Compléments sur l'algorithme CART implémenté dans rpart

Par défaut :

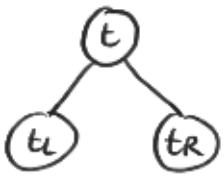
- matrice de coût 0-1
- mesure d'impureté pour le choix de la question binaire est Gini

} argument parms

- Un noeud n'est pas divisé si :

  $n_t < 20$ (minsplit)

  n_L ou $n_R < \text{round}\left(\frac{20}{3}\right) = 6$ (minbucket)



$$cp(t) = \frac{Loss(t) - Loss(t_L) - Loss(t_R)}{Loss(t_1)}$$

t_1 : noeud racine

= proportion d'erreur en moins due à la division de t

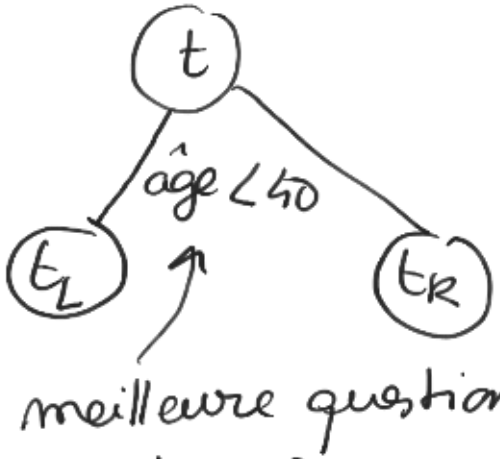
< 0.01 (cp)

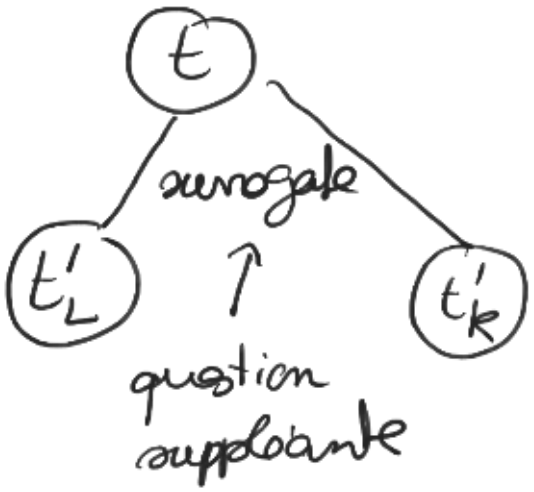
- Validation croisée 10-fold pour l'estimation des erreurs des sous-arbres (xval = 10)

} arguments Control

Gestion des données manquants

Surrogate = suppléant





	t'_L	t'_R
t_L	a	b
t_R	c	d

⇒ Trouver la question suppléante (surrogate) qui donne les mêmes sous-noeuds donc qui minimise $\frac{c+b}{n_t}$

⇒ Si une nouvelle donnée arrive pour laquelle l'âge est absent, c'est la question surrogate qui est utilisée.

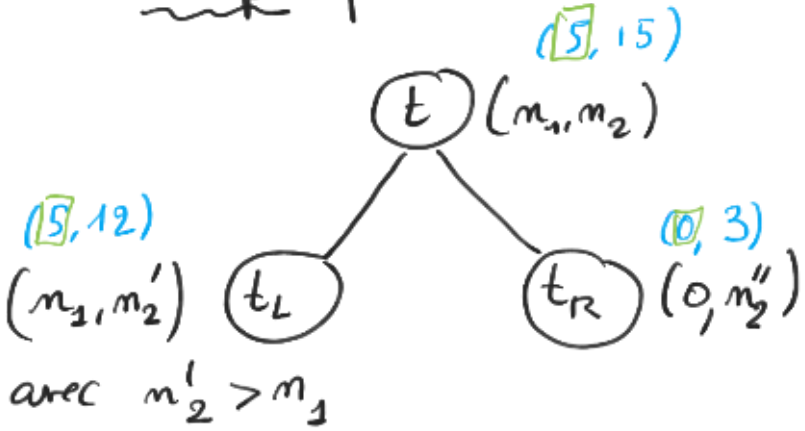
Cas où $C_p = 0$

- $cp(t) = 0$ par convention si t est une feuille
- Sinon on a défini $cp(t) = Loss(t) - Loss(t_L) - Loss(t_R)$

