Correction de l'exercice 1 du TP sur CART

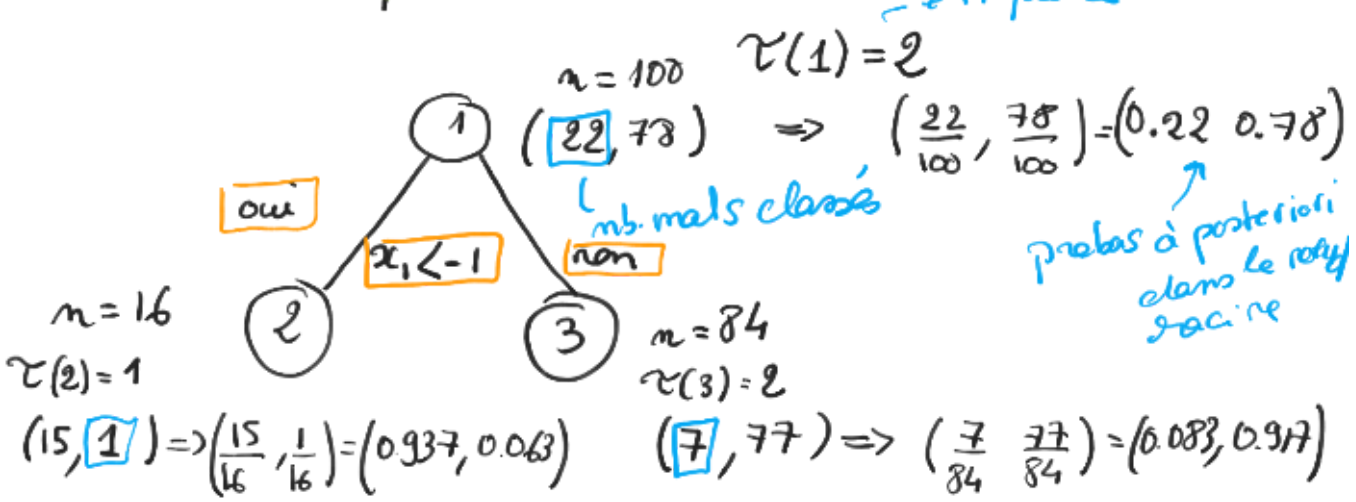
Données synthétiques : $\begin{cases} Y \in \{1,2\} \\ (X^1, X^2) \in \mathbb{R}^2 \end{cases}$

Données d'apprentissage pour la construction de l'arbre :

$\begin{cases} n = 100 \text{ observations} \\ n_1 = 22 \quad (\text{nb d'obs. avec } Y=1) \\ n_2 = 78 \quad (\text{---------- } Y=2) \end{cases}$

① Construction de l'arbre avec les valeurs par défaut de la fonction rpart

Sorties de print(tree)



Le taux d'erreur d'apprentissage de cet arbre est : $\frac{7+1}{100} \Rightarrow$ 8% d'erreur d'apprentissage

Sorties de summary(tree)

Pour le noeud racine (noeud n°1) :

complexity parameter = $cp(1)$

$$= \frac{Loss(1) - Loss(2) - Loss(3)}{Loss(1)}$$

$$= \frac{22 - 1 - 7}{22} = 0.6363$$

$$P(node) = P(1) = \frac{100 - \text{nb. d'obs. dans le noeud}}{100 - n} = 1$$

Predicted class = $\hat{Y}(1) = 2$ (classe majoritaire)

Expected Loss = coût moyen du noeud

$$= \hat{R}(t)$$

$$= \frac{\text{nb mals classés dans } t \text{ (coûts 0-1)}}{n_t}$$

$$= \frac{22}{100} = 0.22 \text{ ici}$$

Improve = diminution de l'impureté (avec Gini ici)

$$= i(t) - \frac{n_L}{n_t} i(t_L) - \frac{n_R}{n_t} i(t_R)$$

$$i(t) = 1 - \sum_{k=1}^{K} p(k/t)^2 \quad \text{(Gini)}$$

Ici :

$$\text{Improve} = (1 - 0.22^2 - 0.78^2) - \frac{16}{100}(1 - 0.938^2 - 0.062^2)$$

$$- \frac{84}{100}(1 - 0.083^2 - 0.917^2)$$

$$\approx 0.1967$$

Soit improve $\approx 19.6\%$

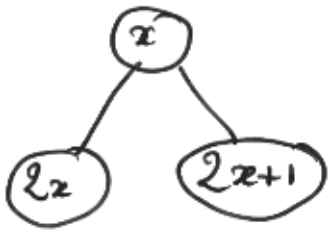
Arrêt des divisions :

