

4 Première brique en Machine Learning : arbres de décision

- Algorithme CART

- Exemples
- Formalisation : construction de l'arbre
- Lien avec le problème de régression classique
- Gestion du surapprentissage : réduction de dimension
- Réponse catégorielle
- Outils et mesures de performance des modèles
- Extensions et conclusion

ALGO ISSU DE L'INTELLIGENCE ARTIFICIELLE

7 principes éthiques de la Commission Européenne pour l'IA :

- 1 contrôle/supervision humaine : l'IA n'a pas de conscience !
- 2 résistance et sécurité des algorithmes : fiabilité pour gérer les erreurs et incohérences ;
- 3 gestion des données, protection de la vie privée : utilisateurs en mesure de contrôler leurs propres données ;
- 4 transparence algo : expliquer ce que fait l'IA, traçabilité
- 5 diversité, non-discrimination et équité ;
- 6 bien-être social et environnemental : l'IA doit être mise au service de la société dans son ensemble ;
- 7 l'"accountability" : principe de responsabilité, mise en place de procédures internes à l'entreprise pour démontrer le respect des règles relatives à la protection des données.

OBJECTIF ARBRE : CLASSIF. DES INDIVIDUS

Regrouper des indiv. hétérogènes en classes homogènes de risque pour résumer l'info d'une BdD gigantesque.

∃ de nombreuses techniques de classification, parmi lesquelles :

- pour la classification **non-supervisée** :
 - les algorithmes dits des k -plus proches voisins (non param.) ;
 - les techniques ascendantes d'arbre de classification (CAH) ;
 - model-based clustering (paramétrique) ;
- pour la classification **supervisée** :
 - modèles paramétrique de choix (LOGIT) ;
 - réseaux de neurones ; SVM (non paramétrique) ;
 - arbres descendants (**CART**, CHAID, ...). Non param.

ARBRE ET CLUSTERING : PREMIERS ÉLÉMENTS

Pour estimer notre quantité d'intérêt, on choisit d'utiliser un arbre...

Mais qu'est-ce qu'un arbre ?

- 1 Une **racine** : contient l'ensemble de la population à segmenter (le portefeuille global) \Rightarrow c'est le point de départ ;
- 2 Un **tronc** et des **branches** : contiennent les règles de division qui permettent de segmenter la population ;
- 3 Des **feuilles** : contiennent les sous-populations homogènes (sur leurs caractéristiques et la réponse) créées, fournissent l'estimation de la quantité d'intérêt.

RÈGLES ET LECTURE D'UN ARBRE CART

Un arbre de classification / régression se lit de la racine vers les feuilles (l'inverse d'une CAH...).

A chaque ramification, une règle de division apparaît : dans CART,

- cette règle (\simeq question) admet une réponse binaire (oui/non),
- elle n'est basée que sur un facteur de risque (une covariable).

Un noeud est l'intersection d'un ensemble de règles. **L'estimation de la quantité d'intérêt se lit dans les noeuds terminaux (feuilles).**

N'importe quel individu de la population initiale appartient à une unique feuille : les **sous-populations** créées sont **disjointes**.

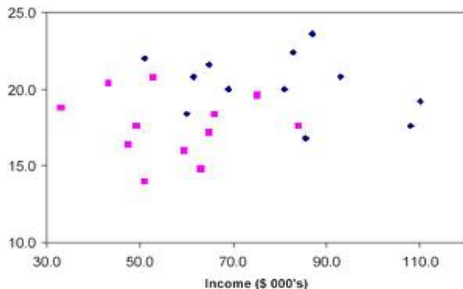
4 Première brique en Machine Learning : arbres de décision

- Algorithme CART
- Exemples
 - Formalisation : construction de l'arbre
 - Lien avec le problème de régression classique
 - Gestion du surapprentissage : réduction de dimension
 - Réponse catégorielle
 - Outils et mesures de performance des modèles
 - Extensions et conclusion

EXEMPLE 1 : ARBRE DE CLASSIFICATION

A travers cet exemple, on veut **intuire comment un arbre se construit...** Cherchons à prévoir “propriétaire” | salaire + surface.

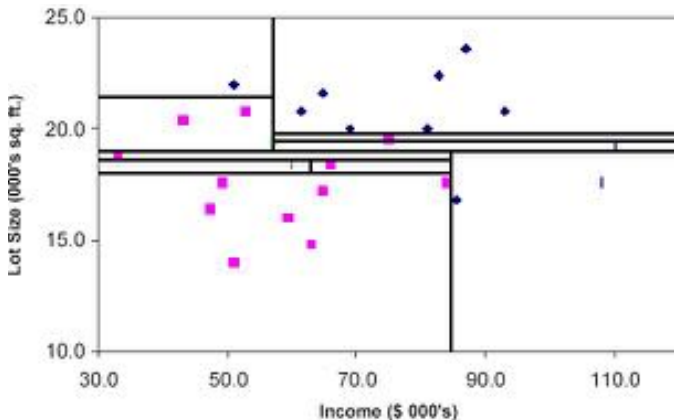
Income (\$ 000's)	Lot Size (000's sq. ft.)	Owners=1, Non-owners=2
60	18.4	1
85.5	16.8	1
64.8	21.6	1
61.5	20.8	1
87	23.6	1
110.1	19.2	1
108	17.6	1
82.8	22.4	1
69	20	1
93	20.8	1
51	22	1
81	20	1
75	19.6	2
52.8	20.8	2
64.8	17.2	2
43.2	20.4	2
84	17.6	2
49.2	17.6	2
59.4	16	2
66	18.4	2
47.4	16.4	2
33	18.8	2
51	14	2
63	14.8	2



