Arbre CART, bagging et boosting

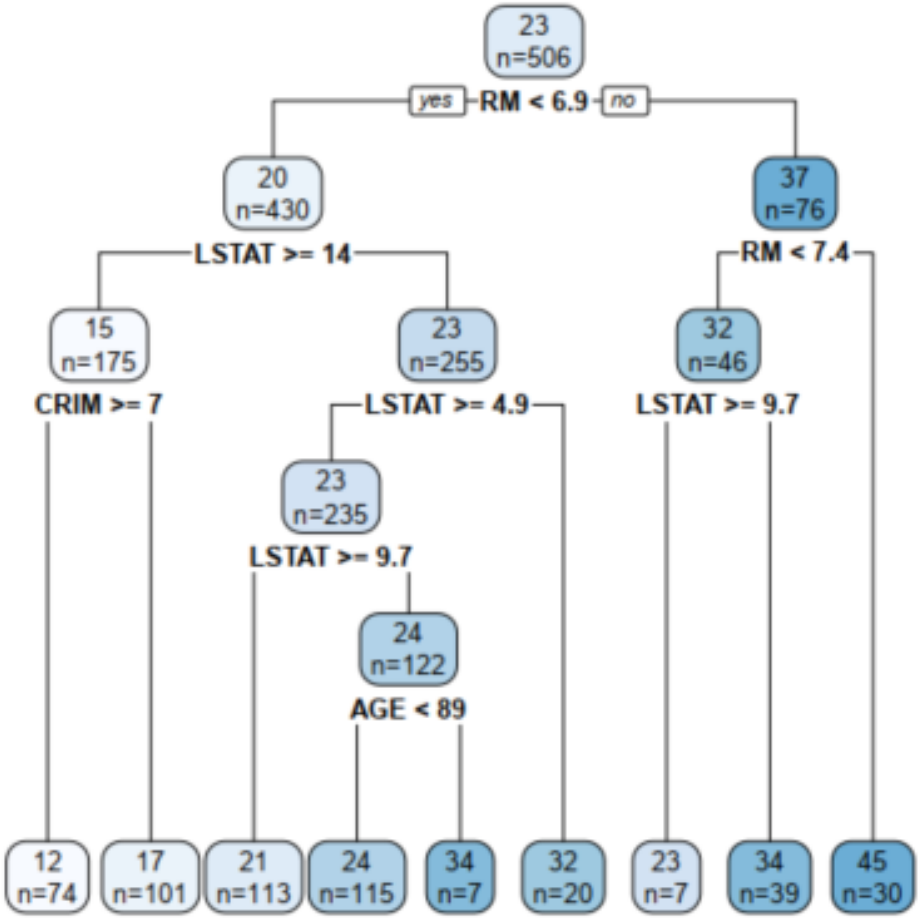
Arnaud Callebaut & Lionel Hertzog

30/04/2025

Au commencement - les arbres CART

Définitions

- Classification And Regresstion Trees (Breiman 1984)
- Partition séquentielle de l'espace des variables explicatives (X) de manière à homogénéiser la variable de réponse (y)
- Méthode algorithmique non-paramétrique, pouvant avoir des variables y/X discrètes et/ou continues

