Présentation de l'algorithme CART

C. Tuleau-Malot

Plan

Plan

Plan

Plan

Plan

Contexte d'étude

On dispose d'un échantillon d'apprentissage $\mathcal{L} = \{(X_i, Y_i)_{i\in\{1,\dots,n\}}\}$ où :

- $(X_i, Y_i)_{i\in\{1,\dots,n\}}$ sont n réalisations indépendantes d'un couple de variables $(X, Y) \in \mathcal{X} \times \mathcal{Y}$
- $\forall\ i \in \{1, \dots, n\}$, $X_i = (x_i^1, \dots, X_i^p)$ vecteur de taille p contenant les observations pour l'individu i des p variables explicatives
- Y_i est l'observation pour l'individu i de la variable à expliquer

Deux cadres d'étude :

- régression : $\mathcal{Y} \subset \mathbb{R}$ et le modèle est $Y = s(X) + \varepsilon$ (ε étant une variable de bruit généralement supposée gaussienne centrée)
- classification : $\mathcal{Y} \subset \mathcal{J}$ et $Y = s(X)$

Contexte d'étude (2)

L'expression de la fonction s est connue

- régression : $s(x) = \mathbb{E}[Y|X = x]$
- classification : $s(x) = \underset{j \in \mathcal{J}}{argmax} P(Y = j|X = x)$

Remarque : il y a le cas particulier de la classification binaire

$$s(x) = 1_{\eta(x) \geq 1/2}$$

où $\eta(x) = P(Y = 1|X = x)$

Mais s est en pratique inconnue $\Rightarrow$ estimation

Etat de l'art

Pour estimer s, il existe différentes méthodes :

- régression : Moindres carrés ordinaires (OLS : projection sur l'espace vectoriel engendré par les variables explicatives de Y), les méthodes Ridge et Lasso (versions pénalisées de OLS) + méthodes non paramétriques
- classification : k-plus proches voisins, SVM, ...

Alternative : arbre de décision $\Rightarrow$ modélisation non linéaire
$\Rightarrow$ CART (**C**lassification **A**nd **R**egression **T**rees)

historique

CART : algorithme développé par Breiman, Friedman, Olshen et Stone (1984)

- estimateurs par histogramme de la fonction cible s
- partitionnement récursif et dyadique de l'espace des observations ($\mathcal{X}$)

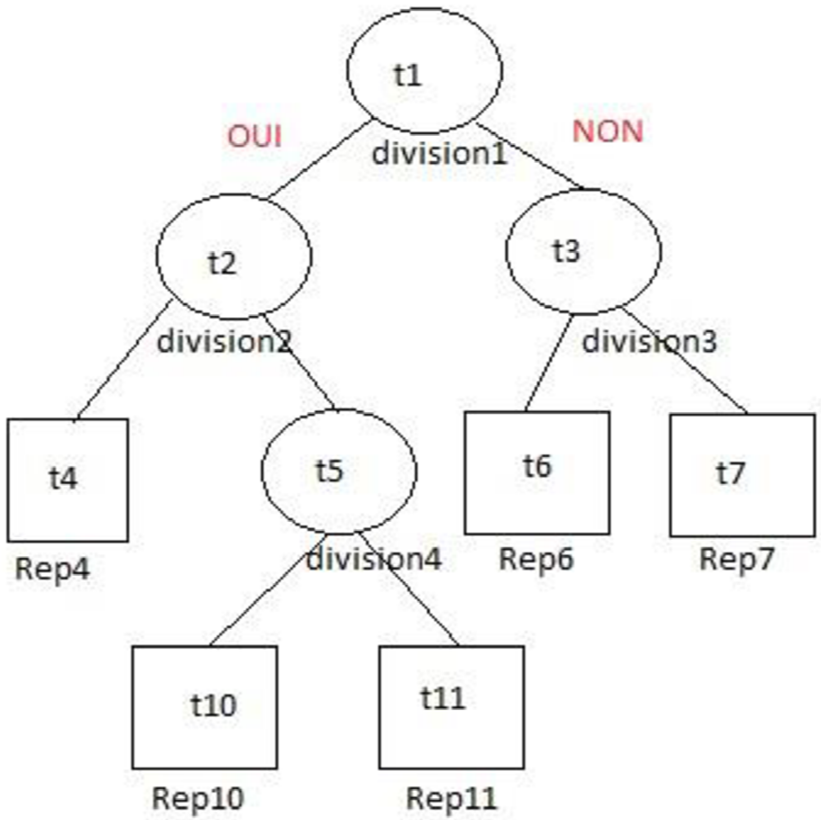
algorithme

Algorithme de construction d'un arbre CART : 3 étapes successives

- Construction de l'arbre maximal
- Elagage
- Sélection finale

représentation

Voici l'aspect d'un arbre CART :



représentation (2)

