alphabet $\mathcal{X}$

- L'incertitude d'un événement x: l'inverse de sa probabilité $log\,\frac{1}{p(x)} = -log\,p(x)$.
- L'entropie de X:

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) log\, p(x)$$

où log s'entend à la base 2.

- L'entropie = nombre de bits requis pour décrire X



- Entropie d' une variable avec deux valeurs possible, p.ex. $\{0, 1\}$.
- ax horizontal: probablité $x = 0$
- $\max H(X) = log|\mathcal{X}|$

Comment on calcule l' entropie de variable cible, *surv*, en Titanic?

Comment on calcule l' entropie de variable cible, *surv*, en Titanic?

$$
\begin{aligned}
X &= \mathbf{surv} \\
p(\mathbf{surv}) &=
\end{aligned}
$$

Comment on calcule l' entropie de variable cible, *surv*, en Titanic?

$$
\begin{aligned}
X &= \mathbf{surv} \\
p(\mathbf{surv}) &= \{p(yes), p(no)\} \\
H(\mathbf{surv}) &=
\end{aligned}
$$

Comment on calcule l' entropie de variable cible, *surv*, en Titanic?

$$
\begin{aligned}
X &= \mathbf{surv} \\
p(\mathbf{surv}) &= \{p(yes), p(no)\} \\
H(\mathbf{surv}) &= -\left(\log(p(yes)) \times p(yes) + \log(p(no)) \times p(no)\right)
\end{aligned}
$$

Comment on fait dans R?

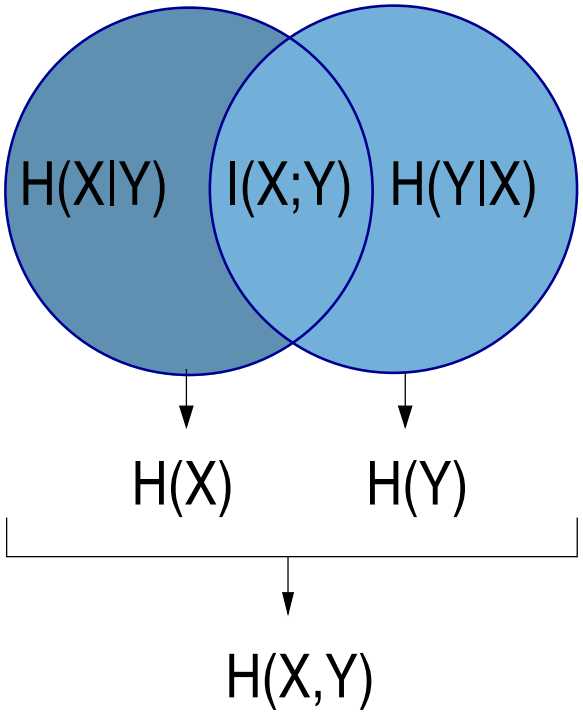
Entropie conditionnelle

- Soient 2 v.a. X et Y.
  L'entropie conditionnelle $H(Y|X)$

$$= \sum_{x \in \mathcal{X}} p(x) H(Y|X=x)$$

$$= -\sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \, log \, p(y|x)$$

- Pour que X serve à prédire Y, il faut
  $I(Y;X) = H(Y) - H(Y|X) > 0$
- On appel $I(Y;X)$ l' Information mutuelle des variables X, Y.



Comment on calcule l' entropie de variable cible *surv* etant donne le variable *sex* ?