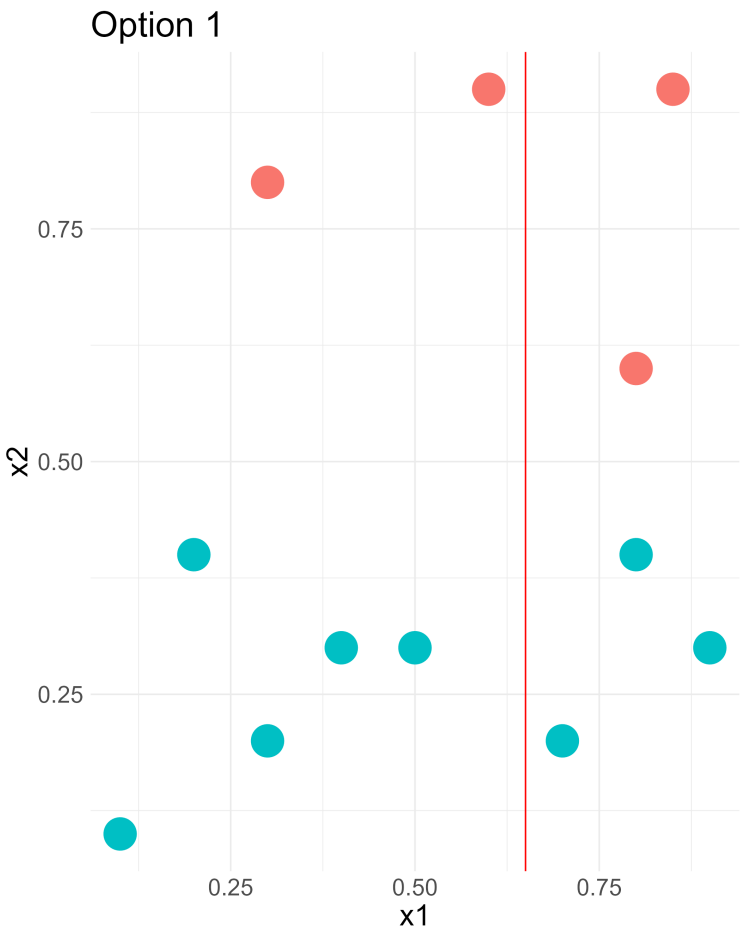


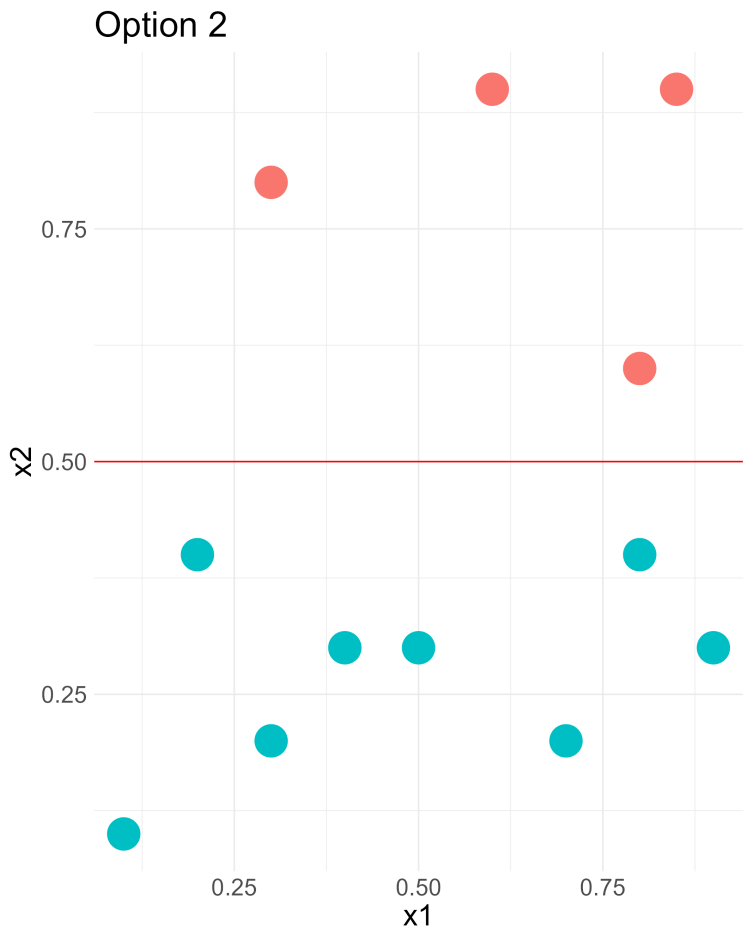
Un arbre - deux éléments

Pour faire un arbre il faut:

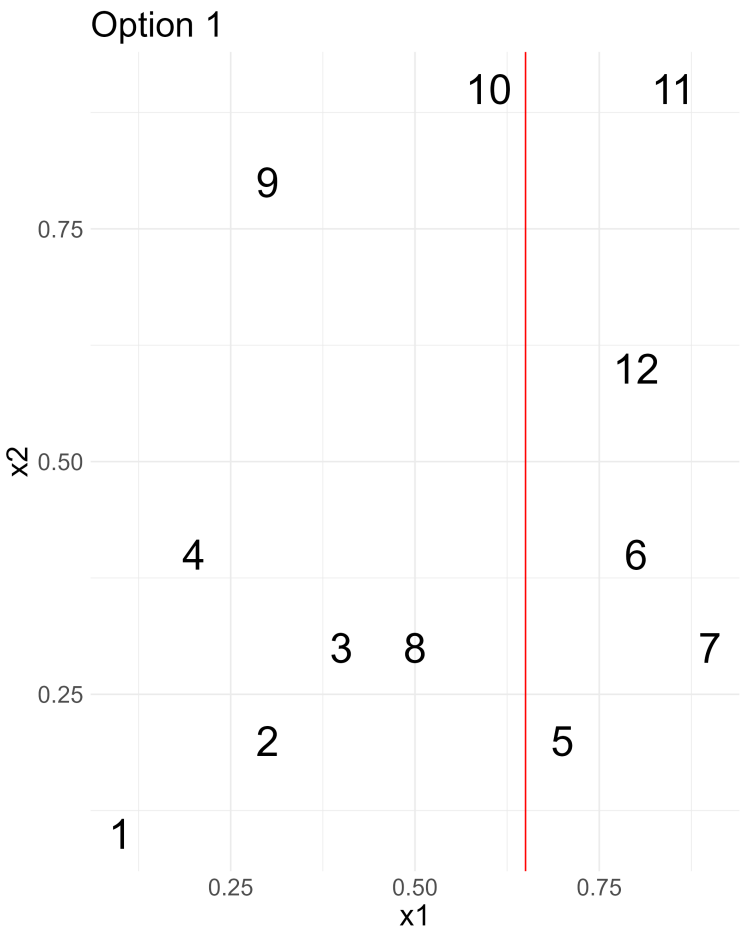
1. Une fonction mesurant l'homogénéité des découpages successifs
2. Un critère d'arrêt pour décider du moment où l'arbre s'arrête