Quelques questions se posent :

- Comment définir les divisions successives ? $\Rightarrow$ critère de construction
- Quand arrêter le principe de division ? $\Rightarrow$ règle d'arrêt
- Comment définir les réponses ? $\Rightarrow$ règle d'assignation

Une difficulté supplémentaire : les réponses peuvent être différentes selon le cadre d'étude

critère de construction

Les divisions :

- question binaire
- de la forme $X^i \leq a$ ou $X^i \in \mathcal{C}$

$\Rightarrow$ détermination de i et de a ou $\mathcal{C}$

idée : tester toutes les divisions possibles et retenir la meilleure selon un critère

critère de construction (2)

Cadre de la classification : $\mathcal{J} = \{1, \ldots, J\}$
Notations :

- π_j probabilité à priori de la classe j. Peut être estimée par $\frac{N_j}{n}$ avec $N_j = Card\{(x_k, y_k)|y_k = j\}$
- soit t un noeud de l'arbre, $N(t) = Card\{(x_k, y_k)|x_k \in t\}$ nombre d'observations de $\mathcal{L}$ dans t
- soit t un noeud de l'arbre et $j \in \mathcal{J}$, $N_j(t) = Card\{(x_k, y_k)|x_k \in t \text{ et } y_k = j\}$ nombre d'observations de $\mathcal{L}$ dans t et de classe j

critère de construction (3)

Estimations :

- $P(j,t)$: probabilité qu'une observation soit dans le noeud t et de classe j
  $\Rightarrow$ estimée par $p(j,t) = \pi_j \frac{N_j(t)}{N_j}$
- $P(t)$: probabilité qu'une observation soit dans le noeud t
  $\Rightarrow$ estimée par $p(t) = \sum_j p(j,t)$
- $P(j|t)$: probabilité a posteriori dans t de la classe j
  $\Rightarrow$ estimée par $\frac{p(j,t)}{p(t)}$

critère de construction (4)

définition :

Soit h une fonction de $\{(p_1, \ldots, p_J) | p_i \geq 0, \ \sum_j p_j = 1\}$ dans $\mathbb{R}$. h est une fonction dite d'hétérogénéité si :

- h est symétrique en $p_1, \ldots, p_J$
- h est maximale en $(\frac{1}{J}, \ldots, \frac{1}{J})$
- h est minimale en $(1, 0, \ldots, 0)$, $(0, 1, 0, \ldots, 0)$,..., $(0, \ldots, 0, 1)$

définition :

Soit t un noeud. On définit l'hétérogénéité de t par :

$$i(t) = h(p(1|t), \ldots, p(J|t))$$

avec h une fonction d'hétérogénéité

Exemples de fonction d'hétérogénéité :

- Gini $h(p_1, \ldots, p_J) = \sum_{i \neq j} p_i p_j$
- Shanon $h(p_1, \ldots, p_J) = -\sum_j p_j log(p_j)$

critère de construction (5)

Principe de construction d'une division :

Soit t un noeud de l'arbre. Soit t_d son descendant droite et t_g son descendant gauche, descendants engendrés par une division δ.
On note $p_g = \frac{p(t_g)}{p(t)}$ et $p_d = \frac{p(t_d)}{p(t)}$: proportion d'observations envoyées respectivement dans t_g et t_d.
La variation d'hétérogénéité générée par δ est définie par :

$$\Delta i(\delta, t) = i(t) - p_g i(t_g) - p_d i(t_d)$$

La division optimale du noeud t est donnée par :

$$\delta^*(t) := \delta^* = \underset{\delta \ division}{argmax} \ \Delta i(\delta, t)$$

critère de construction (6)

La construction ne sera valide que si h est une fonction concave $\Rightarrow \Delta i(\delta, t) \geq 0$ (*)

Exercices :