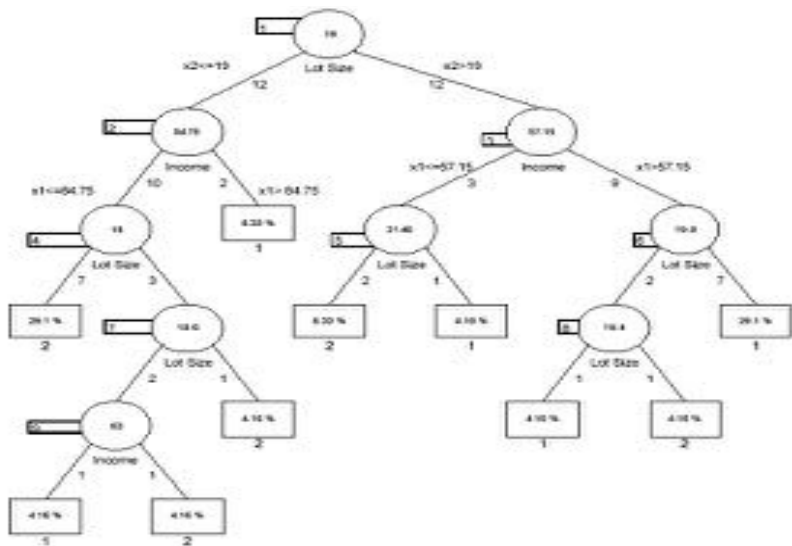
CHOISIR LA SEGMENTATION DE L'ESPACE

- ① Choisir une var. explicative j donnée à m valeurs : soit elle est
 - numérique ou catégorielle ordonnée : partitionnements de l'espace associé à cette covariable se situent entre 2 de ses valeurs successives observées $\Rightarrow m - 1$ possibilités ;
 - catégorielle non ordonnée : partitionnements de χ_j sont toutes les combinaisons de modalités, au nb de $2^m - 1$;
- ② Je teste tous ces partitionnements : j'y associe un critère d'homogénéité par rapport à ma quantité d'intérêt (réponse) ;
- ③ Je choisis le partitionnement qui conduit à la **plus grande homogénéité** dans les sous-espaces créés ;
- ④ Je répète les étapes (1)-(3) pour chacune des covariables dont je dispose : j'obtiens une **liste de k homogénéités max.** ;
- ⑤ Je choisis à la fin la covariable et son partitionnement qui **maximise l'homogénéité globalement.**

PARTITIONNEMENT ET ARBRE MAXIMAL



Partitionnement qui maximise l'homogénéité dans chq rectangle.



4 Première brique en Machine Learning : arbres de décision

- Algorithme CART
- Exemples
- **Formalisation : construction de l'arbre**
- Lien avec le problème de régression classique
- Gestion du surapprentissage : réduction de dimension
- Réponse catégorielle
- Outils et mesures de performance des modèles
- Extensions et conclusion

NOTATIONS

- $i \in \llbracket 1, n \rrbracket$: identifiant de l'individu / l'assuré ;
- $j \in \llbracket 1, k \rrbracket$: identifiant du facteur de risque (continu ou discret) ;
- Y_i : réponse **OBSERVEE** du $i^{\text{ème}}$ individu (continue/discrète) ;
- $\mathbf{X}_i = (X_{i1}, \dots, X_{ik})$: vecteur des facteurs de risque de l'indiv. i ;
- \mathcal{X} : espace des covariables (facteurs de risque) ;
- $l \in \llbracket 1, L \rrbracket$: identifiant des feuilles de l'arbre ;
- \mathcal{X}_l : ensemble de la partition correspondant à la feuille l .

ARBRE DE RÉGRESSION AVEC Y CONTINUE

En régression, la quantité d'intérêt est

$$\pi_0(\mathbf{x}) = E_0[Y | \mathbf{X} = \mathbf{x}] \quad (1)$$

En supposant une relation lin. (se restreignant à une classe d'estimateurs), on a

$$\hat{\pi}(\mathbf{x}) = \hat{\beta}_0 + \mathbf{x}^T \hat{\beta},$$

et on estime les paramètres de régression par MCO.

En toute généralité, on ne peut pas considérer ts les estimateurs potentiels de $\pi_0(\mathbf{x}) \Rightarrow$ arbres sont **1 autre classe d'estimateurs** : ce sont des **fonct. constantes par morceaux**.

Construire un arbre maximal génère une suite d'estimateurs selon une procédure spécifique : divisions successives de l'espace \mathcal{X} .

CONSTRUCTION DE L'ARBRE : CRITÈRE DE DIVISION

La ramification de l'arbre est **basée sur la définition d'un critère d'homogénéité**, cohérent avec l'estimation de la quantité d'intérêt.