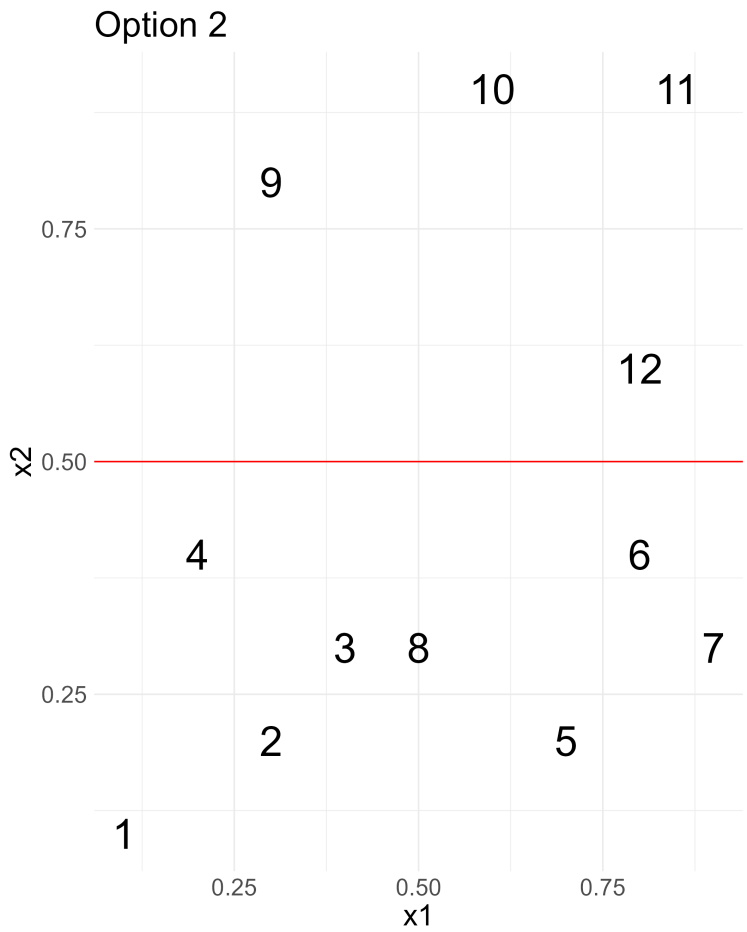
Le découpage - données discrètes





Le découpage - données continues





Les fonctions d'homogénéité

Ces fonctions quantifient l'homogénéité des feuilles crées par les découpages potentiels.

Données continues

$$H(d) = \frac{1}{N_d} * \sum_{i:Y \in d} (Y_i - \bar{Y}_d)^2 \quad (1)$$

En régression, l'homogénéité d'une découpe d est la variance.

Données discrètes

$$H(d) = 1 - \sum_{i=1}^{M} (\frac{N_i}{N})^2 \quad (2)$$

En classification (M classes), par défaut la fonction de gini est utilisée.

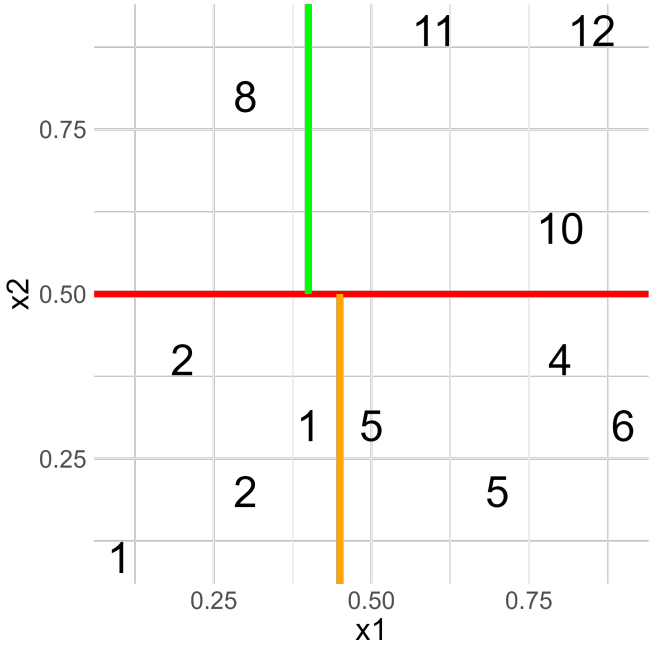
Le découpage

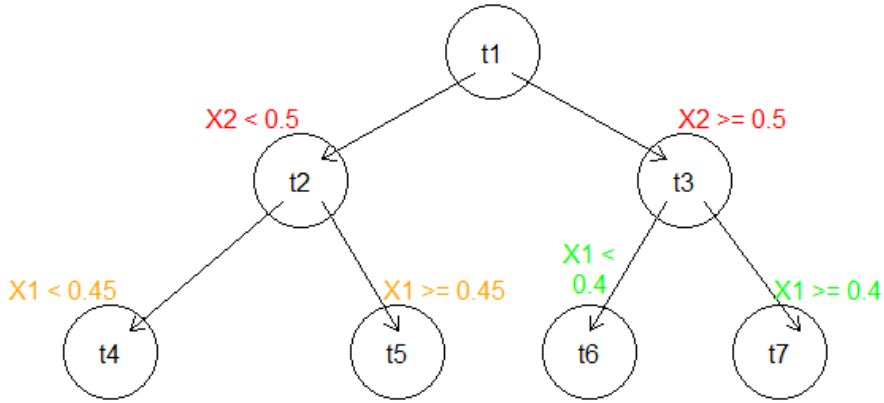
Pour sélectionner un découpage l'algorithme suivant est utilisé:

1. Calcul de l'homogénéité avant découpage
2. Calcul de l'homogénéité de la feuille droite
3. Calcul de l'homogénéité de la feuille gauche
4. Calcul de l'augmentation de l'homogénéité (pondérée) après découpage

L'algorithme sélectionne le découpage menant à la plus grande augmentation de l'homogénéité.

Découpage séquentiel





Note

Le découpage marche aussi sur des variables explicatives discrètes

Jusqu'où faire monter l'arbre ?