Comment on calcule l' entropie de variable cible *surv* etant donne le variable *sex* ?

$$Y \quad =$$

Comment on calcule l' entropie de variable cible *surv* etant donne le variable *sex* ?

$$Y = \mathbf{surv}$$
$$X =$$

Comment on calcule l' entropie de variable cible *surv* etant donne le variable *sex* ?

$$
\begin{aligned}
Y &= \mathbf{surv} \\
X &= \mathbf{sex} \\
H(\mathbf{surv}|\mathbf{sex}) &=
\end{aligned}
$$

Comment on calcule l' entropie de variable cible *surv* etant donne le variable *sex* ?

$$
\begin{aligned}
Y &= \mathbf{surv} \\
X &= \mathbf{sex} \\
H(\mathbf{surv}|\mathbf{sex}) &= p(male)H(\mathbf{surv}|male) \\
&+ p(female)H(\mathbf{surv}|female) \\
H(\mathbf{surv}|male) &=
\end{aligned}
$$

Comment on calcule l' entropie de variable cible *surv* etant donne le variable *sex* ?

$$
\begin{aligned}
Y &= \mathbf{surv} \\
X &= \mathbf{sex} \\
H(\mathbf{surv}|\mathbf{sex}) &= p(male)H(\mathbf{surv}|male) \\
&+ p(female)H(\mathbf{surv}|female) \\
H(\mathbf{surv}|male) &= -(p(yes|male) \times \log(p(yes|male)) \\
&+ p(no|male) \times \log(p(no|male)) \\
H(\mathbf{surv}|female) &= \dots
\end{aligned}
$$

Information mutuelle et choix de la variable

- L'information mutuelle entre X et Y = quantité d'info sur Y apportée par la connaissance de X et vice-versa = **gain d'information**

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(X) - H(X|Y) \\ &= 0 \Leftrightarrow X \text{ et } Y \text{ indépendantes} \end{aligned}$$

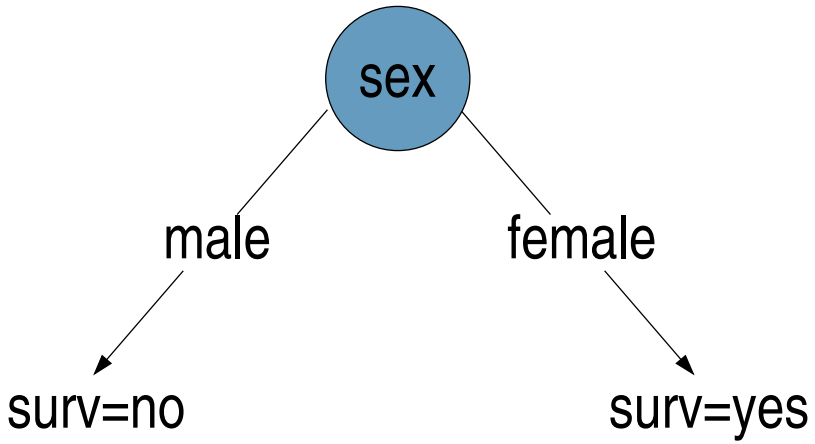
- Inconvénient : $I(X;Y) \nearrow$ avec $|\mathcal{X}|$: favorise les variables ayant bcp de valeurs distinctes
- Solution dans C5 : normaliser le gain d'info par l'entropie de la variable prédictive $X \rightarrow$ **rapport de gain** (critère par défaut)