Dans l'estimation de (1), *MCO* tjs utilisé car solution donnée par

$$\pi_0(\mathbf{x}) = \arg \min_{\pi(\mathbf{x})} E_0[\Phi(Y, \pi(\mathbf{x})) | \mathbf{X} = \mathbf{x}], \quad (2)$$

où $\Phi(Y, \pi(\mathbf{x})) = (Y - \pi(\mathbf{x}))^2$.

La fonction de perte Φ correspond donc à l'**erreur quadratique** (fn. convexe), et le critère est la **minimisation de l'EQM**.

La \neq est ici que l'on va estimer $\pi_0(\mathbf{x})$ en **plusieurs étapes** !

ETAPES DE CONSTRUCTION DE L'ARBRE

On résume donc l'enchaînement des étapes de construction de l'arbre :

- ➊ on part de la racine ;
- ➋ on cherche la meilleure première segmentation (donnant le meilleur gain d'homogénéité) ;
- ➌ on segmente ;
- ➍ on itère sur chacun des 2 noeuds fils ;
- ➎ on itère sur les fils des noeuds fils, et ainsi de suite...

Par construction l'hétérogénéité diminue à chaque segmentation, pour atteindre sa valeur minimale sur l'arbre maximal.

4 Première brique en Machine Learning : arbres de décision

- Algorithme CART
- Exemples
- Formalisation : construction de l'arbre
- **Lien avec le problème de régression classique**
- Gestion du surapprentissage : réduction de dimension
- Réponse catégorielle
- Outils et mesures de performance des modèles
- Extensions et conclusion

LIEN ENTRE RÉGRESSION ET ARBRE

Arbre = ensemble de règles. Pour chaque noeud m , une règle R_m est associée à un sous-ensemble $\mathcal{X}_m \subseteq \mathcal{X}$.

Notation : dans la suite, $E_n[Y]$ désigne la moyenne empirique de Y , et $\mathcal{X}_{pa(m)}$ est le sous-ensemble associé au noeud parent de m .

L'arbre est associé à la fonction de régression

$$\hat{\pi}(\mathbf{x}) = \sum_{m=1}^M \hat{\beta}_m^{tree} R_m(\mathbf{x}) \quad (3)$$

où $\hat{\beta}_m^{tree} = E_n[Y | \mathbf{x} \in \mathcal{X}_m] - E_n[Y | \mathbf{x} \in \mathcal{X}_{pa(m)}]$ si $m \neq \text{racine}$,
 $\hat{\beta}_m^{tree} = E_n[Y]$ sinon.

Cela équivaut en régression classique à chercher

$$\hat{\beta}^{tree} = \arg \min_{\beta^{tree}} E_n \left[\left(Y - \sum \beta_m^{tree} R_m(\mathbf{x}) \right)^2 \right].$$

Depuis (3), en \sum sur ts les noeuds, il reste les feuilles... :

$$\hat{\pi}(\mathbf{x}) := \hat{\pi}^L(\mathbf{x}) = \sum_{l=1}^L \hat{\gamma}_l R_l(\mathbf{x}) \quad (4)$$

\Rightarrow Décomposition en bases fonctionnelles de $\mathbf{x} \Rightarrow$ **non-param** !

- L est le nombre de **feuilles** de l'arbre, l leur indice,
- $R_l(\mathbf{x}) = \mathbb{1}(\mathbf{x} \in \mathcal{X}_l)$: règle d'appartenance au ss-ensemble \mathcal{X}_l ,
- $\hat{\gamma}_l = E_n[Y | \mathbf{x} \in \mathcal{X}_l]$: moyenne empirique de Y dans la feuille l ,
- Ss-ensembles $\mathcal{X}_l \subseteq \mathcal{X}$ disjoints ($\mathcal{X}_l \cap \mathcal{X}_{l'} = \emptyset, l \neq l'$) et exhaustifs ($\mathcal{X} = \cup_l \mathcal{X}_l$).

(4) généralisable qlq soit la quantité d'intérêt. Ainsi, **tout arbre peut être vu comme un estimateur par morceaux.**

→ Interprétation :

- chaque morceau est une feuille, dont la valeur est la moyenne empirique des valeurs de Y de cette feuille (cas quantitatif),
- chq div. d'1 noeud t minimise la \sum variances intra-noeuds résultantes \Rightarrow **max.** \searrow hétérogénéité $H_t = 1/|t| \sum_{i \in t} (y_i - \bar{y}_t)^2$:

$$\max_{div.} (H_t - (H_{t_g} + H_{t_d})) \Leftrightarrow \min \left(\frac{|t_g|}{n} \sum_{i \in t_g} (y_i - \bar{y}_{t_g})^2 + \frac{|t_d|}{n} \sum_{i \in t_d} (y_i - \bar{y}_{t_d})^2 \right)$$

où t_g et t_d désignent respectivement les fils gauche et droite du noeud parent t .

La construction étant **récursive**, on génère une suite d'estimateurs depuis le nd racine : soit une suite $\{\Pi^K\}$ de ss-espaces t.q. $\Pi^K \subseteq \Pi$,

$$\Pi^K = \left\{ \pi^L(.) = \sum_{l=1}^L \gamma_l R_l(.) : L \in \mathbb{N}^*, L \leq K \right\}. \quad (5)$$

A K fixé, on cherche $\pi_0^K(\mathbf{x}) = \arg \min_{\pi(\mathbf{x}) \in \Pi^K} E_0[\Phi(Y, \pi(\mathbf{x})) | \mathbf{X} = \mathbf{x}]$.