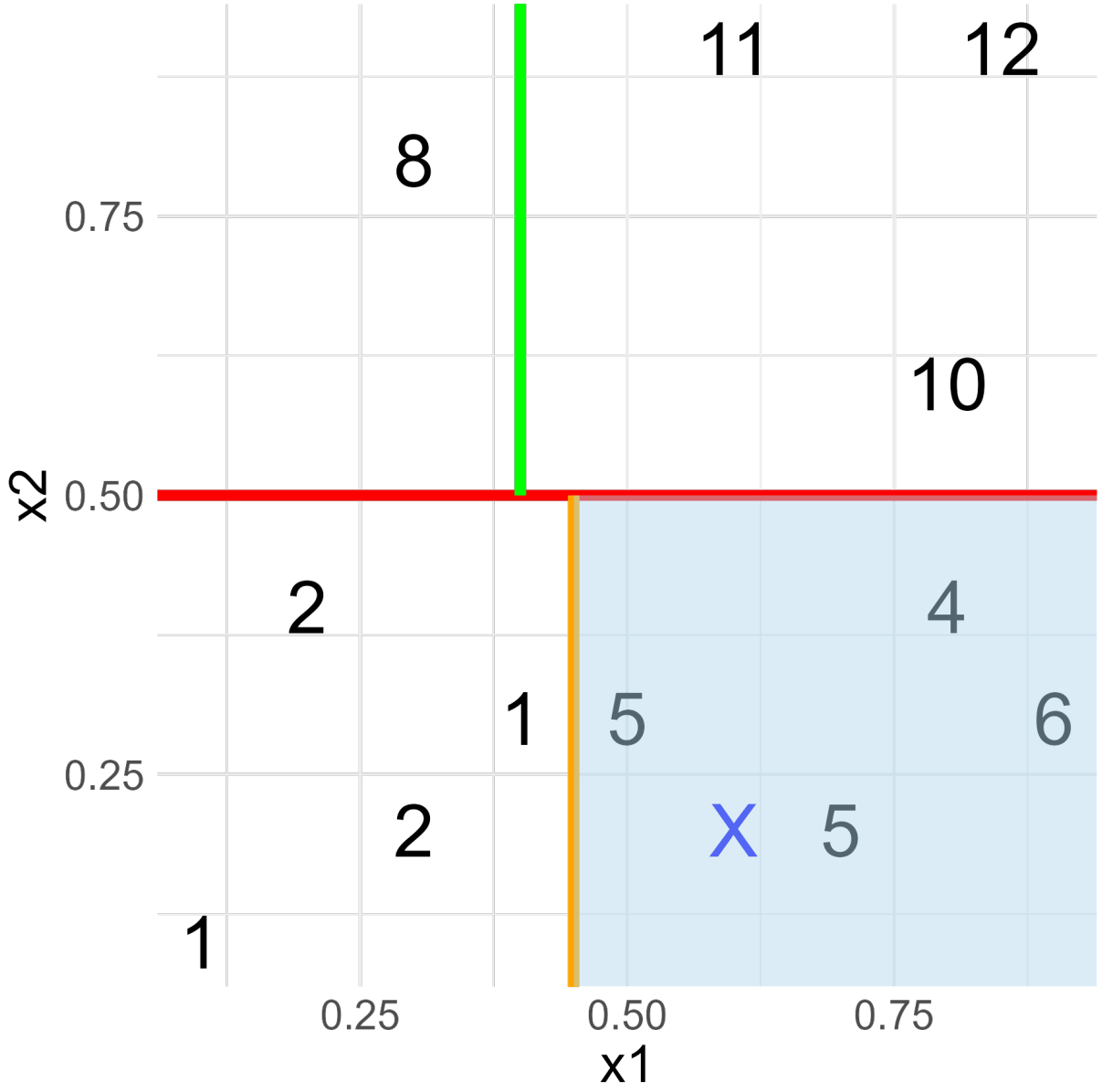
Critère d'arrêt

L'algorithme CART crée des feuilles jusqu'à ce qu'un critère d'arrêt défini a priori soit atteint

- Le nombre d'observation dans les feuilles terminales sont inférieurs à un seuil fixé (quid de nmin $= 1$?)
- Si le meilleur découpage des feuilles terminales mènent à une augmentation de l'homogénéité inférieur à un seuil fixé
- Si l'homogénéité dans les feuilles terminales est supérieur à un seuil fixé

