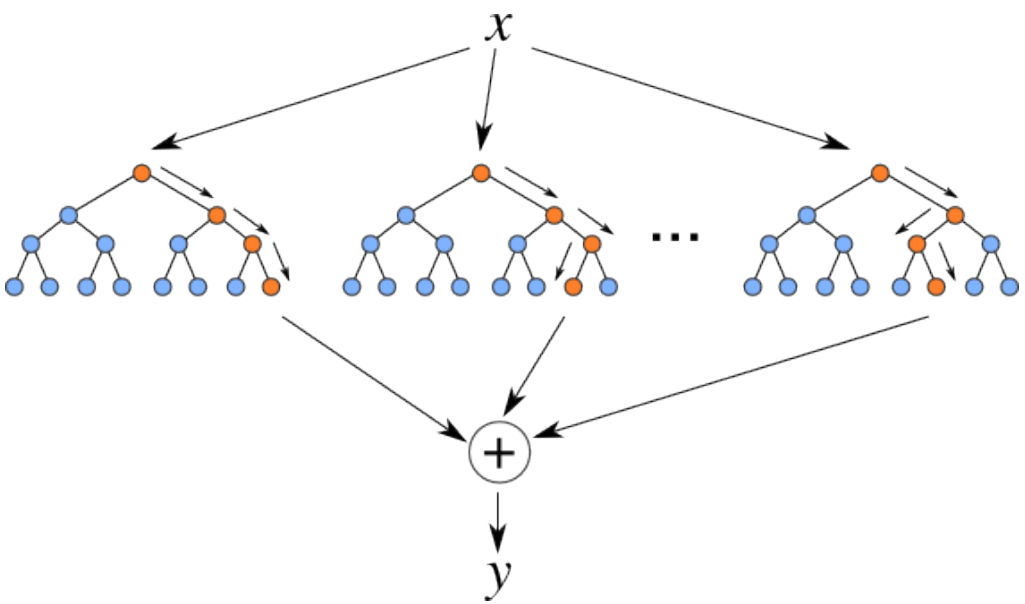
Apprentissage supervisé
Forêts aléatoires.

Marie Chavent

Université de Bordeaux

Le Bagging

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



La méthode :

- On construit q échantillons bootstrap $E^{(1)}, ..., E^{(q)}$ à partir de l'échantillon E des observations $(X_i, Y_i)_{i=1,...,n}$.
- 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 $\ell = 1, ...q$, on constuit l'arbre CART g_ℓ à partir de l'échantillon bootstrap $E^{(\ell)}$.
- On aggréger les prédictions des arbres $g_1, \ldots, g_q$ par vote majoritaire :

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

Forêts aléatoires

La méthode :

- On construit q échantillons bootstrap $E^{(1)}, ..., E^{(q)}$ à partir de l'échantillon E des observations $(X_i, Y_i)_{i=1,...,n}$
- Sur chaque échantillon boostrap $E^{(\ell)}$ on applique une variante de CART appelée RI (Random Input) :
  - à chaque division, on tire aléatoirement m variables variables parmi p (sans remise) pour constuire les questions binaires.
  - on construit l'arbre de longueur maximale (pas d'élagage).
- On aggréger des arbres $g_1, \ldots, g_q$ ainsi obtenus avec cette variante par vote majoritaire :

$$g_{\mathrm{RF}}(x) = \underset{k \in \{1,\ldots,K\}}{\arg\max} \sum_{\ell=1}^{q} \mathbb{1}_{g_\ell(x)=k}$$

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), $i = 1, \ldots, n$:
  - 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 aggrège 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 : $errOOB = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\hat{Y}_i \neq Y_i}$.

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

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

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 sur $E_{oob}^{(\ell)}$.
  - on permute aléatoirement les valeurs de X^j selon une permutation ϕ :

$$\mathbf{x}_{oob}^{(\ell)}(\phi) = \begin{array}{c} \\ 1 \\ 2 \\ \vdots \\ n_l \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & \cdots & j & \cdots & p \end{array} \\ \begin{pmatrix} \cdots & X^j_{\phi(1)} & \cdots \\ \cdots & X^j_{\phi(2)} & \cdots \\ & \vdots & \\ \cdots & X^j_{\phi(n_l)} & \cdots \end{pmatrix} \end{array}$$

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

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