$$IGR = \frac{I(X;Y)}{H(X)}$$

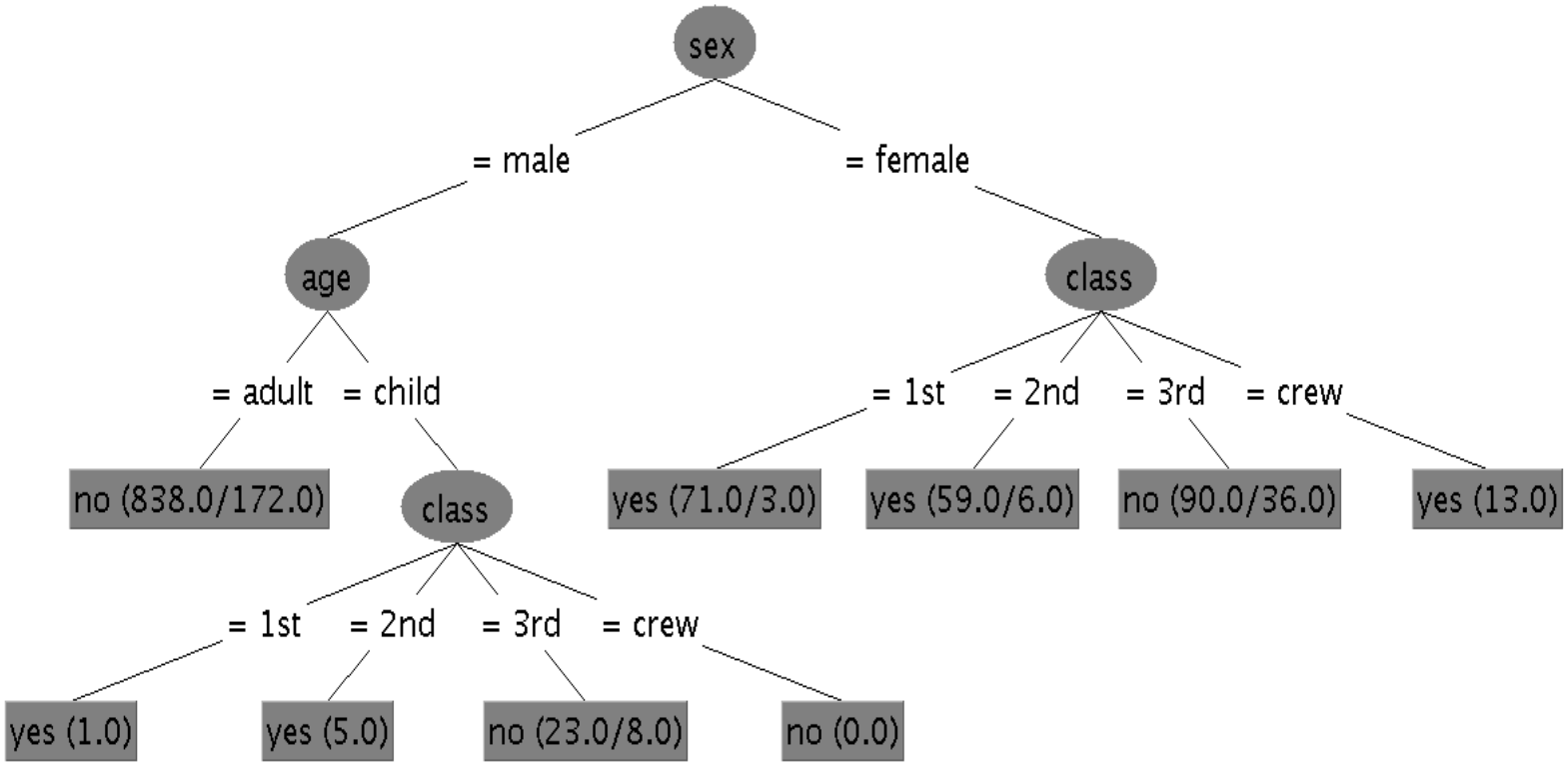
Exemple : choix de la variable racine

Sur l'ensemble d'apprentissage TRN

$$
\begin{aligned}
H(surv) &= 0.908 \\
I(class; surv) &= 0.052 \\
I(age; surv) &= 0.005 \\
I(sex, surv) &= 0.139
\end{aligned}
$$