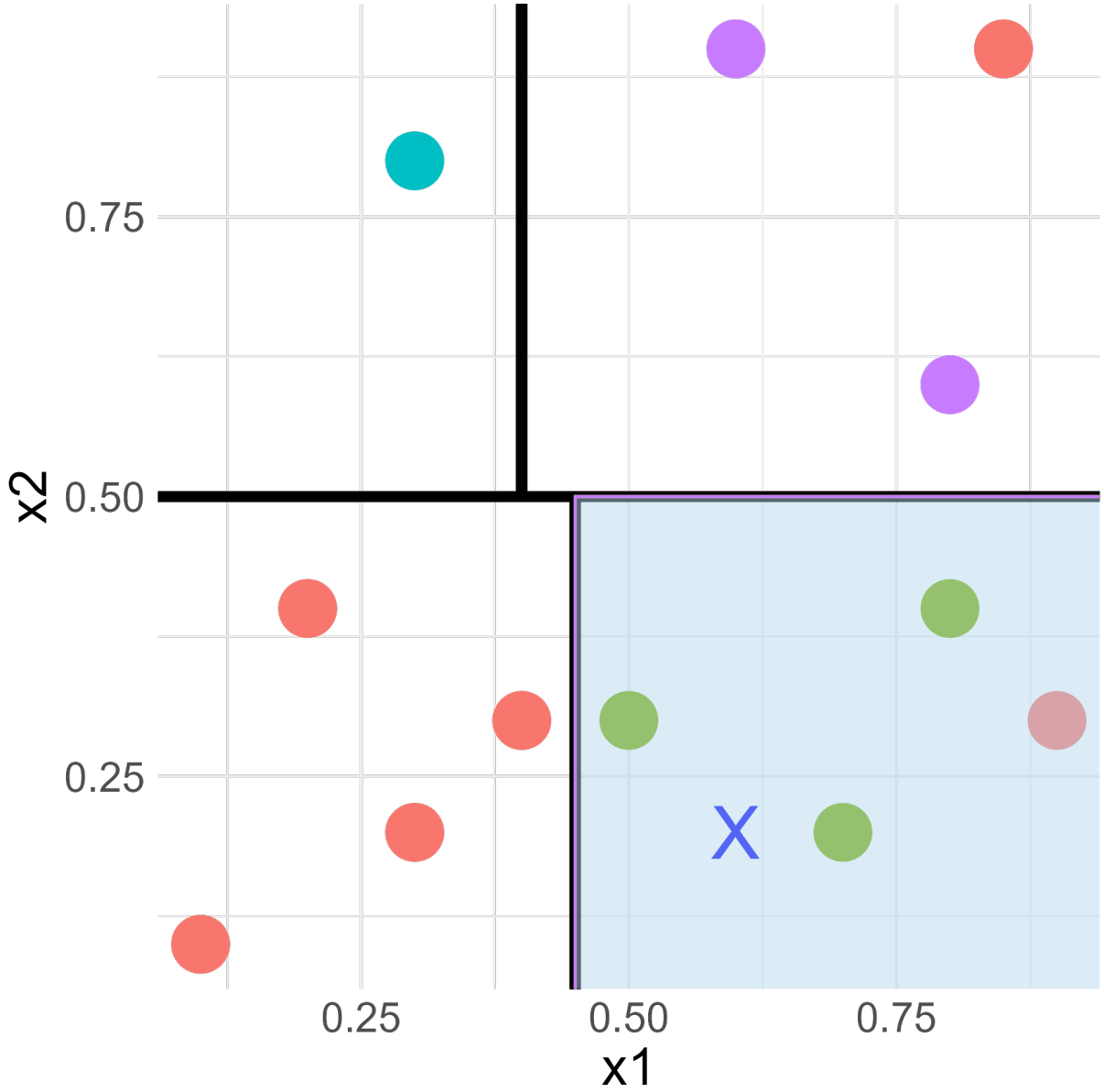
Prédire à partir d'un arbre CART

- Donnée continue : la moyenne des observations tombant dans la feuille terminale



Prédire à partir d'un arbre CART

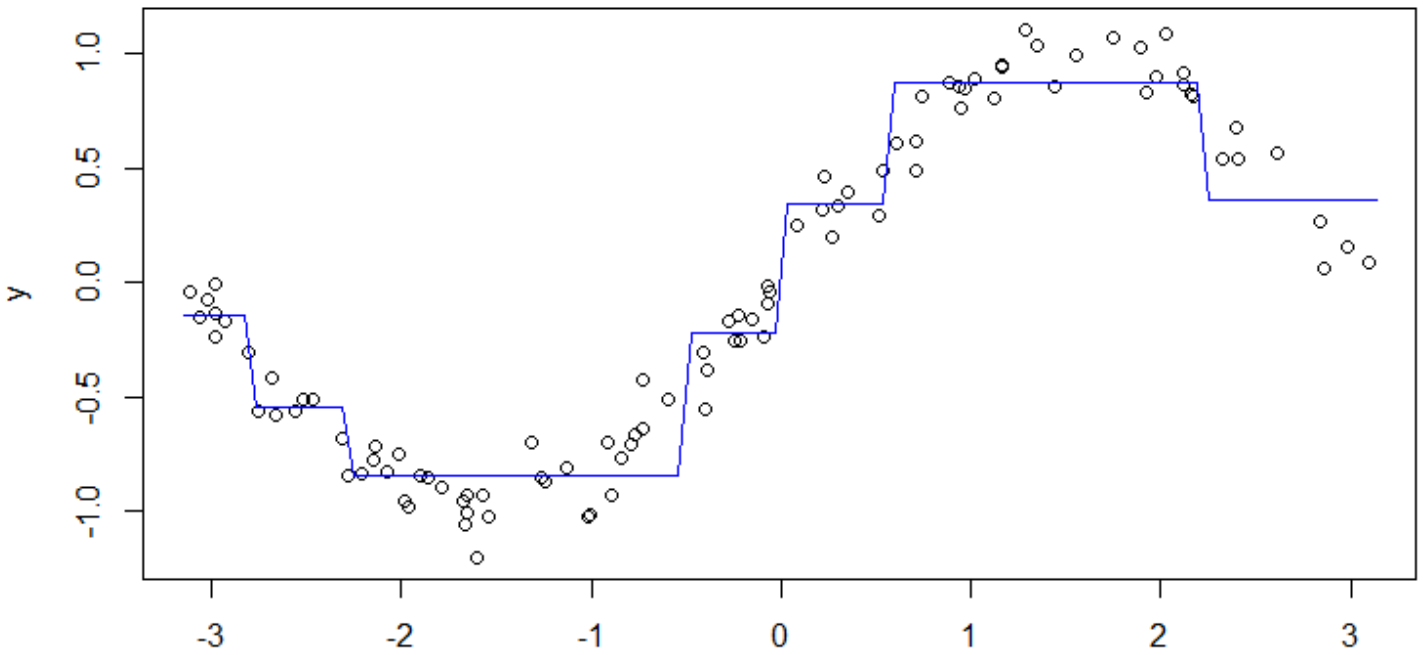
- Donnée continue : la moyenne des observations tombant dans la feuille terminale
- Donnée discrète : la classe majoritaire des observations tombant dans la feuille terminale



Méthode d'ensemble

Problème avec CART

- Le partitionnement de l'espace des variables crée des relations entre X et y par palier
- Un modèle unique peut souffrir de sur- ou sous-apprentissage et donc avoir de faibles capacités de généralisation



Méthode d'ensemble

Idées principales

- Construire q prédicteurs (modèles CART p.ex.) puis agréger les prédictions
- Agréger permet de lisser les relations mais également de réduire la variance des prédictions
- Cela nécessite de créer des prédicteurs avec une corrélation faible