\Rightarrow **Version empirique** $\hat{\pi}^K : \hat{\pi}^K(\mathbf{x}) = \arg \min_{\pi(\mathbf{x}) \in \Pi^K} E_n[\Phi(Y, \pi(\mathbf{x}))]$. Ou :

$$\boxed{\hat{\pi}^K(\mathbf{x}) = \arg \min_{\gamma=(\gamma_1, \dots, \gamma_L)} E_n[\Phi(Y, \pi^L(\mathbf{x}))]}. \quad (6)$$

CART ne cherche pas ts les estimateurs possibles avec $L \leq K$:
 approche ce minimum petit à petit.

ARRÊT DE LA PROCÉDURE DE SEGMENTATION

Comme déjà évoqué, l'algorithme CART **ne fixe pas de règle d'arrêt arbitraire** pour la procédure de division de l'espace.

L'algorithme arrête ainsi de diviser les feuilles quand :

- il n'y a qu'une observation dans la feuille, ou
- les individus de la feuille ont les mêmes valeurs de facteurs de risque (covariables **X**).

On construit ainsi l'*arbre "maximal"*, qui sera ensuite élagué.

Arbre maximal : estimateur par morceaux le + complexe de la suite d'estimateurs construits → **CV garantie** (Breiman et al. 1984).

ILLUSTRATION ESTIMATEUR PAR MORCEAUX : EXEMPLE 2

Exemple en assurance : prévision de décès et modélisation des taux de mortalité. Résultats de l'article EAJ Olbricht (2012).

Portefeuille de SwissRe avec les caractéristiques suivantes :

- comprenant 1 463 964 enregistrements,
- couvrant une période de 4 ans,
- les variables explicatives en jeu sont le sexe et l'âge.

Les résultats obtenus par CART sont comparés à la table de mortalité actuelle "German standard life table DAV 2008 T".

ARBRE ÉLAGUÉ (PAS MAXIMAL !)

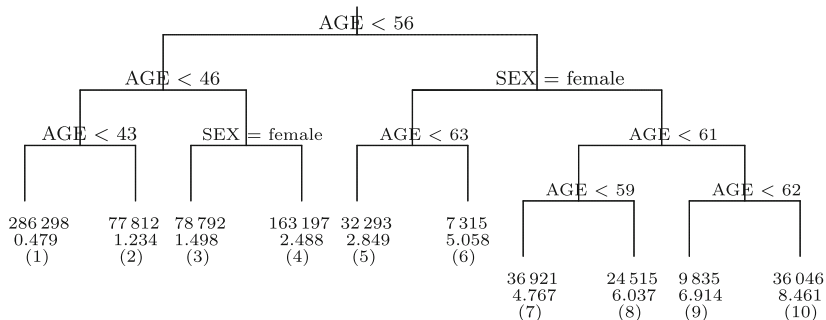
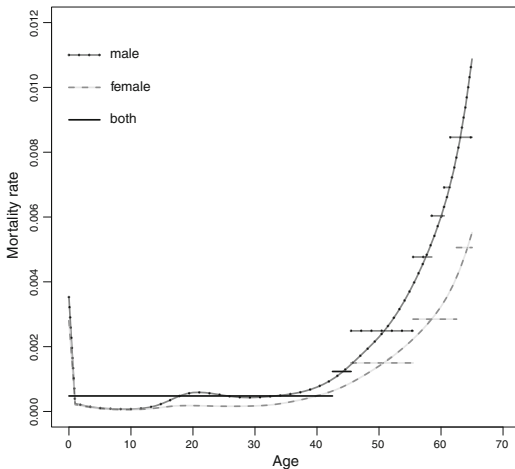


Fig. 8 Final tree for the standard life table example. For each terminal node the number of cases and the mortality rate (per mille) are given (the numbers in *brackets* are the labels for the nodes used in Table 6)

COURBES DE MORTALITÉ CORRESPONDANTES



Courbe continue : table réglementaire ; par morceaux : CART.

REMARQUE IMPORTANTE

Notez la différence majeure qu'il existe entre ce type de modélisation et une modélisation dite paramétrique.

En effet, **on s'autorise toute forme de dépendance ici**, alors qu'un modèle paramétrique (ex : GLM) impose une forme de dépendance entre Y et \mathbf{X} ...

⇒ Peut s'avérer inadapté dans de nombreux cas pratiques ! (ex : tarification d'un contrat auto en incluant l'âge ds le modèle de fréquence, ss forme de classes d'âge).

