

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

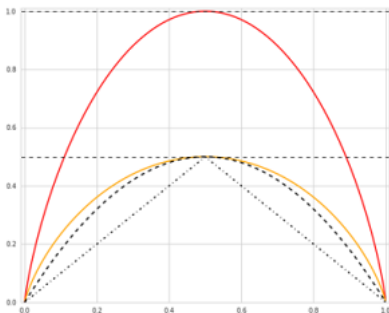
$$H_t = \sum_{l \in \{A, B\}} p_t^l (1 - p_t^l).$$

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) :

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

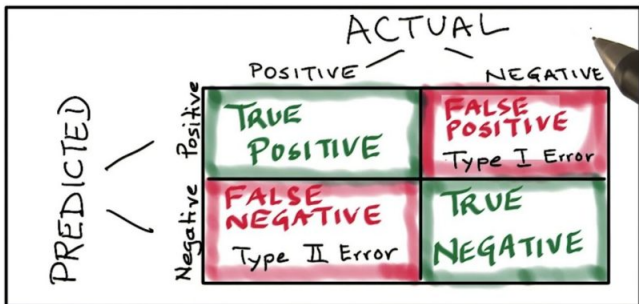
- **l'Erreur Absolue Moyenne** (EAM, ou MAE) :

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

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...).

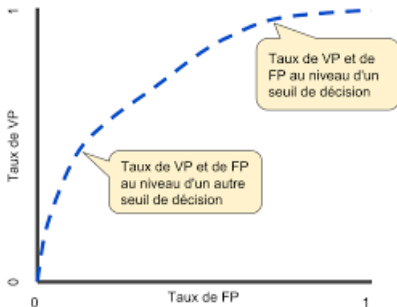
LIMITES DE CETTE MESURE

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.

UN MOT SUR LA ROBUSTESSE PREDICTIVE

Certaines techniques ont été développées afin de **stabiliser la prévision donnée par un estimateur arbre**.

En effet, la construction d'un arbre optimal **peut varier fortement quand bien même le jeu de données initial varie peu...**

⇒ Proposer des estimateurs agrégés ⇒ \searrow variance estimateur !

Pour éviter de corréler les estimateurs simples qui composeront l'estimateur agrégé, on peut intégrer par exemple

- 1 choix aléatoire des covariables considérées lors d'1 division ;
- 2 tirage aléatoire de sous-jeux de données.

UN MOT SUR LES STRATEGIES D'AGREGATION

Deux stratégies s'opposent dans le raisonnement :

→ **Stratégie d'agrégation aléatoire** (**bagging** : bootstrap aggregating) : créer des échantillons, construire le modèle sur chq échantillon, combiner les modèles (ex : type forêts aléatoires).

→ **Stratégie alternative, apprentissage incrémental** (**boosting**) : apprentissage sur 1 paquet, prévision sur paquet 2, puis apprendre des exemples mal prédits du paquet 2, actualiser modèle, puis recommencer sur les paquets suivants ⇒ **apprendre, mémoriser** (ex : GBM).

LE BAGGING PLUS EN DETAIL

[FH00], [Bre94]

Le bagging conduit structurellement à diminuer la variance d'un estimateur.

En effet, n'importe quelle estimateur peut s'écrire à l'aide d'un développement de Taylor...Le premier terme étant la partie linéaire, les termes suivants étant des termes d'ordre supérieur. Le bagging ne touche pas au premier terme, mais considère l'espérance des termes suivants... Faisant ainsi diminuer la variance !

Conclusion : plus la quantité à estimer est linéaire (problème simple et dimension raisonnable), moins le bagging est efficace !

LE BOOSTING PLUS EN DETAIL

[FS97], [Fri01], [Sha03]

Le bagging conduit structurellement à diminuer la variance d'un estimateur.

En effet, n'importe quelle estimateur peut s'écrire à l'aide d'un développement de Taylor...Le premier terme étant la partie linéaire, les termes suivants étant des termes d'ordre supérieur. Le bagging ne touche pas au premier terme, mais considère l'espérance des termes suivants... Faisant ainsi diminuer la variance !

Conclusion : plus la quantité à estimer est linéaire (problème simple et dimension raisonnable), moins le bagging est efficace !

5 Bagging + randomization de CART : forêts aléatoires

- Principe
- Construction de la forêt aléatoire
- Force, corrélation et erreur de la forêt
- Interprétabilité de modèles ensemblistes

PRINCIPES DES RANDOM FORESTS

Agrégation d'estimateurs CART.