Soit $\hat{f}_1$, $\hat{f}_2$, ..., $\hat{f}_q$, q prédicteurs avec une variance σ^2 et une corrélation ρ :

$$\mathbb{E}[\bar{f}] = \mathbb{E}[\hat{f}_1] \tag{3}$$

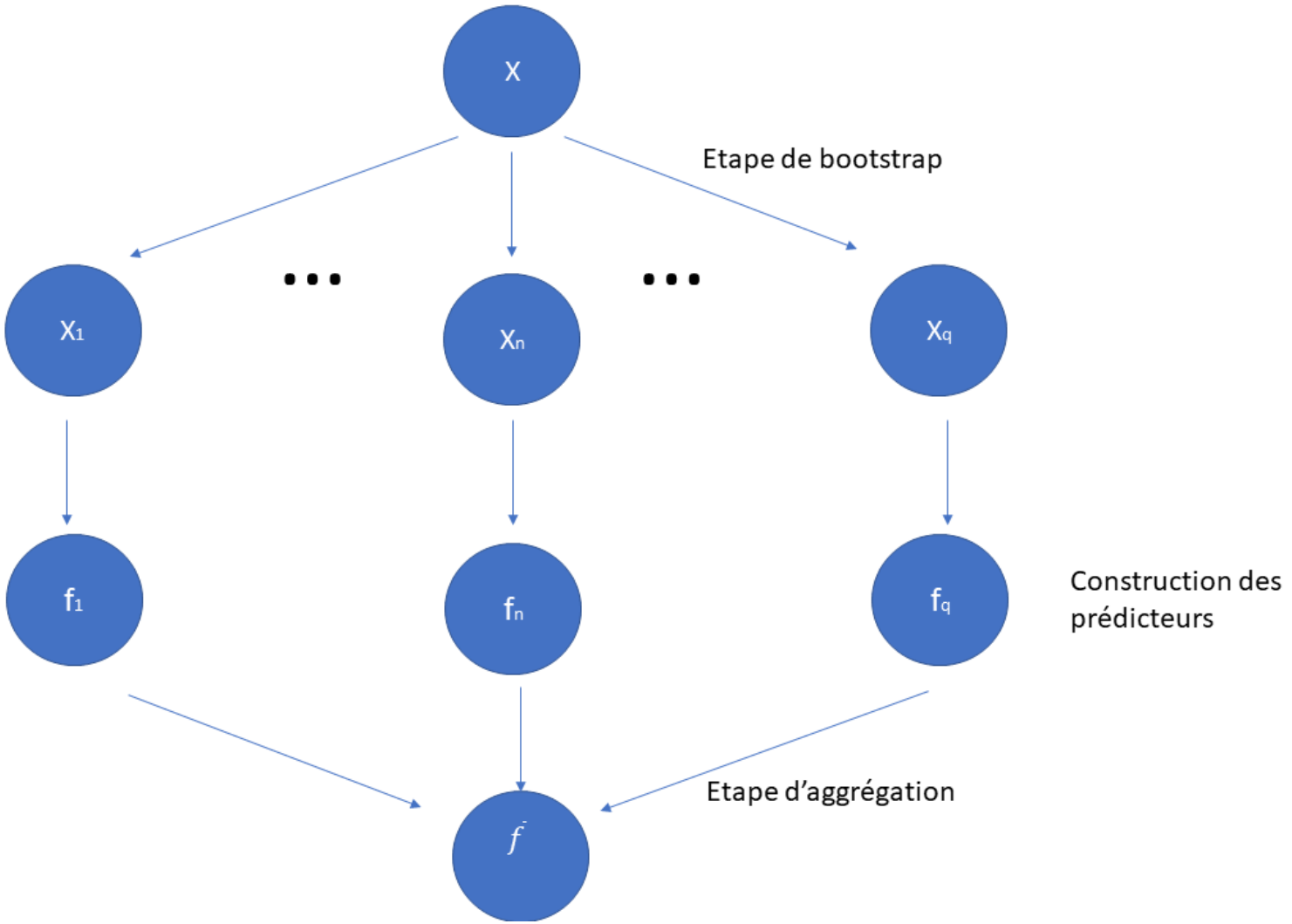
$$Var(\bar{f}) = \sigma^2 * \rho + \frac{1-\rho}{q} * \sigma^2 \tag{4}$$