En revanche, dans l'exemple de mortalité ici, il serait **préférable d'avoir un modèle paramétrique...**

PERFORMANCE DE LA PRÉVISION CART

La performance s'évalue sur le "test set" à droite du tableau :

Table 6 Performance of the tree from Fig. 8

Node	Learning set			Independent test set			
	No. of elements in node	No. of deaths in node	Estimated mortality rate (per mille)	No. of elements in node	No. of deaths in node	Tree prediction (Fig. 8)	Classical prediction (DAV 2008 T)
1	286,298	137	0.479	254,995	143	122	127
2	77,812	96	1.234	75,882	60	94	79
3	78,792	118	1.498	79,202	146	119	116
4	163,197	406	2.488	155,912	361	388	389
5	32,293	92	2.849	33,163	119	94	96
6	7,315	37	5.058	7,440	26	38	36
7	36,921	176	4.767	41,759	163	199	188
8	24,515	148	6.037	20,708	118	125	118
9	9,835	68	6.914	8,354	59	58	55
10	36,046	305	8.461	33,525	219	284	299
Total	753,024	1,583		710,940	1,414	1,521	1,503

4 Première brique en Machine Learning : arbres de décision

- Algorithme CART
- Exemples
- Formalisation : construction de l'arbre
- Lien avec le problème de régression classique
- Gestion du surapprentissage : réduction de dimension
- Réponse catégorielle
- Outils et mesures de performance des modèles
- Extensions et conclusion

SÉLECTION DE MODÈLE (α FIXÉ)

L'arbre maximal construit (de taille $K(n)$) génère une suite d'estimateurs $(\hat{\pi}^K(\mathbf{x}))_{K=1,\dots,K(n)} \Leftrightarrow$ chaque sous-arbre.

But : éviter estimateur trop complexe (surapprentissage) \Rightarrow trouver meilleur sous-arbre selon un **arbitrage adéquation / prévision** :

$$R_\alpha(\hat{\pi}^K(\mathbf{x})) = E_n[\Phi(Y, \hat{\pi}^K(\mathbf{x}))] + \alpha(K/n),$$

où α param. de complexité, K dim. de l'estimateur (nb de feuilles).

Pour α fixé, l'estimateur final optimise un critère coût-complexité :

$$\hat{\pi}_\alpha^K(\mathbf{x}) = \arg \min_{(\hat{\pi}^K)_{K=1,\dots,K(n)}} R_\alpha(\hat{\pi}^K(\mathbf{x})). \quad (7)$$

RESULTATS REMARQUABLES

→ Pour α fixé, l'arbre $\hat{\pi}_\alpha^K(\mathbf{x})$ est **unique** et le calcul est **rapide** !

Exemples :

- $\alpha = \infty$: le modèle sélectionné sera la racine ;
- $\alpha = 0$: le modèle sélectionné sera l'arbre maximal.

→ Puisque n'importe quelle ***suite de sous-arbres emboîtés*** de l'arbre maximal a au max. K membres, toutes les valeurs possibles de α peuvent être groupées en m intervalles ($m \leq K$) :

$$I_1 = [0, \alpha_1] \quad I_2 = (\alpha_1, \alpha_2] \quad \dots \quad I_m = (\alpha_{m-1}, +\infty]$$

⇒ Chaque $\alpha \in I_i$ partage le **même sous-arbre optimal**.

PROCEDURE D'ELAGAGE

Raisonnement : impossible de parcourir ts les sous-modèles de l'arbre max. (nb sous-arbres exponent. ↗ avec nb feuilles) \Rightarrow

- 1 on part de l'arbre maximal construit ;
- 2 on considère **une 1^{ère} valeur de α** : conduit à sélectionner un sous-arbre optimal de l'arbre maximal (cf équation (7)).
- 3 **à partir de ce sous-arbre optimal**, on prend une autre valeur de α (+ grande) qui conduit à sélectionner un sous-arbre optimal de ce sous-arbre.
- 4 Et ainsi de suite... Cela crée une suite croissante de α_z !

⇒ Par construction, on obtient une **suite ↘ de sous-arbres optimaux** emboîtés (de l'arbre maximal vers la racine).

Dans cette liste d'estimateurs, on choisit finalement $\hat{\alpha}$ (et l'arbre optimal qui va avec) tel que

$$\hat{\pi}_{\hat{\alpha}}^K(\mathbf{x}) = \arg \min_{(\hat{\pi}_{\alpha_Z}^K)_{\alpha=\alpha_1, \dots, \alpha_Z}} R_{\alpha_Z}(\hat{\pi}_{\alpha_Z}^K(\mathbf{x})). \quad (8)$$

Remarque : en pratique,

→ il faut déterminer les valeurs possibles de α !


→ et $\hat{\alpha}$ est choisi en regardant cette erreur, mais moyennée via une **validation croisée** (pr minimiser une erreur de généralisation).

Consistance : Gey et Nedelec (2005) ; Molinaro, Dudoit et VanDerLaan (2004).

PROPOSITION DES VALEURS DE α

La suite des valeurs de α est obtenue lors de la construction de l'arbre maximal, avec le raisonnement suivant :

Ancien nombre de feuilles = N
 ↓
 Après segmentation, nouveau nombre de feuilles : $N+1$



$$R_\alpha(\hat{\pi}^{N+1}) = \mathbb{E}_n[\phi(y, \hat{\pi}^{N+1})] + \alpha \frac{K}{n}$$

⇒ Avant segmentation : $R_\alpha(\hat{\pi}^N) = \mathbb{E}_n[\phi(y, \hat{\pi}^N)] + \alpha \frac{K}{n}$
 Après segmentation : $R_\alpha(\hat{\pi}^{N+1}) = \mathbb{E}_n[\phi(y, \hat{\pi}^{N+1})] + \alpha \frac{N+1}{n}$

⇒ Seg. ⇒ Gain Haurgeniste ⇒ $\downarrow \mathbb{E}_n[\cdot]$
 Mais \uparrow pénalité de $\frac{\alpha}{n}$

⇒ Idée : Comparer $\mathbb{E}_n[\phi(y, \hat{\pi}^{N+1})] - \mathbb{E}_n[\phi(y, \hat{\pi}^N)]$
 avec $\alpha \frac{1}{n}$ ⇒ point de tracer α critique à chaque round

TUNING : CHOIX DE L'HYPERPARAMETRE α

- **Tuning** du modèle : sélection du paramètre de complexité α .
- **Elagage** : sélection de modèle pour un α fixé.