Arbre C4.5 sur le Titanic



Paramètre m (nb min d'ex par feuille): contrôle la complexité

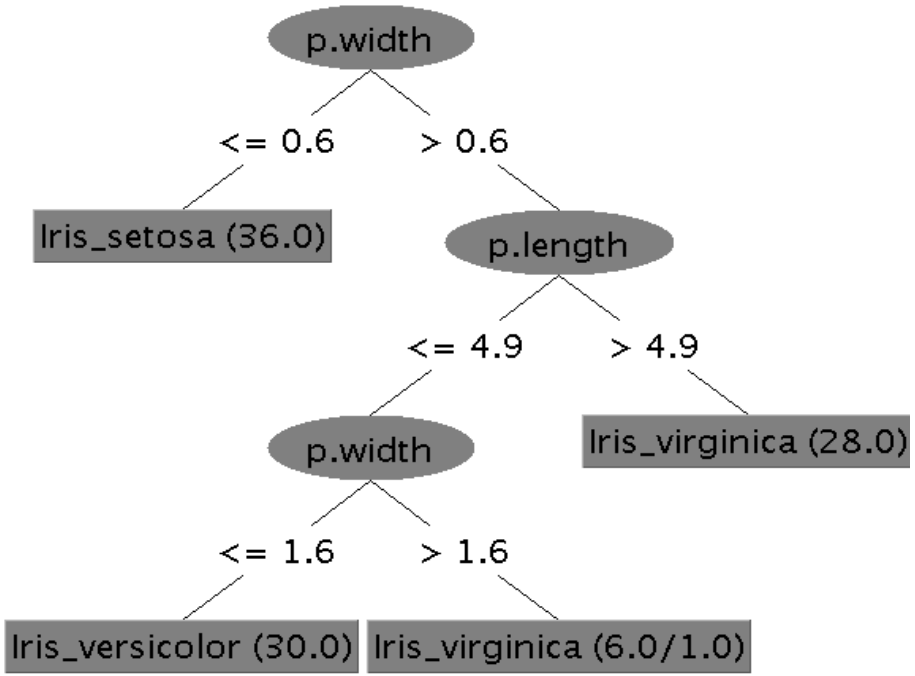
Ici: $m = 2$ $\neq$arbre p. 6: $m = 20$

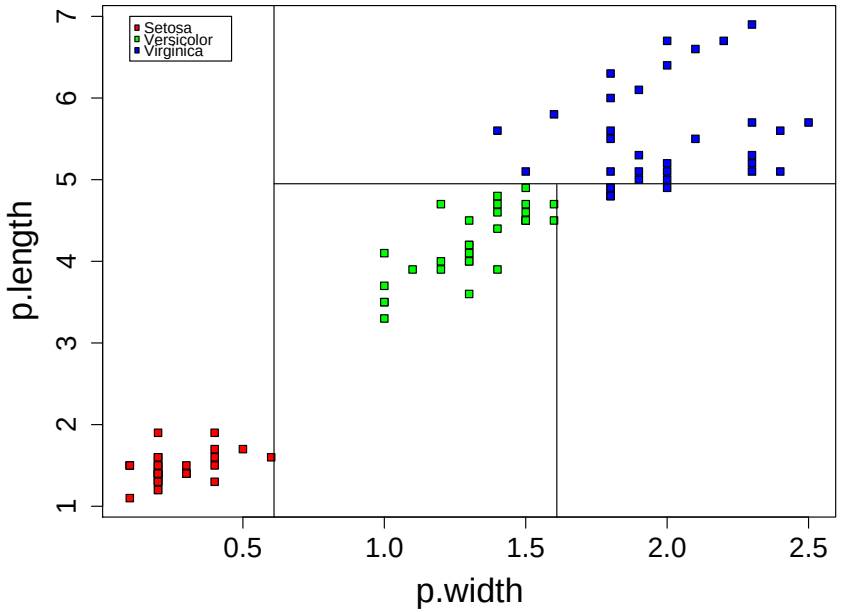
Partitionnement binaire sur les variables continues

Idée: transformer une variable continue X en variable booléenne: trouver un seuil t permettant d'avoir 2 groupes homogènes par rapport à la variable cible

1. Identifier les seuils potentiels parmi les valeurs distinctes x_i de X
   - Trier les exemples dans l'ordre croissant des x_i
   - Les seuils potentiels = x_i adjacents ayant des classes différentes
     Ex. iris: petal.length : 36 valeurs distinctes, mais 6 seuils potentiels
2. Choisir la partition qui maximise le critère choisi (gain d'info ou rapport de gain en C4.5)