Méthodes d'ensembles

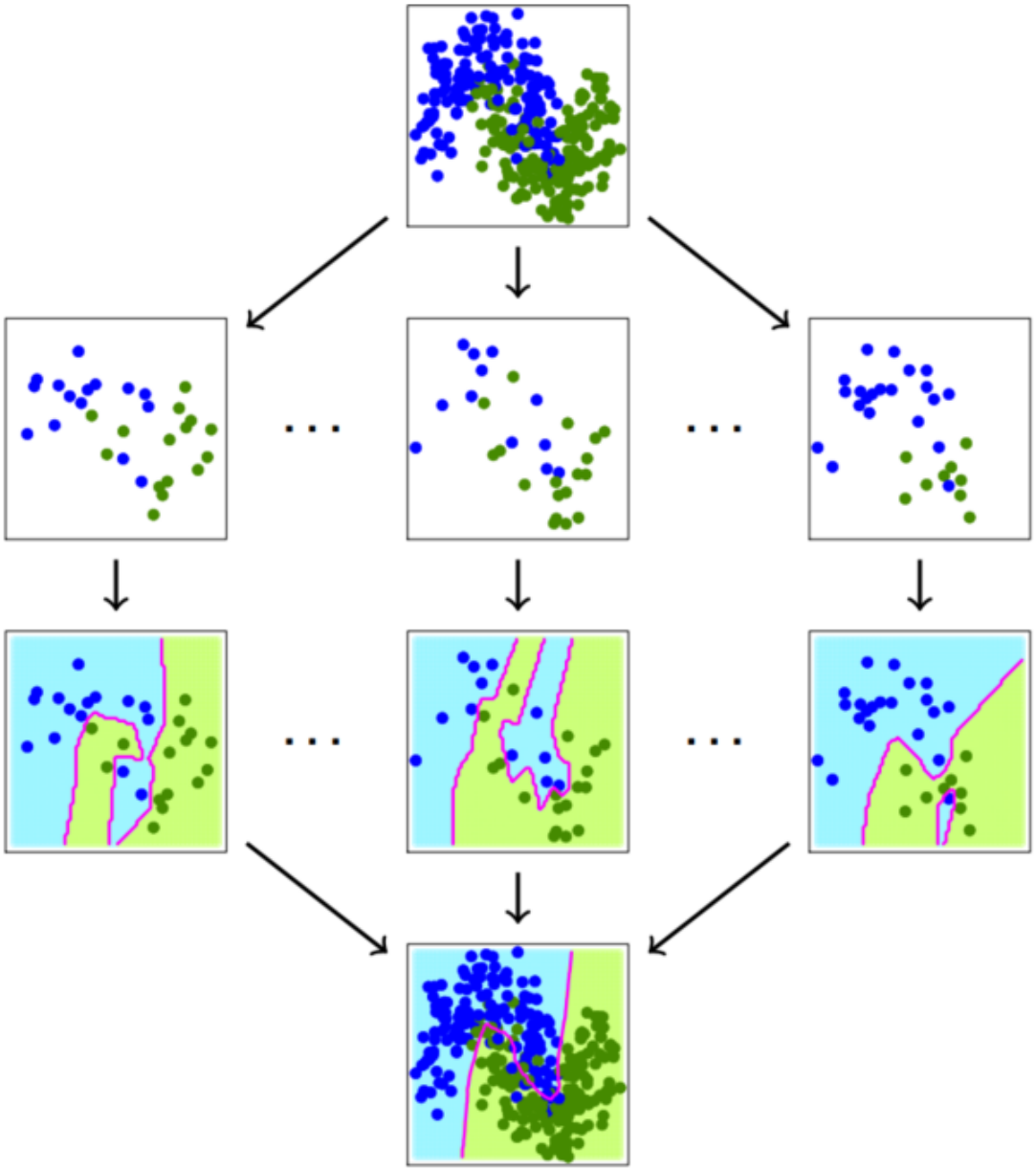
Deux méthodes potentielles

- Bagging
- Boosting

Bagging


X
Etape de bootstrap
X1
Xn
Xq
f1
fn
fq
Construction des prédicteurs
Etape d'aggrégation
f^

Bagging



Bagging en détails

Etape 1 : Bootstrap

Pour chaque prédicteur, on tire aléatoirement n observations **avec remise** a.k.a. on fait un bootstrap. Ceci crée de l'aléatoire (réduit la corrélation) entre les prédicteurs

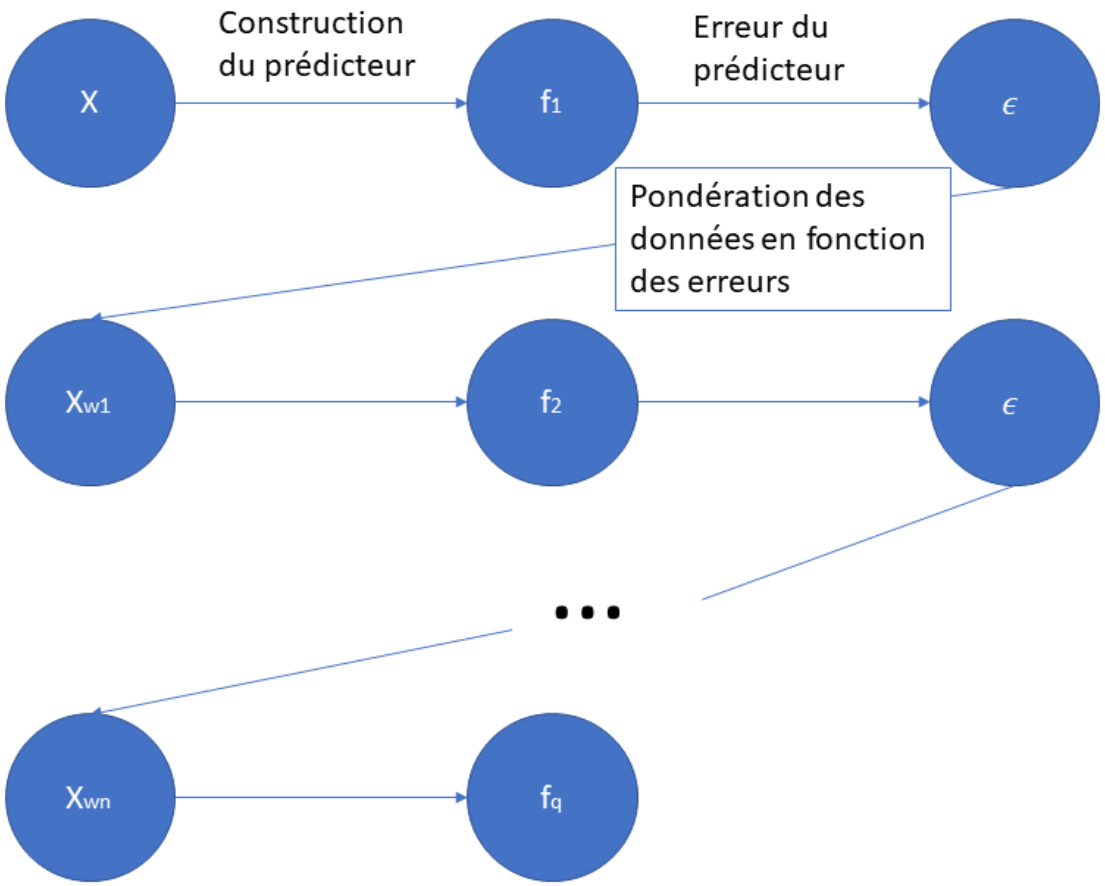
Etape 2 : le modèle

Sur les données bootstrap on entraine q modèle. Cela peut être n'importe quel type de modèle (lm, kNN, CART ...).