- Montrer que la focntion d'hétérogénéité de Gini est concave
- Montrer que si h est concave alors $\Delta i(\delta, t) \geq 0$

La propriété (*) est essentielle pour valider la procédure d'optimisation globale par des étapes locales.

définition :

Soit T un arbre. On note $\tilde{T}$ l'ensemble des feuilles de T.
L'hétérogénéité de l'arbre T est définie par :

$$I(T) = \sum_{t \in \tilde{T}} i(t)p(t)$$

critère de construction (7)

propriété :

Soit T un arbre. Soit $\tilde{t}$ une feuille de T. Soit $\tilde{\delta}$ une division possible de $\tilde{t}$ et soit T' l'arbre résultant de la division $\tilde{\delta}$.
Alors, on a :

$$I(T) - I(T') = p(\tilde{t})\Delta i(\tilde{\delta}, \tilde{t})$$

Exercice : Montrer cette propriété.

règle d'arrêt

Par récursivité, la construction de l'arbre devient évidente.

règle d'arrêt :
Un noeud t d'un arbre est déclaré terminal si :

- Une seule observation dans le noeud t
- Que des observations dans t avec un même label

$\Rightarrow$ Partition minimale de l'espace $\mathcal{X}$.

règle d'assignation

définition :
Soit t un noeud dont la réponse associée est $j(t)$.
La probabilité de mauvais classement du noeud t, évaluée par substitution, est définie par :

$$r(t) = \sum_{j \neq j(t)} p(j|t)$$

définition :
Soit t un noeud.
La réponse $j(t)$ associée est définie par :

$$j(t) = \underset{j \in \mathcal{J}}{argmax}\ p(j|t)$$

règle d'assignation (2)

propriété :
Cette définition de $j(t)$ est telle que $r(t)$ est minimale.

Exercice :

- Montrer la propriété précédente.
- Soit $R(T) = \sum_{t \in \tilde{T}} p(t)r(t)$. Soit T' un arbre issu de T (un arbre plus développé). Montrer que $R(T') \leq R(T)$.

critère de construction (8)

Cadre de la régression : $\Rightarrow$ Cadre beaucoup plus simple

définition :

Soit d un prédicteur de s induit par un arbre T. Son erreur est :

$$R^*(d) = \mathbb{E}[(d(X) - Y)^2]$$

Cette erreur est estimée par :

$$R(d) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - d(X_i))^2$$

propriété : Le prédicteur $\tilde{d}$ issu de l'arbre T est celui qui minimise $R(d)$ parmi l'ensemble des prédicteurs issus d'un arbre. Donc

$$\tilde{d}(x) = \sum_{t\in\tilde{T}} a(t)1_{x\in t}$$

avec $a(t) = \frac{1}{N(t)}\sum_{X_i\in t} Y_i := \bar{Y}_t$.

Exercice : Montrer la propriété précédente

critère de construction (9)

Construction d'une division

On définit pour un arbre T :

$$R(T) = \sum_{t \in \tilde{T}} R(t)$$

où $R(t) = \frac{1}{n} \sum_{X_i \in t} (Y_i - \bar{Y}_t)^2$.

Soit T un arbre et t une feuille de T.
Soit δ une division de t en t_g et t_d.
On pose :

$$\Delta R(t, \delta) = R(t) - R(t_g) - R(t_d).$$

On définit la division optimale de t, notée δ^* par :

$$\delta^* = \underset{\delta}{argmax}\ \Delta R(t, \delta)$$

critère de construction (10)

Exercices :

- Montrer que $\Delta R(t, \delta) \geq 0$

- Soit T un arbre. Soit $\tilde{t}$ une feuille de T. Soit $\tilde{\delta}$ une division possible de $\tilde{t}$ et soit T' l'arbre résultant de la division $\tilde{\delta}$. Alors, on a :

$$R(T) - R(T') = \Delta R(\tilde{\delta}, \tilde{t})$$

Pourquoi

Nous avons à présent un arbre maximal T.
Problème : arbre en général inexploitable

- nombre de feuilles trop grand
- abre trop fidèle aux données d'apprentissage

$\Rightarrow$ Créer une suite de sous-arbres à l'aide d'un critère pénalisé

Sous suite de Breiman

Notations :

- Soit T un arbre et t un noeud non terminal de T. Elaguer T à partir de t consiste à créer un nouvel arbre T^* qui n'est autre que T privé de tous les descendants de t.
- Tout arbre T' obtenu par élagage de T est un sous-arbre de T, ce que l'on note $T' \prec T$.