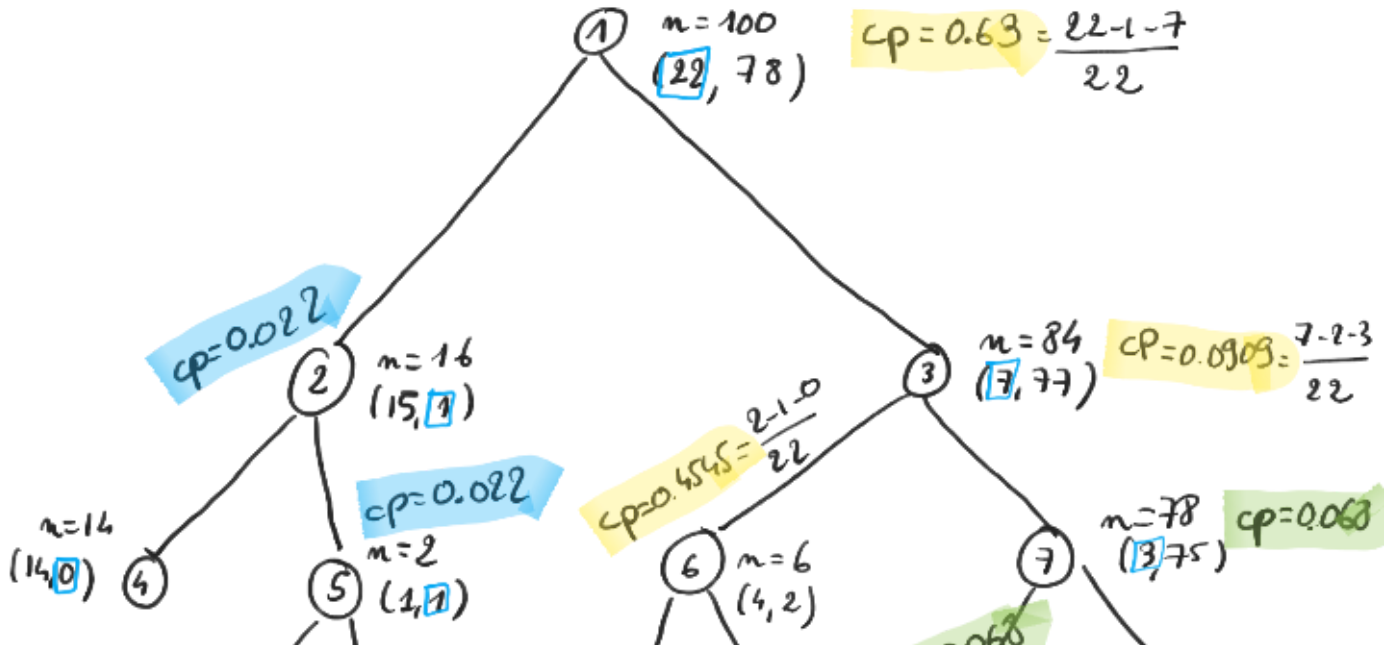
* noeud 2 est une feuille car il contient 16 observations et $n_{min} = 20$ par défaut
* noeud 3 est une feuille car probablement $cp(3) < 0.01$ (valeur par défaut pour arbr.)

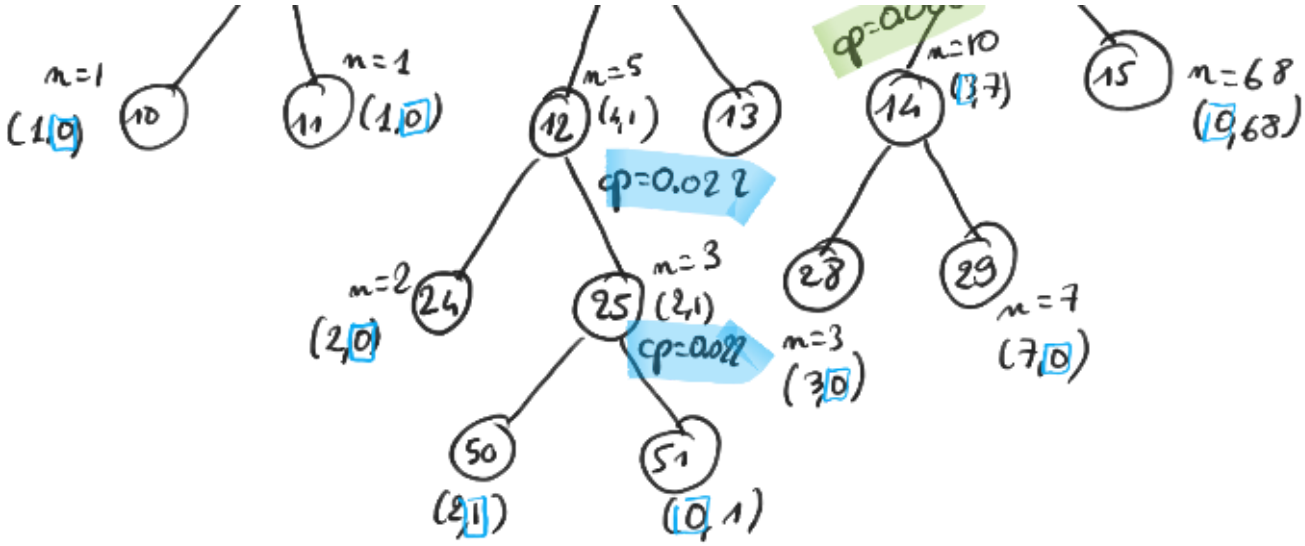
(2) Construction de l'arbre de longueur maximale suivie d'une étape d'élagage