L'objectif des forêts aléatoires est de proposer un estimateur "moyenné" afin d'améliorer la robustesse de l'estimation de la quantité d'intérêt (\searrow **variance** estimateur agrégé).

Il s'agit d'**intégrer** une multitude de prévisions obtenues dans une estimation finale. Approche intéressante pour 2 raisons principales :

- on peut **dégager un classement robuste du pouvoir explicatif de chacun des facteurs de risque**,
- sa consistance a été démontrée dans plusieurs articles.

MAIS N'OUBLIEZ PAS...

“RF is an example of a tool that is useful in doing analyses of scientific data.”

“But the cleverest algorithms are no substitute for human intelligence and knowledge of the data in the problem.”

“Take the output of random forests not as absolute truth, but as smart computer generated guesses that may be helpful in leading to a deeper understanding of the problem.”

Leo BREIMAN.

PRÉVISIONS : Y CONTINUE VS Y DISCRÈTE

Soit \hat{Y}_i l'estimateur obtenu pour l'indiv. i par un CART **maximal** (pour diminuer le biais).

On construit N arbres CART en modifiant l'échantillon à chaque fois. Pour chaque obs. i , l'estimateur forêts aléatoires vaut :

- une **moyenne** dans le cas où Y est continue :

$$\hat{Y}_i^{RF} = \frac{1}{N} \sum_{n=1}^N \hat{Y}_i^{CART}$$

- un **vote majoritaire** si Y est discrète :

$$\hat{Y}_i^{RF} = \arg \max_{k=A,B} (\# \hat{Y}_i^{CART} = k)$$

5 Bagging + randomization de CART : forêts aléatoires

- Principe
- Construction de la forêt aléatoire
- Force, corrélation et erreur de la forêt
- Interprétabilité de modèles ensemblistes

RF = BAGGING + RANDOMIZATION

Les forêts aléatoires étaient basées sur plusieurs arbres CART.
Chacun de ces arbres est construit comme suit.

- ① Construire un échantillon bootstrap de même taille que l'apprentissage (répliquer l'éch. selon mesure empirique) ;
- ② Construire l'arbre CART sur cet échantillon bootstrap : considérons qu'il y a k facteurs de risque, avec $m \ll k$:
 - à chaque noeud, on tire aléatoirement m facteurs de risque parmi les k disponibles ;
 - on cherche la division optimale basée sur ces m covariables ;
 - où s'arrête-t-on dans la construction (cf slide suivante) ?
- ③ agréger ces arbres pour construire l'estimateur forêt.

Remarque : m ne change pas entre les \neq arbres de la forêt.

TROIS STRATÉGIES D'ÉLAGAGE

Chaque arbre est-il élagué ? On distingue 3 stratégies \neq

- ➊ Laisser construire l'arbre maximal pour chacun des échant..
→ Bon compromis volume des calculs / qualité des prév. : faible biais et grande variance de chaque estimateur.
- ➋ Construire un arbre d'au plus q feuilles → Cf plus loin...
- ➌ Construire l'arbre maximal à chaque fois, puis l'élaguer par validations croisées → pénalise lourdement la quantité de calculs sans gain substantiel de qualité de prévision...

Rq : stratégie (1) implémentée par défaut dans `randomForest(.)`.

BOOTSTRAP, AGGRÉGATION ET RANDOM FORESTS

On fait B échantillons bootstrap (nombre d'arbres dans la forêt).

$$\hat{\pi}^{RF}(x) = \frac{1}{B} \sum_{b=1}^B \hat{\pi}_b^{OBT}(x)$$

$$\begin{aligned} \Rightarrow \text{Var}(\hat{\pi}^{RF}(x)) &= \text{Var}\left(\frac{1}{B} \sum_{b=1}^B \hat{\pi}_b^{OBT}(x)\right) \\ &= \frac{1}{B^2} \text{Var}\left(\sum_{b=1}^B \hat{\pi}_b^{OBT}(x)\right) \\ &\leq \frac{1}{B^2} B \times \max_{b=1, \dots, B} (\text{Var}(\hat{\pi}_b^{OBT}(x))) \end{aligned}$$

$$\leq \frac{1}{B} \max(\text{Var}(\cdot)) \Rightarrow \downarrow \text{Variance de l'estimateur}$$

Mais on conserve le même ordre de grandeur pour le biais puisque l'espérance est linéaire.