Arbre C4.5 sur les iris





Post-élagage de l'arbre C4.5

Critère : réduction de l'erreur

- Partitionner les données en ensemble d'apprentissage TRN et ensemble de validation VAL
- Construire un arbre en utilisant TRN
- Convertir un noeud interne en feuille si son erreur sur VAL n'est pas supérieur à la somme d'erreur de ses fils

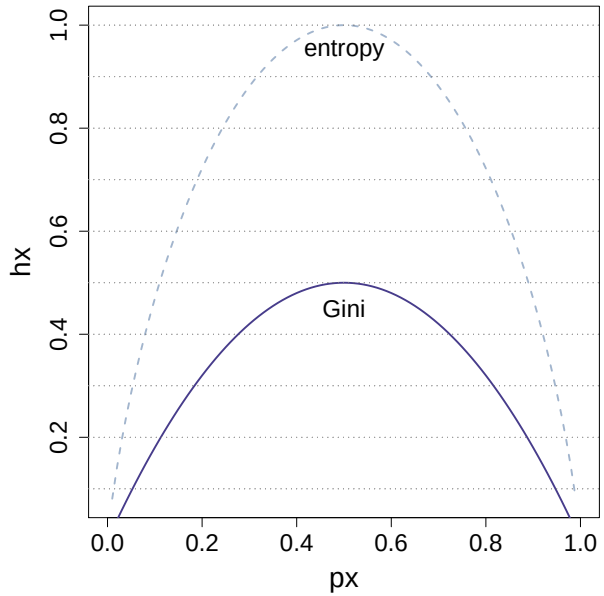
Plan

Mesure d'impureté d'une variable X discrete

- L'index Gini $G(X)$ mesure l'impureté de la variable discrete X prenant ses valeurs dans l'alphabet $\mathcal{X}$; il est calcule comme:

$$G(X) = \sum_{x \in \mathcal{X}} P(x)(1 - P(x)) \tag{1}$$

- $Min(G) = 0 \Leftrightarrow$ tous les examples ont la même valeur de la variable cible
- $Max(G) = 1 - \frac{1}{|\mathcal{X}|} \Leftrightarrow$toutes les valeurs x équiprobables



Mesure d'impureté d'une variable X continuous (régression)

- L'index Gini $R(X)$ mesure l'impureté de la variable continue X prenant ses valeurs dans ; il est calcule comme:

$$R(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)^2 \tag{2}$$

- μ_x est la moyenne des x_i
- $R(X)$ measure quoi????

Mesure d'impureté d'une variable X continuous (régression)

- L'index Gini $R(X)$ mesure l'impureté de la variable continue X prenant ses valeurs dans ; il est calcule comme:

$$R(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)^2 \tag{2}$$

- μ_x est la moyenne des x_i
- $R(X)$ measure quoi???? la variance de X.

Gini Conditionnelle, Y discrete (classification)

- Soient 2 v.a. discretes X et Y. Similaire a l' entropie conditionnelle le Gini conditionnelle est donne par:

$$G(Y|X) \quad =$$

Gini Conditionnelle, Y discrete (classification)

- Soient 2 v.a. discretes X et Y. Similaire a l' entropie conditionnelle le Gini conditionnelle est donne par:

$$\begin{aligned} G(Y|X) &= \sum_{x \in \mathcal{X}} p(x) G(Y|X=x) \\ &= \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x)(1-p(y|x)) \end{aligned}$$

- comme avec l' entropie confitionnelle pour que X serve à préedire Y, il faut que le *Gini gain*:

$$\Delta G(Y,X) = G(Y) - G(Y|X)$$

soit positif, c.a.d $\Delta G(Y,X) > 0$

Gini Conditionnelle, Y continuous (régression)