(nb de mals classés)

Le seul cas (a priori) où $cp = 0$ est le cas où tous les mals classés du noeud t vont dans un même sous-noeud où ils ne sont pas majoritaires

Exemple pour Y binaire:



avec $n_2' > n_1$

$\Rightarrow cp(t) = 0$

⚠ Dans ce cas, une valeur de $cp(t) \neq 0$ est affectée au noeud. Cette valeur est liée à la définition de la suite de paramètres de complexités α associés aux arbres élagués optimaux.

On reprend le point 2. de la procédure de définition de ces paramètres de complexités :

Soit T_L un arbre à L feuille et α_L la valeur du paramètre de complexité à partir duquel T_L minimise $C_\alpha(T)$.

T_L minimise $C_\alpha(T)$ tant que

$$C_\alpha(T_L) < C_\alpha(T_{L'})$$

où $T_{L'}$ est un arbre à L' feuilles et $L' < L$.

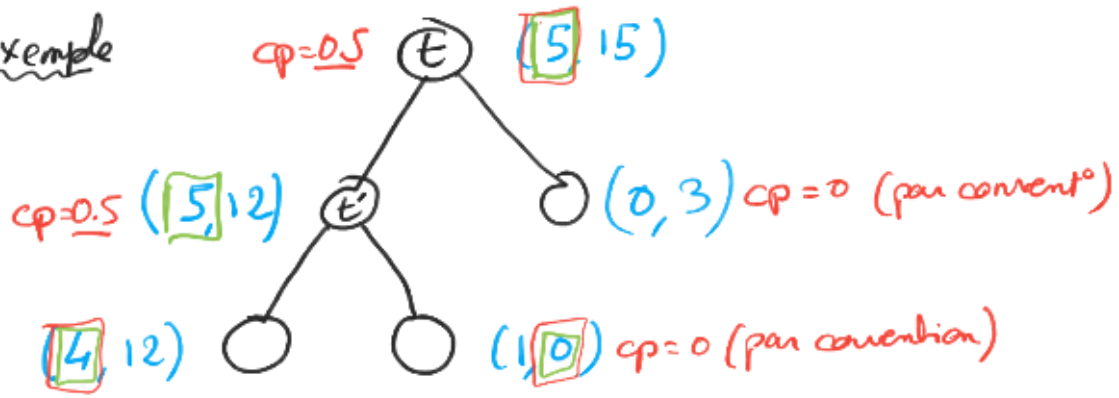
Cette condition s'écrit également

$$n\hat{R}(T_L) + \alpha L < n\hat{R}(T_{L'}) + \alpha L'$$

$$\Leftrightarrow \quad \alpha < \frac{n\hat{R}(T_{L-1}) - n\hat{R}(T_{L'})}{L - L'}$$

$$\Rightarrow \quad \alpha_{L'} = \frac{\text{Reduction du nombre de mals classés}}{\text{nombre de feuilles ajoutées (divisions)}}$$

Exemple



$$cp(t) = cp(t) = \frac{5 - 4 - 0}{2} = 0.5$$

⚠ dans Rpart : $cp(t) = \dfrac{Loss(t) - Loss(t_R) - Loss(T_L)}{Loss(t_1)}$

↑ noeud racine.

Notion de question concurrente (competitor)

A chaque division d'un noeud de l'arbre, c'est la meilleure question binaire qui est sélectionnée c.a.d celle qui maximise la reduction de l'impureté

Une question binaire concurrente (compétiteur) à la meilleure question retenue, donne une réduction de l'"impureté" proche de la réduction optimale".

La première question concurrente est la seconde question qui maximise la réduction de l'impureté

Etc...

Cela donne une idée des autres variables qui pourraient être importantes pour diviser ce noeud.