\Rightarrow on baisse l'erreur de l'estimateur.

5 Bagging + randomization de CART : forêts aléatoires

- Principe
- Construction de la forêt aléatoire
- Force, corrélation et erreur de la forêt
- Interprétabilité de modèles ensemblistes

ERREUR DE LA FORÊT

L'erreur associée à la forêt dépend de 2 paramètres :

- la **corrélation** entre les arbres de la forêt : + cette corrélation ↗, + l'erreur est grande ;
- la capacité de chq arbre ds la forêt à donner une estimation proche de réalité (**force**) : + l'arbre est précis, – erreur gde.

Par rapport au paramètre de tuning “ m ”, on observe que

- abaisser m réduit la corrélation et la force,
- agrandir m augmente la corrélation et la force.

⇒ Arbitrage à trouver sur $m \rightarrow$ minimiser erreur $O(ut)-O(f)-B(ag)$

Rq : l'autre paramètre de tuning est le nombre d'arbres de la forêt.

L'ERREUR OOB

Au sein de la construction de chaque arbre CART de la forêt, on ne considère qu'une portion de l'échantillon bootstrap correspondant \Rightarrow le reste constitue les **données "out-of-bag"**.

C'est sur ces données "out-of-bag" que sont calculées :

- une estimation non-biaisée de l'erreur de l'arbre,
- une estimation de l'importance des facteurs de risque.

Ici, pas de validation croisée pour avoir une estimation non-biaisée de l'erreur : on prend les obs. et prévisions chaque fois qu'elles sont dans l'éch. OOB \rightarrow calcul erreur indiv. \rightarrow moy. erreurs indiv.

TIRAGE ALÉATOIRE COVARIABLES CHAQUE ÉTAPE

→ Randomization permet de diminuer la corrélation entre les arbres (rappel : les arbres sont ensuite agrégés), et de traiter le pb de covariables corrélées qui induisent un biais.

La variance de la moyenne de B estimateurs \perp (v.a.) vaut

$$\text{Var}\left(\frac{1}{B} \sum_{b=1}^B X_b\right) = \frac{1}{B^2} \text{Var}\left(\sum_{b=1}^B X_b\right) \simeq \frac{\sigma_b^2}{B}.$$

En revanche, si ces arbres sont corrélés 2 à 2, de coefficient de corrélation ρ , on obtient :

$$\text{Var}\left(\frac{1}{B} \sum_{b=1}^B X_b\right) \simeq \rho \sigma_b^2 + \frac{1-\rho}{B} \sigma_b^2.$$

Ainsi,

- si $\rho \rightarrow 0$, alors on retrouve le cas initial,
- si $\rho \rightarrow 1$, alors on a beau $\nearrow B$, il restera toujours $\rho\sigma^2$.

Cela limite donc fortement l'avantage du bagging... !

La procédure de bagging est encore plus fructueuse si p (nb de facteurs de risque) est grand !

Conclusion : lors de l'agrégation, on \searrow ainsi la variance de l'estimateur tout en conservant le même ordre de grandeur pour le biais...l'erreur globale de l'estimateur diminue donc !

Grâce à cette randomization, la stratégie d'élagage peut être + élémentaire qu'en pur bagging (avec d'autres modèles), on pourrait adopter la stratégie (2) d'élagage...

→ Pourquoi "randomiser"? prenons la variance de la moyenne de B variables iid;
 chacune de variance σ^2 . $\text{Var}\left(\frac{1}{B} \sum_{b=1}^B X_b\right) = \frac{1}{B^2} \text{Var}\left(\sum_{b=1}^B X_b\right) = \frac{B}{B^2} \text{Var}(X_b) = \frac{\sigma^2}{B}$
 Si les variables sont i.i.d., mais corrélées 2 à 2 avec corrélation ρ :

$$\text{Var}\left(\frac{1}{B} \sum X_b\right) = \frac{1}{B^2} \text{Var}\left(\sum X_b\right) = \frac{1}{B^2} \left[\underbrace{\text{Var}(X_1) + \dots + \text{Var}(X_B)}_{\sigma^2} + 2 \underbrace{\text{Cov}(X_1, X_2) + \dots + \text{Cov}(X_{B-1}, X_B)}_C \right]$$