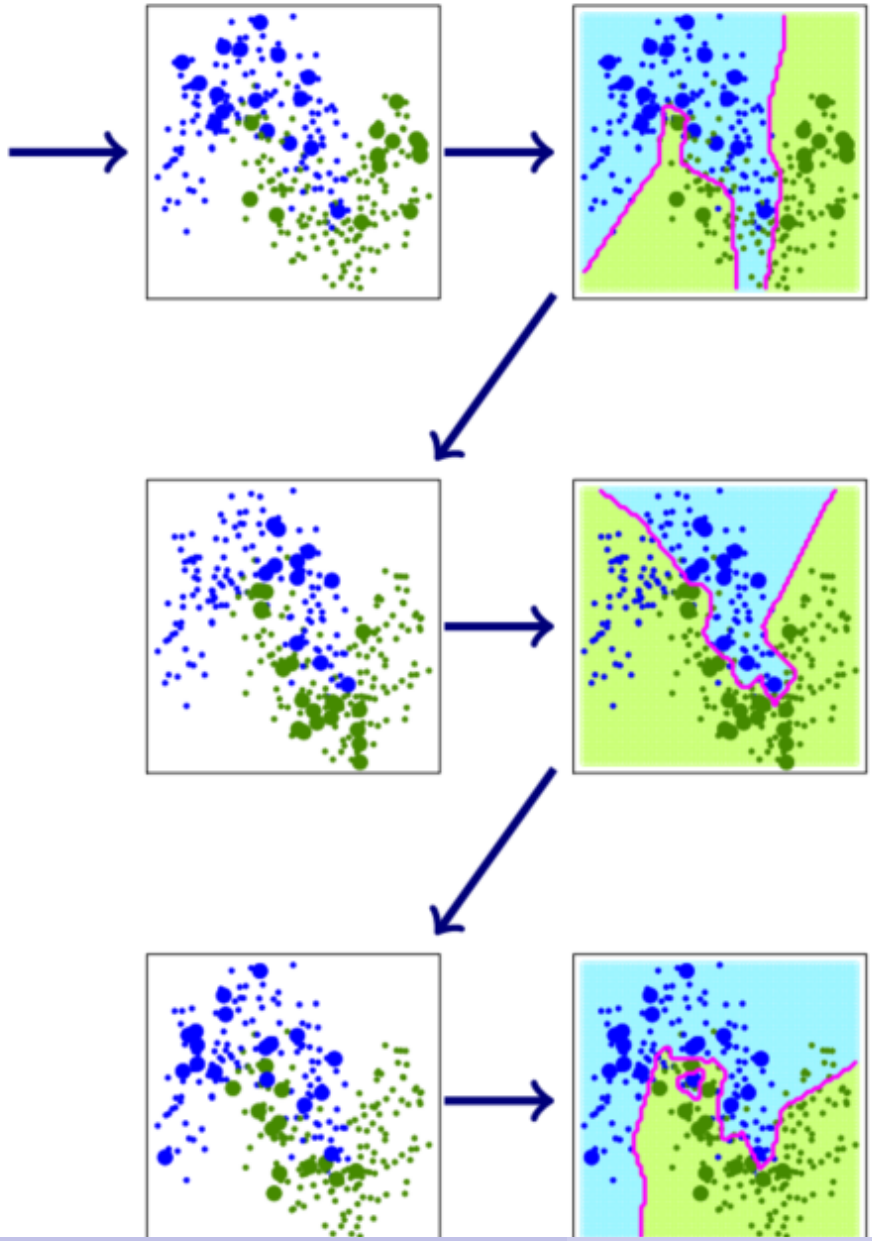
Etape 3 : aggrégation

Les prédictions des q prédicteurs sont aggrégées pour produire une prédiction. Données continues : moyenne, données discrètes : classe majoritaire

Boosting


X
Construction du prédicteur
f_1
Erreur du prédicteur
ϵ
Pondération des données en fonction des erreurs
X_{w1}
f_2
ϵ
...
X_{wn}
f_q

Boosting



Boosting en détails

Le boosting est une approche séquentielle en 3 étapes se répétant :

1. Construction d'un modèle (lm, kNN, CART) pondéré
2. Extraction des erreurs du modèle
3. Pondération des observations en fonction des erreurs (retour à l'étape 1)

Le modèle est graduellement amélioré en se focalisant sur les observations mal prédites à l'étape précédente.

What we saw so far

- CART : approche algorithmique et non-paramétrique permettant de modéliser des relations entre des variables X (continus et discret) et une variable y (continu ou discrète)
- Cette approche se base sur une partition séquentielle de l'espace de variable de manière à maximiser l'homogénéité de la réponse dans les feuilles
- Les méthodes d'ensemble permettent d'agréger des prédicteurs et d'obtenir des prédictions plus robustes et avec une variance plus faible
- Le bagging est une méthode d'ensemble basé sur une étape de bootstrap des données, d'une construction en parrallèle de plusieurs prédicteurs et de l'aggrégation des prédictions.
- Le boosting est basé sur une amélioration séquentielle d'un modèle par pondération succesive des observations en fonction des erreurs du modèle.