Comment choisir le meilleur paramètre de tuning α ?

Application à CART : une particularité... En effet, la validation croisée induit des séquences d'arbres emboîtés différentes.

⇒ L'erreur moyenne n'est pas calculée pour chaque sous-arbre avec un nb de feuilles donné, mais pour chaque valeur α_z fixée *issue de la séquence produite initialement par tout l'échantillon.*

Le choix de α répond à l'équation (8) (où l'erreur est moyennée) ⇒ fournit le bon α et donc l'arbre optimal !

→ En pratique, choisis 1^{er} point en-dessous de min+1SE
(Therneau : An Introduction to Recursive Partitioning Using the RPART Routines).

FORMULATION ALGORITHMIQUE (V-fold)

- ① Construction de l'arbre maximal T_{max} ;
- ② Construction de la séquence T_K, \dots, T_1 d'arbres emboîtés associée à une séquence de valeurs (α_z) ;
- ③ Pour $v = 1, \dots, V$ (où v désigne le segment de l'échantillon initial servant à la validation),
 - pr chq nouvel éch. d'apprentissage, construire T_{max} et estimer la séquence d'arbres associée à la séq. des pénalisations α_z ,
 - estimation de l'erreur sur la partie validation de l'échantillon ;
- ④ Calcul de la séquence des moyennes de ces erreurs ;
- ⑤ L'erreur minimale désigne la pénalisation α_{opt} optimale ;
- ⑥ Retenir l'arbre associé à α_{opt} ds la suite initiale T_K, \dots, T_1 .

VALIDATIONS CROISÉES DANS rpart

Pour amener plus de robustesse au choix du paramètre de complexité α , on procède par validations croisées.

Principe de la validation croisée : meilleur compromis biais / variance. On diminue la variance de l'estimateur en recherchant une valeur réaliste de l'erreur basée sur plusieurs calibrations.

Dans le cadre de l'algorithme CART, cela consiste en les étapes :

- ➊ Construire l'arbre maximal (modèle complet) sur l'échantillon ;
- ➋ Dédire les intervalles I_1, I_2, \dots, I_m à partir des α_z .
- ➌ Construire la suite (β_z) (pour se placer dans les intervalles $]\alpha_k, \alpha_{k+1}]$) telle que

$$\begin{aligned}
 \beta_1 &= 0 \\
 \beta_2 &= \sqrt{\alpha_1 \alpha_2} \\
 \dots &= \dots \\
 \beta_{m-1} &= \sqrt{\alpha_{m-2} \alpha_{m-1}} \\
 \beta_m &= \infty
 \end{aligned}$$

- ④ Diviser l'échantillon d'origine en s sous-groupes G_1, G_2, \dots, G_s de taille s/n (n est la taille de l'échantillon de base).
- ⑤ Sur chaque sous-groupe i :
 - construire l'arbre maximal sur l'ensemble des sous-groupes sauf le groupe i , et déterminer les sous-arbres $T_{\beta_1}, T_{\beta_2}, \dots, T_{\beta_m}$,
 - prédire la quantité d'intérêt pour chaque observation du groupe i dans chaque modèle $T_{\beta_z}, 1 \leq z \leq m$;
 - calculer l'erreur pour chaque sous-arbre.
- ⑥ Pour chaque β_z , sommer les erreurs des G_i . Prendre le paramètre de complexité β d'erreur minimale, et choisir T_β comme meilleur sous-arbre sur l'échantillon de base.

4 Première brique en Machine Learning : arbres de décision

- Algorithme CART
- Exemples
- Formalisation : construction de l'arbre
- Lien avec le problème de régression classique
- Gestion du surapprentissage : réduction de dimension
- Réponse catégorielle
- Outils et mesures de performance des modèles
- Extensions et conclusion

ARBRE DE CLASSIFICATION : Y DISCRÈTE

Supposons que $Y \in \{A, B\}$.

Dans le cas discret, la quantité d'intérêt est

$$\pi_0(\mathbf{x}) = E_0[1_{Y=A} | \mathbf{X} = \mathbf{x}] = \mathbb{P}(Y = A | \mathbf{X} = \mathbf{x})$$

Ici il faut **adapter le critère d'homogénéité**, donc la perte Φ .

On considère classiquement de

- l'indice de Gini,
- l'entropie.

ENTROPIE

La **fonction d'entropie** est classiquement définie pour $p \in [0, 1]$ par

$$f(p) = -p \log(p).$$

Appliqué aux CART, dans un pb à 2 classes $\{A, B\}$ pour Y , on définit l'hétérogénéité du noeud t (convention $0 \log(0) = 0$) comme

$$H_t = -2 \sum_{l \in \{A, B\}} |t| p_t^l \log(p_t^l),$$

où p_t^l est la proportion de la classe l dans le noeud t .