$$= \frac{1}{B^2} \left[B \sigma^2 + 2 \sum_{1 \leq i < j \leq B} \text{Cov}(X_i, X_j) \right] \quad \text{avec} \quad \left(\text{Cov}(X_i, X_j) = E[(X_i - E(X_i))(X_j - E(X_j))] \right)$$

$$= \frac{\sigma^2}{B} + \frac{1}{B^2} \times 2 \times \sum_{1 \leq i < j \leq B} \sigma^2 \underbrace{\text{Corr}(X_i, X_j)}_C$$

$$\left(\text{Corr}(X_i, X_j) = \frac{\text{Cov}(X_i, X_j)}{\sqrt{\text{Var}(X_i)} \sqrt{\text{Var}(X_j)}} = \frac{\text{Cov}(X_i, X_j)}{\sigma^2} \right)$$

$$= \frac{\sigma^2}{B} + \frac{1}{B^2} \times 2 \times \rho \sigma^2 \times [1 + 2 + \dots + (B-2) + (B-1)]$$

nb de termes de la somme = $\sum_{j=1}^{B-1} j = (B-1) \times \frac{(1+B-1)}{2}$

$$= \frac{\sigma^2}{B} + \frac{1}{B^2} \times 2 \times \rho \sigma^2 \times (B-1) \frac{1+(B-1)}{2}$$

$$= \frac{\sigma^2}{B} + \frac{\rho \sigma^2}{B^2} [(B-1) + (B-1)^2] = \frac{\sigma^2}{B} + \rho \sigma^2 \left[\frac{1}{B} - \frac{1}{B^2} + \left(\frac{B^2}{B^2} - \frac{2B}{B^2} + \frac{1}{B^2} \right) \right]$$

$$= \frac{\sigma^2}{B} + \rho \sigma^2 \left[\frac{1}{B} + 1 - \frac{2}{B} \right] = \frac{\sigma^2}{B} + \rho \sigma^2 \left[1 - \frac{1}{B} \right] = \frac{\sigma^2}{B} + \rho \sigma^2 - \frac{\rho \sigma^2}{B}$$

$$= \left[\rho \sigma^2 + \frac{(1-\rho)}{B} \rho \sigma^2 \right], \text{ à comparer avec le cas } \underline{\underline{1}} = \frac{\sigma^2}{B}!$$

• si $\rho \rightarrow 1$, on a beau augmenter B , il restera toujours $\rho \sigma^2$! \Rightarrow limite
 l'avantage du bagging... \Rightarrow motive la randomisation par $\downarrow \rho$.

Preuve :

REMARQUES ADDITIONNELLES

→ La randomization permet de gérer également les covariables corrélées.

→ L'importance des facteurs de risque peut être calculée de 2 façons différentes.

- Mesurer l'importance par permutation des covariables (**shuffling**).

→ Permutation aléatoire des valeurs de la covariable entre individus, puis on prédit : + la qualité de prévision est dégradée, + le facteur de risque est important.

En pratique, on calcule l'erreur OOB du b^e arbre sans et avec permutation de la covariable, puis on regarde l'écart. Puis on moyenne sur tous les arbres.

- Utiliser pour chaque variable dans chaque arbre la valeur de décroissance de l'indice de Gini.

En pratique : il est + simple de moyenner la \searrow de Gini car elle est déjà calculée lors de la construction de l'arbre.

→ Gestion des **données manquantes** : imputées comme suit,

- échantillon d'apprentissage : moyenne ou proximités ;
- échantillon de valid. : \neq suivant que l'on observe Y ou non.

Rq : le fichier d'aide de `randomForest` détaille tout cela...

RÉSUMÉ SUR LE BAGGING

Finalement, le principe du bagging présente des avantages et des inconvénients...

- (+) Simple à mettre en oeuvre et à comprendre ;
- (+) Se programme facilement, qlq soit la méthode ;
- (+) Diminue la variance de l'estimateur ;
- (-) Temps de calcul parfois important : nécessité d'agréger un grand nombre de modèles avant de stabiliser l'erreur OOB ;
- (-) stockage de tous les modèles (mémoire...) ;
- (-) Perte de l'interprétabilité, sorte de boîte noire.

5 Bagging + randomization de CART : forêts aléatoires

- Principe
- Construction de la forêt aléatoire
- Force, corrélation et erreur de la forêt
- Interprétabilité de modèles ensemblistes

INTERPRETABILITE

Méthode de LIME

Méthode de SHAP (Shapley)