Critère d'élagage :

$$crit_\alpha(T) = R(T) + \alpha|\tilde{T}|$$

Sous suite de Breiman (2)

Proposition :

Pour chaque valeur de α, il existe un unique sous-arbre de T_{max}, noté T_α, tel que :

- $T_\alpha = \underset{T \prec T_{max}}{argmin}\ crit_\alpha(T)$
- si $crit_\alpha T_\alpha = crit_\alpha(T)$, alors $T_\alpha \prec T$.

proposition :

Soit α_1 et α_2 deux réels positifs avec $\alpha_1 \leq \alpha_2$. Alors $T_{\alpha_2} \prec T_{\alpha_1}$.

Sous suite de Breiman (3)

Détermination du premier élément de la suite : Soit T_0 le premier élément de la suite associé à $\alpha = 0$.
T_0 est le sous-arbre de T_max obtenu en élagant tous les noeuds t pour lesquels $R(t) = R(t_g) + R(t_d)$.

Ainsi T_0 satisfait :

- $crit_0(T_0) = 0$
- pour tout noeud t de T_0, $R(t) > R(t_d) + R(t_g)$.

Sous suite de Breiman (4)

Détermination du deuxième élément de la suite : Par définition de T_0, on a pour tout noeud t de T_0, $crit_0(t) > crit_0(T_0^t)$, soit $R(t) > R(T_0^t)$.
De ce fait, tant que α demeurera suffisamment petit, on aura :

$$R(t) + \alpha > R(T_0^t) + \alpha |T_0^t|$$

α suffisamment petit signifie $\alpha < \frac{R(t)-R(T_0^t)}{|T_0^t|-1} = s(t, T_0^t)$ pour tout noeud t de T_0.
Lorsque α atteint un seuil, cela signifie que $crit_\alpha(t) = crit_\alpha(T_0^t)$ et que le noeud t devient préférable à la branche issue de t.
Posons $\alpha_1 = \min_{t \text{ noeud de } T_0^t} s(t, T_0^t)$ et définissons $T_1 := T_{\alpha_1}$ comme étant le sous-arbre de T_0 obtenu en élagant toutes les branches issues de noeuds minimisant $s(t, T_0^t)$.
T_1 est le deuxième élément de la sous-suite.

Sous suite de Breiman (6)

Théorème :

Il existe une suite finie strictement croissante de réels positifs notée $(\alpha_k)_{\{k\in\{0,\dots,K\}\}}$ telle que :

$$\forall k \in \{0, \dots, K-1\},\ si\ \alpha \in [\alpha_k,\ \alpha_{k+1}[,\ T_\alpha = T_{\alpha_k} := T_k$$

Pourquoi

A la fin de la phase d'élagage, nous disposons de plusieurs sous-arbres et donc de plusieurs estimateurs.
$\Rightarrow$ En sélectionner un

Deux méthodes :

- Echantillon test
- Validation croisée

Echantillon test

Nouvel échantillon : $\mathcal{L}' = \{(X'_1, Y'_1), \ldots, (X'_m, Y'_m)\}$

On évalue l'erreur associée à chacun des sous-arbres :

- en régression : $R(T_k) = \frac{1}{m} \sum_{i=1}^{m} (Y'_i - T_k(X'_i))^2$
- en classification : $R(T_k) = \frac{1}{m} \sum_{i,j \in \mathcal{J}} N^k_{i,j}$ où $N^k_{i,j}$ est le nombre d'observations de $\mathcal{L}'$ de classe i classées j par le sous-arbre T_k.

L'arbre final retenu est celui avec la plus faible erreur estimée.

Echantillon test (2)

Inconvénient : il faut pouvoir diviser l'échantillon en deux

Validation croisée

Principe : chaque observation sert à la fois à apprendre et à tester !

On découpe $\mathcal{L}$ en une partition équilibrée à V morceaux.
A l'étape i on utilise tous les morceaux sauf le numéro i, pour construire une sous-suite dont on évalue l'erreur à l'aide du morceaux i. On fait cela pour i variant de 1 à V et on fait la moyenne des erreurs.