

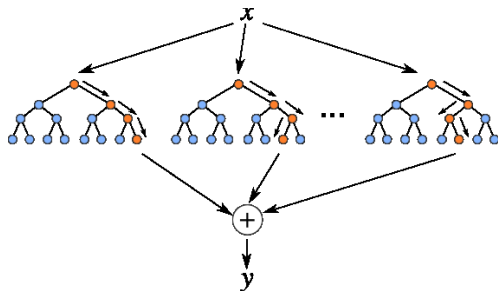
Apprentissage supervisé Forêts aléatoires.

Marie Chavent

Université de Bordeaux

Le Bagging

- ▶ En 1996 Breiman introduit le **Bagging**, une des premières **méthodes d'ensemble** qui consiste à **agrèger une collection de classifieurs** pour en obtenir un meilleur.
- ▶ En classification, on agrège les prédictions par **vote majoritaire**.
- ▶ Bagging=**Bootstrap aggregating** : principe non limité aux arbres de classification mais **CART** est le classifieur de base le plus couramment utilisé.



La méthode :

- ▶ On construit q **échantillons bootstrap** $E^{(1)}, \dots, E^{(q)}$ à partir de l'échantillon E des observations $(X_i, Y_i)_{i=1, \dots, n}$.
- ▶ Un échantillon bootstrap $E^{(\ell)}$ est obtenu par **tirage avec remise** de n observations dans E , chaque observation ayant une probabilité $\frac{1}{n}$ d'être tirée à chaque tirage.
- ▶ Pour $\ell = 1, \dots, q$, on **constitue l'arbre CART** g_ℓ à partir de l'échantillon bootstrap $E^{(\ell)}$.
- ▶ On **aggrège les prédictions** des arbres g_1, \dots, g_q par vote majoritaire :

$$g_{\text{bag}}(x) = \arg \max_{k \in \{1, \dots, K\}} \sum_{\ell=1}^q \mathbb{1}_{g_\ell(x)=k}$$

La méthode :

- ▶ On construit q échantillons bootstrap $E^{(1)}, \dots, E^{(q)}$ à partir de l'échantillon E des observations $(X_i, Y_i)_{i=1, \dots, n}$
- ▶ Pour $\ell = 1, \dots, q$, on construit une variante de l'arbre CART appelée RI (Random Input) et à partir de l'échantillon bootstrap $E^{(\ell)}$. L'arbre g_ℓ^{RI} est obtenu de la manière suivante :
 - à chaque division, on tire aléatoirement m variables parmi p (sans remise) pour construire les questions binaires.
 - on construit l'arbre de longueur maximale (pas d'élagage).
- ▶ On agrège des arbres g_1, \dots, g_q ainsi obtenus avec cette variante par vote majoritaire :

$$g_{RF}(x) = \arg \max_{k \in \{1, \dots, K\}} \sum_{\ell=1}^q \mathbb{1}_{g_\ell^{RI}(x)=k}$$

Le taux d'erreur OOB (Out Of Bag) de la forêts aléatoire est calculé de la manière suivante :

- ▶ Pour chaque observation (X_i, Y_i) , $i = 1, \dots, n$:
 - on sélectionne les échantillons bootstrap $E^{(\ell)}$ ne contenant pas (X_i, Y_i) . Pour ces échantillons, on dit que l'observation (X_i, Y_i) est OOB.
 - on prédit cette observation avec tous les arbres construits sur ces échantillons bootstrap.
 - on agrège ces prédictions par vote majoritaire et on note \hat{Y}_i la prédiction OOB de X_i .
- ▶ Le taux d'erreur OOB de g_{RF} est alors : $err_{OOB} = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\hat{Y}_i \neq Y_i}$.

L'échantillon OOB d'un arbre g_ℓ^{RI} est l'ensemble des observations de E qui ne sont pas dans $E^{(\ell)}$. On notera :

- n_ℓ le nombre d'observations OOB de $E^{(\ell)}$,
- $E_{oob}^{(\ell)}$ l'échantillon OOB de l'arbre g_ℓ^{RI} .

L'importance d'une variable X^j est calculée de la manière suivante :

- ▶ Pour chaque échantillon $E_{oob}^{(\ell)}$:
 - ▶ on calcule $errOOB^{(\ell)}$, le taux d'erreur de g_{ℓ}^{RI} sur $E_{oob}^{(\ell)}$.
 - ▶ on permute aléatoirement les valeurs de X^j selon une permutation ϕ :

$$\mathbf{x}_{oob}^{(\ell)}(\phi) = \begin{matrix} 1 \\ 2 \\ \vdots \\ n_j \end{matrix} \begin{pmatrix} 1 & \cdots & j & \cdots & p \\ \cdots & & X_{\phi(1)}^j & & \cdots \\ \cdots & & X_{\phi(2)}^j & & \cdots \\ \vdots & & \vdots & & \vdots \\ \cdots & & X_{\phi(n_j)}^j & & \cdots \end{pmatrix}$$

et on calcule $errOOB_j^{(\ell)}$ le taux d'erreur de g_{ℓ}^{RI} sur $E_{oob,j}^{(\ell)}$ l'échantillon $E_{oob}^{(\ell)}$ permuté sur X^j .

- ▶ on calcule l'augmentation du taux d'erreur induite par la permutation des valeurs de X^j : $errOOB_j^{(\ell)} - errOOB^{(\ell)}$.
- ▶ L'importance VI de la variable X^j est alors une mesure de lien entre Y et X^j définie par :

$$VI(X^j) = \frac{1}{q} \sum_{\ell=1}^q (errOOB_j^{(\ell)} - errOOB^{(\ell)}).$$

Remarque : $VI(X^j)$ est parfois la moyenne des augmentations divisée par leur écart-type.