- Soient 2 v.a. X discrete et Y continuous. Comme dans les cas de classification le Gini conditionnelle est donne par::

$$\begin{aligned} R(Y|X) &= \sum_{x \in \mathcal{X}} p(x) R(Y|X=x) \\ &= \sum_{x \in \mathcal{X}} p(x) \frac{1}{n_x} \sum_{y_i | x_i = x} (y_i - \mu_{y_x})^2 \end{aligned}$$

- μ_{y_x} est le moyenne chez les instances pour lesquelles la variable X prend le valeur x.
- n_x c'est le nombre des instances qui prende le valeure x pour la variable X.
- A quoi correspond alors le term: $\frac{1}{n_x} \sum_{y_i | x_i = x} (y_i - \mu_{y_x})^2$?
- Pour que X ammenne une reduction de impurite de Y:

Gini Conditionnelle, Y continuous (régression)

- Soient 2 v.a. X discrete et Y continuous. Comme dans les cas de classification le Gini conditionnelle est donne par::

$$\begin{aligned} R(Y|X) &= \sum_{x \in \mathcal{X}} p(x) R(Y|X=x) \\ &= \sum_{x \in \mathcal{X}} p(x) \frac{1}{n_x} \sum_{y_i | x_i = x} (y_i - \mu_{y_x})^2 \end{aligned}$$

- μ_{y_x} est le moyenne chez les instances pour lesquelles la variable X prend le valeur x.
- n_x c'est le nombre des instances qui prende le valeure x pour la variable X.
- A quoi correspond alors le term: $\frac{1}{n_x} \sum_{y_i | x_i = x} (y_i - \mu_{y_x})^2$?
- Pour que X ammenne une reduction de impurite de Y:

$$\Delta R(Y, X) = R(Y) - R(Y|X) > 0$$

Choix de la variable X

- *Attention*: CART marche seleument quand la variable X prende *deux* valeurs. *Par contre* le gini index marche aussi avec des variables *discrete* qui prennent plusiers valeurs.
- Comment CART fait?

Choix de la variable X

- *Attention*: CART marche seleument quand la variable X prende *deux* valeurs. *Par contre* le gini index marche aussi avec des variables *discrete* qui prennent plusiers valeurs.
- Comment CART fait?
- Pour chaque variable candidate X (dom. $\mathcal{X}$) en lice au noeud t
  - calculer $\Delta G|\Delta R$ pour tous les tests binaires possibles

    var continue $x \leq t\,?$ ($t =$seuil potentiel, voir p. 13)
    var discrète $x \in \mathcal{X}\prime \subset \mathcal{X}\,?$
  - choisir la dichotomie qui maximise $\Delta G|\Delta R$
- Choisir la variable concernée par la dichotomie ayant $\Delta G|\Delta R$ maximale.