On maximise ↘ **hétérogénéité**, soit $\max_{div.} H_t - (H_{t_g} + H_{t_d})$.

CONCENTRATION DE GINI

La **concentration de Gini** est définie pour $p \in [0, 1]$ par

$$f(p) = p(1 - p).$$

Appliqué aux CART, on définit l'hétérogénéité comme

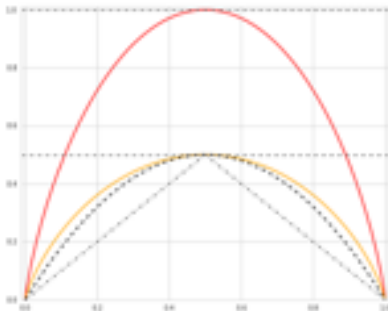
$$H_t = \sum_{l \in \{A, B\}} p_t^l (1 - p_t^l).$$

Rq :

- La concentration de Gini est la variance d'une Bernoulli...
- Proportions remplaçable par des proba. conditionnelles si proba. a priori des classes connues (\neq proba. observées). Sinon, proba. de chq classe estimées sur l'éch. (revient à prendre proportion).

GRAPHIQUE DE L'ERREUR

Ds tous les cas, la quantité à optimiser sera convexe/concave.



⇒ Zones intéressantes : extrémités de $[0, 1]$.

AFFECTATION POUR PREVISION

Concernant l'affectation de l'observation à prédire à l'une des classes, il y a donc 3 distinctions possibles en fonction de l'information à disposition :

- soit on affecte la classe la plus représentée dans la feuille,
- soit on affecte la classe a posteriori la plus probable (au sens bayésien) si l'on dispose de probabilités a priori (pas les proba. de représentation dans l'échantillon) des classes,
- soit on affecte la classe la moins coûteuse si des coûts de mauvais classement sont donnés.

4 Première brique en Machine Learning : arbres de décision

- Algorithme CART
- Exemples
- Formalisation : construction de l'arbre
- Lien avec le problème de régression classique
- Gestion du surapprentissage : réduction de dimension
- Réponse catégorielle
- Outils et mesures de performance des modèles
- Extensions et conclusion

REPONSE QUANTITATIVE

Les **mesures classiques de performance** d'un modèle si Y est quantitative sont :

- **l'Erreur Quadratique Moyenne** (EQM, ou MSE) :

$$MSE(\hat{\pi}^K(\mathbf{x})) = \sum_i (Y_i - \hat{\pi}^K(\mathbf{x}_i))^2$$

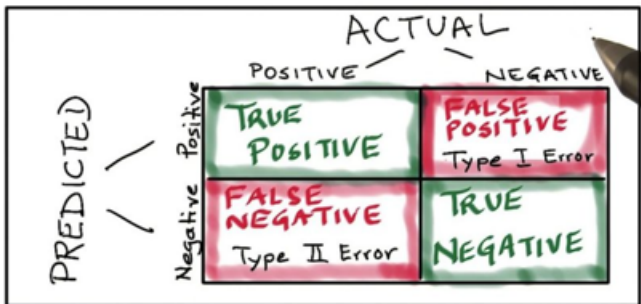
- **l'Erreur Absolue Moyenne** (EAM, ou MAE) :

$$MAE(\hat{\pi}^K(\mathbf{x})) = \sum_i |Y_i - \hat{\pi}^K(\mathbf{x}_i)|$$

Rq : évidemment, ces erreurs se mesurent sur un échant. test, pas des échant. ayant servi à construire et tuner/optimiser le modèle...

REPONSE CATEGORIELLE : MATRICE DE CONFUSION

Dans un pb de classif., on utilise svnt la [matrice de confusion](#) comme mesure de performance \Rightarrow résume les indiv. mal classés et ceux bien classés par le modèle :



A hand-drawn diagram of a confusion matrix. The vertical axis is labeled 'PREDICTED' and the horizontal axis is labeled 'ACTUAL'. The matrix is divided into four quadrants: True Positive (green), False Positive (red, labeled 'Type I Error'), False Negative (red, labeled 'Type II Error'), and True Negative (green). A pencil is pointing to the top right corner of the matrix.

		ACTUAL	
		POSITIVE	NEGATIVE
PREDICTED	Positive	TRUE POSITIVE	FALSE POSITIVE Type I Error
	Negative	FALSE NEGATIVE Type II Error	TRUE NEGATIVE

REMARQUES

En utilisant cet outil, on peut calculer facilement :

- le **taux de mauvaise classification** :

$$(FP + FN)/(FP + FN + TP + TN)$$

- l'indice de **sensibilité** : $TP/(TP + FN)$
- l'indice de **spécificité** : $TN/(TN + FP)$

Ds la pratique, on optimise svt le modèle par rapport à 1 des 2 indices, qui mène à la prudence du modèle (svt la spécificité, qui mesure la prédiction d'un événement rare...).