Sortie de summary pour tracer l'arbre maximale

⚠ numérotation des noeud :







CP = 0.022 = $\frac{1-0-1}{22 \times 2}$: on élague directement { 2 et 5 / 25 et 12

=> ↓ (2) feuilles à chaque fois

CP = 0.068 = $\frac{3-0-0}{22 \times 2}$: on élague directement 7 et 14

=> ↓ de (2) feuilles

Sortie de printcp

	CP	nsplit	rel error (erreur relative de l'arbre)	xerror	xstd
Six entrées / arbres	0.6363	0	22/22 = 1		
	0.0909	1	(7+1)/22 = 0.36	⚠	⚠
	0.0681	2	(1+2+3)/22 = 0.27		
	0.0545	4	.	les valeurs de ces colonnes peuvent varier. Elles dépendent des découpages x-fold.	
	0.0227	5 divisions => 6 feuilles	.		
	0	9 divisions => 10 feuilles	.		

Remarque : Pour passer de l'arbre à 10 feuilles (arbre maximal

a l'arbre a 6 feuilles, tous les nœuds ayant une valeur de cp = 0.0227 ont été élagués (nœuds 2, 5, 12, 25)
→ 4 divisions ont été supprimées.

xerror = erreur relative moyenne de validation croisée x-plds de l'arbre élagué

xstd = écart-type des erreurs relatives de validation croisé de l'arbre élagué

Graphique de plotcp

Ce graphique est obtenu avec les résultats de printcp
En abscisse : paramètre de complexité calculés lors de la procédure de validation croisée

$\beta_j = \sqrt{\alpha_j \, \alpha_{j-1}}$ ⇒ caractéristique de $[\alpha_j, \alpha_{j-1}[$

Soit :

$\alpha_1 = 0.6363$	$\beta_1 = +\infty$
$\alpha_2 = 0.0909$	$\beta_2 = \sqrt{0.0909 \times 0.6363} = 0.2405$
$\alpha_3 = 0.068$	$\beta_3 = \sqrt{0.068 \times 0.0909} = 0.079$
$\alpha_4 = 0.04545$	$\beta_3 = 0.056$
$\alpha_5 = 0.0227$	$\beta_4 = 0.032$
$\alpha_6 = 0$	$\beta_5 = 0$

Size of tree : chaque valeur β_j (en abscisse) correspond à un sous-arbre élagué. L'axe en haut donne le nombre de feuilles de ce sous-arbre

le nombre de feuille de ces sous-arbres

En ordonnée : l'ordonnée indique l'erreur relative de validation croisée 10-fold de chaque sous-arbre (relativement à l'erreur du noeud racine). Les points sont les moyennes des erreurs relatives des 10 folds : colonne xerror (moyenne)
Les intervalles correspondent à ± 1 écart type des 10 erreurs relatives : colonne xstd (écart-type)
Ligne pointillée : ligne $y = S$ où en posant

$$j^* = \arg\min_{\beta_j} xerror_j$$

$$S = xerror_{j^*} + xstd_{j^*}$$

Choix du paramètre de complexité pour élaguer

Cas 1 : On choisit β_{j^*} qui minimise l'erreur

Cas 2 : On choisit la plus grande valeur de β_j pour laquelle $\beta_j \leq S$

C'est la règle des 1-SE. On prend la valeur du cp qui diminue le + possible la complexité de l'arbre tout en conservant une erreur à moins de 1 écart-type de l'erreur "optimale".