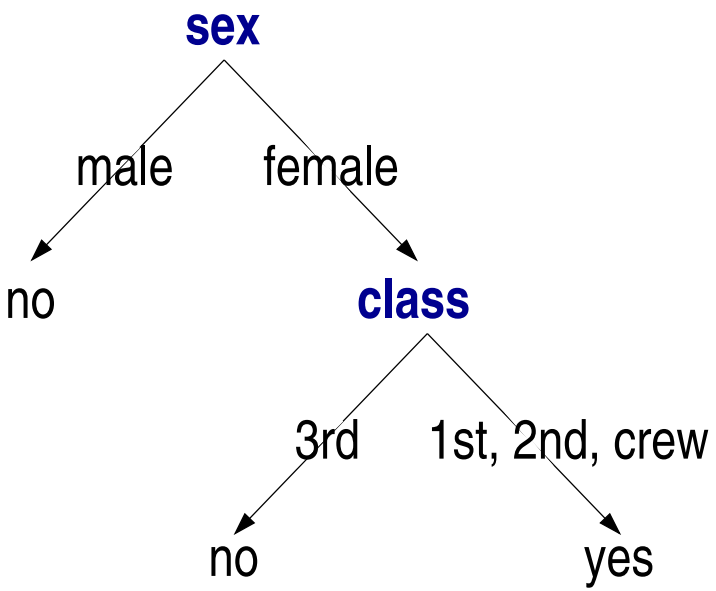
Post-élagage de l'arbre CART

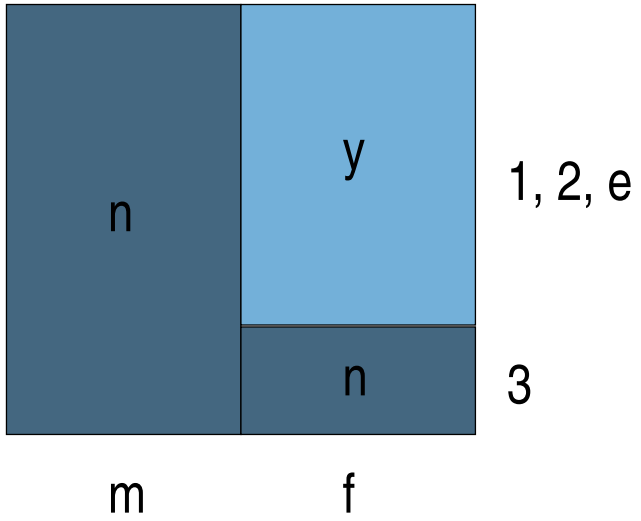
- Critère : compromis entre erreur et complexité d'un arbre T :

$$C_\lambda(T) = E(T) + \lambda|T|$$

- $E(T)$ = erreur totale (TBC ou MSE) aux feuilles de T
- $|T|$ = nombre de feuilles de l'arbre T: mesure de la complexité du modèle
- λ = un paramètre de **régularisation** qui contrôle le compromis entre performance et complexité
- la valeur de λ est déterminée empiriquement (validation croisée, à voir plus tard)