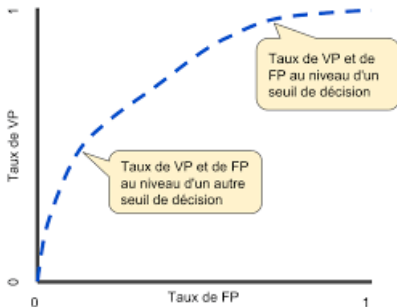
LIMITES DE CETTE MESURE

Principalement 2 limites à l'utilisation de cette matrice :

- 1 **dépendante d'un seuil d'affectation** : pour classer les prév. du modèle, on définit ce seuil. Dans un pb à 2 classes, svt 0,5 \Rightarrow bien connu que ce n'est svt pas seuil optimal (\Rightarrow ROC).
- 2 ds un pb où classes de Y sont largement **disproportionnées**, le modèle prédira tjs la même classe et donnera 1 erreur de classif. globalement très faible... Peu réaliste, car souvent c'est l'événement rare qu'il nous intéresse de prédire... Donc en fait l'erreur sur cette prévision est maximale, puisque l'événement en question n'est jamais prédit !

COURBE ROC ET AUC

ROC (Receiving Operator Curve) résume taux de VP (sensibilité) et FP (1-spécificité) pour ts les seuils d'affectation :



AUC (Area Under Curve) : $\in [0,5 \text{ (modèle aléatoire)} ; 1 \text{ (parfait)}]$.

AUTRES OUTILS : C-INDEX, F_1 -SCORE

Au lieu d'utiliser la matrice de confusion pour optimiser un modèle, on peut aussi utiliser une mesure différente qui répond à une autre logique...

- le C-index (descendant de l'AUC...) : cf thèse Anani
- ex : article Pierrick ;
- F_1 score...permet de tuner les hyperparamètres en optimisant ce score ! (cf article Yohan Le Faou)

4 Première brique en Machine Learning : arbres de décision

- Algorithme CART
- Exemples
- Formalisation : construction de l'arbre
- Lien avec le problème de régression classique
- Gestion du surapprentissage : réduction de dimension
- Réponse catégorielle
- Outils et mesures de performance des modèles
- Extensions et conclusion

EXTENSIONS : AUTRES FONCTIONS DE PERTE Φ

$$\pi_0(\mathbf{x}) = \arg \min_{\pi(\mathbf{x})} E_0[\Phi(Y, \pi(\mathbf{x})) | \mathbf{X} = \mathbf{x}]$$

→ **Estimation de moyenne** : $\pi_0(\mathbf{x}) = E_0[Y | \mathbf{X} = \mathbf{x}]$

Critère de division (MCO) : $\Phi(Y, \pi(\mathbf{x})) = (Y - \pi(\mathbf{x}))^2$.

→ **Quantile** : $\pi_0(\mathbf{x}) = Q_Y(\alpha | \mathbf{X} = \mathbf{x}) = \inf\{y : F(y | \mathbf{X} = \mathbf{x}) \geq \alpha\}$

$\Phi_\alpha(y, \pi(\mathbf{x})) = \alpha|y - \pi(\mathbf{x})|\mathbb{1}(y > \pi(\mathbf{x})) + (1 - \alpha)|y - \pi(\mathbf{x})|\mathbb{1}(y \leq \pi(\mathbf{x}))$

→ **Estimation de densité** de la loi de Y :

$\Phi(Y, \pi(\mathbf{x})) = -\log \pi(Y, \mathbf{x})$, avec π la densité jointe de (Y, \mathbf{X}) .

⇒ En pratique, **version empirique** de ces mesures par l'estimateur !

DONNEES MANQUANTES : LES SURROGATE SPLITS

Dans la pratique, on n'observe pas certaines variables explicatives pour certains individus \Rightarrow on ne peut pas les faire descendre dans l'arbre pour en déduire une prévision...

Dans ce cas, on impute la donnée manquante ou on utilise une **surrogate split** (obligatoirement basée sur une autre covariable!).

Correspond à la division **la** + **voisine** de celle initialement choisie, en termes de **concordance des individus envoyés dans chacun des noeuds** fils \Rightarrow imite au mieux la meilleure d'origine, mesurée par une mesure d'association entre 0 et 1 (1 est un clône).

PROBLEMATIQUES CLASSIQUES A GERER

Problème de biais de l'estimateur CART lorsqu'une variable explicative catégorielle contient trop de modalités... Tendance à attirer la règle de division à cette variable notamment.

Problème lorsque unbalanced response : on se retrouve qu'avec la racine et on ne segmente pas ! Que faire si on a juste la racine ?...
cf <https://stats.stackexchange.com/questions/28029/training-a-decision-tree-against-unbalanced-data>

Problème de censure, troncature...

CONCLUSION SUR CART

- + Algorithme simple, résultat facile à interpréter (règles, fournit pouvoir discriminant facteurs de risque).
- + Procédure statistique consistante théoriquement.
- + Méthode non-paramétrique, et invariante par transformation monotone des covariables (rangs utilisés) \Rightarrow robustesse.
- + Adapté à la gestion de bc de var. explic. : sélection variables “intégrée” à l’algo. et interactions implicitement considérées.
- + Extensions possibles avec adaptation de la perte.
- Algo récursif : peut passer à côté de l’optimum global...
- Instabilité aux données d’apprent. (variance estimateur) du fait de structure hiérarchique \Rightarrow gagner en robustesse.