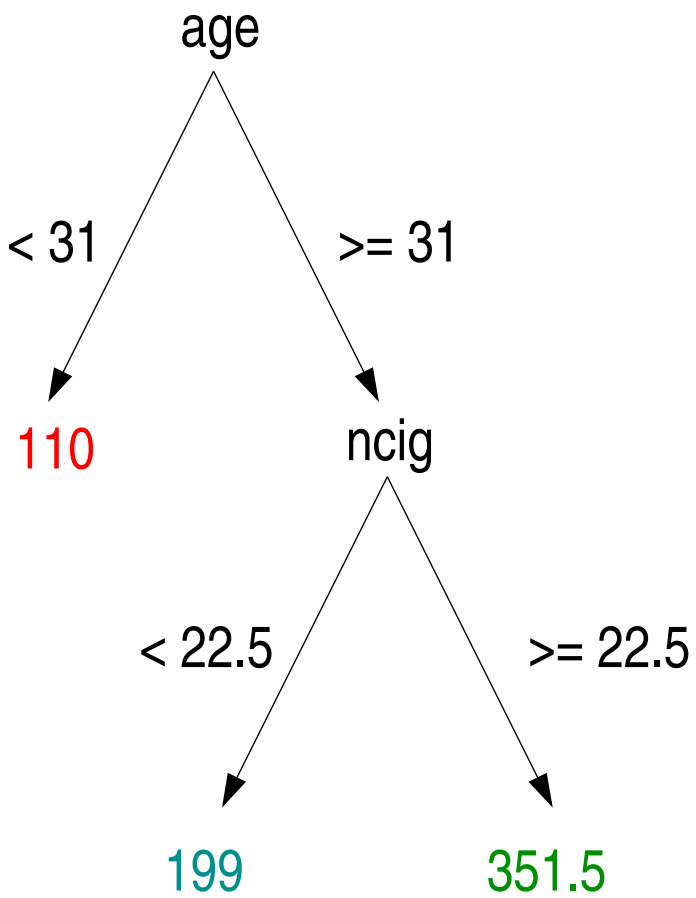
Arbre de classification CART

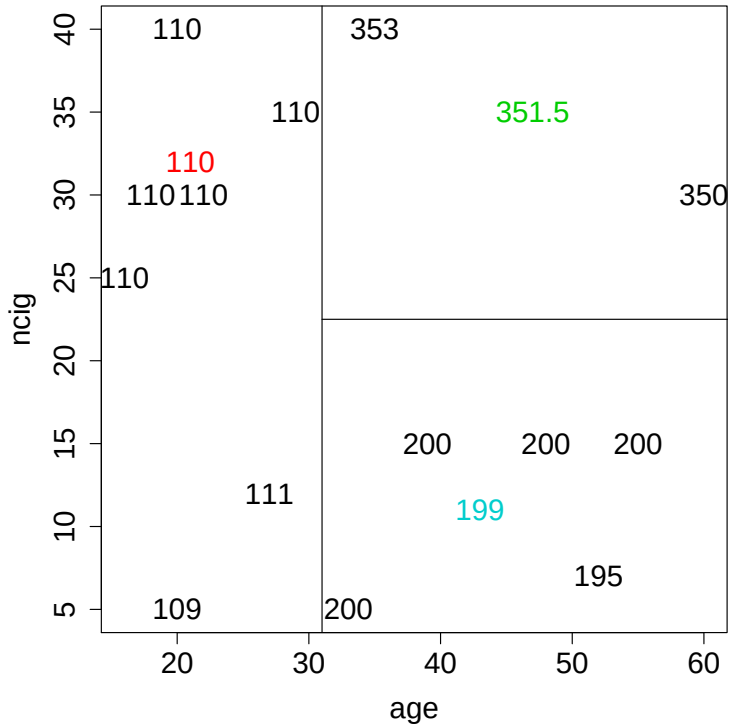




$m = 20$. cf. l'arbre C4.5 page 6

Arbre de régression CART





Plan

Principes de base de l'évaluation

Pour évaluer l'efficacité d'un prédicteur, il faut

1. une mesure de performance
   - classification : $TBC = \frac{\text{nb de cas bien classés}}{\text{nb total de cas}}$
   - régression : $MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$
2. une référence de base (baseline)
   - le classifieur par défaut : prédit toujours la classe majoritaire
   - le régresseur par défaut : prédit toujours la moyenne de la var. cible
3. un ensemble de test indépendant de l'ensemble d'apprentissage

Répartition des données

- Pour éviter le surapprentissage, nous répartissons les données en 2 sous-ensembles (proportions suivant la taille des données)
  - l'ensemble d'entraînement (TRN) : servira à construire nos modèles
  - l'ensemble de test (TST) : servira à valider les modèles construits
- Les partitions doivent être stratifées : doivent conserver la distribution d'origine de la variable cible

	tout	trn	tst	%
surv=oui	711	356	355	32
surv=non	1490	744	746	68
total	2201	1100	1101	100

Evaluation des modèles sur le Titanic

Classification			
TBC	TRN [n=1100]	TST [n=1101]	CPX
Baseline	68%	68%	
J48, m=2	79.5%	78.564%	|T|=9
J48, m=20	79.0%	77.66%	|T|=5
CART, m=20	79.0%	77.66%	|T|=3
Régression			
MSE	TRN [n=14]	TST [n=6]	CPX
Baseline	6768.78	11749.89	
CART, m=5	1.89	3809.25	|T|=3

Bilan

- Avantages
  - apprentissage très rapide
  - compréhensibilité du modèle
  - robustesse aux variables non pertinentes
- Inconvénients
  - instabilité : très sensible aux variations des données
  - incapacité à détecter les interactions entre variables
  - puissance de représentation assez limitée : découpes orthogonales trop peu adaptées aux problèmes demandant des frontières obliques et lisses