

# INTRODUCTION AU MACHINE LEARNING

## ARBRE ET ENSEMBLE

**Théo Lopès-Quintas**

BPCE Payment Services,  
Université Paris Dauphine

2022-2024

# ARBRE

## MODÉLISATION

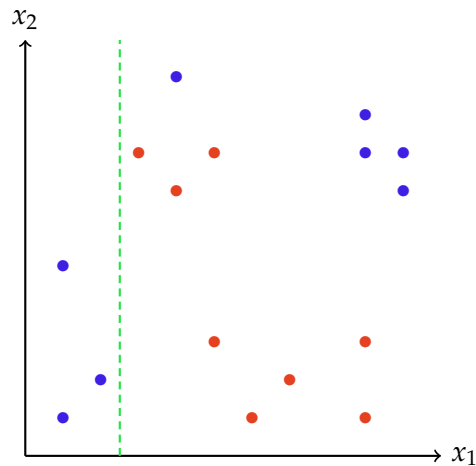
Probabilité de la classe d'intérêt dans la partition  $P$

$$f_{\theta}(x) = \sum_{P \in \theta} \mu_P \mathbb{1}_{\{x \in P\}}$$

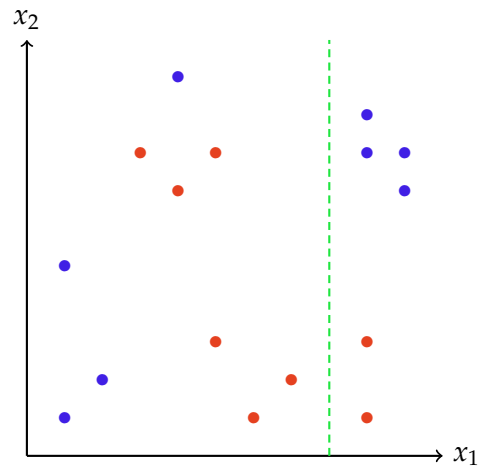
Partition de l'espace

# ARBRE

## MEILLEURE SÉPARATION



(a) Séparation 1



(b) Séparation 2

**Figure** – Sélection de la meilleure séparation de l'espace pour une information donnée

# ARBRE

## EXERCICES

### Exercice 1 (Difficulté à extrapoler)

*Exhiber un exemple de problème de régression (à construire) où une régression linéaire réussit à extrapoler, mais pas un arbre de décision.*

# ARBRE

## EXERCICES

### Exercice 1 (Difficulté à extrapoler)

*Exhiber un exemple de problème de régression (à construire) où une régression linéaire réussit à extrapoler, mais pas un arbre de décision.*

### Exercice 2 (Faire communiquer deux algorithmes)

*On souhaite prédire des prix de certaines crypto-monnaies qui sont connues pour être particulièrement volatiles. L'enjeu d'estimer au mieux le prix est donc fort : une bonne prédiction peut donner lieu à un grand gain, et une mauvaise prédiction une grande perte. On sollicite notre équipe de data-scientists, et ils nous présentent deux algorithmes :*

- ▶ **Régression Linéaire** : marche plutôt bien, et reste raisonnablement correcte pour les grandes variations de prix
- ▶ **Arbre de régression** : marche beaucoup mieux quand les prix sont dans les moyennes, mais est très mauvais dès qu'on sort des prix moyens

*Expliquer succinctement pourquoi les comportements relatifs étaient prévisibles, et proposer des solutions pour utiliser les deux algorithmes ensembles et faire mieux que les deux séparément.*

# MÉTHODES ENSEMBLISTES

## BAGGING

$$\text{MSE}(y, \hat{f}(x)) = \left( \text{Bias} [\hat{f}(x)] \right)^2 + \mathbb{V} [\hat{f}(x)] + \sigma^2 \quad (\text{Trade off biais-variance})$$

### Exercice 3 (Intérêt du bagging)

Supposons que l'on traite un problème de régression, que l'on dispose de  $m$  régresseurs  $(f_k)_{k \leq m}$  chacun entraîné sur  $m$  échantillons issus de la distribution engendrée par le dataset. On construit un régresseur fort à partir de ces modèles :

$$F(x) = \frac{1}{m} \sum_{k=1}^m f_k(x)$$

1.  $\mathbb{E} [F(x)] = \mathbb{E} [f_k(x)]$  pour n'importe quel  $k \leq m$ .
2.  $\mathbb{V} [F(x)] = \frac{1}{m} \mathbb{V} [f_k(x)]$  pour n'importe quel  $k \leq m$ .
3. Conclure sur l'intérêt de la méthode proposée.

# MÉTHODES ENSEMBLISTES

## BAGGING

Supposons que l'on traite un problème de régression, que l'on dispose de  $m$  régresseurs  $(f_k)_{k \leq m}$  chacun entraîné sur  $m$  échantillons issus de la distribution engendrée par le dataset. On construit un régresseur fort à partir de ces modèles :

$$F(x) = \frac{1}{m} \sum_{k=1}^m f_k(x)$$

Alors si l'on suppose que les variables  $(f_k(x))_{k \leq m}$  sont indépendantes et identiquement distribuées, on a :

$$\mathbb{V}[F(x)] = \frac{1}{m} \mathbb{V}[f_k(x)]$$

En réalité, les datasets ne sont pas parfaitement indépendants et identiquement distribués. Si l'on considère qu'ils sont corrélé d'une valeur  $\rho$ , alors :

$$\mathbb{V} \left[ \frac{1}{m} \sum_{k=1}^m f_k(x) \right] = \frac{1}{m} (1 - \rho) \sigma^2 + \rho \sigma^2$$

### Exercice 4 (Cohérence)

Vérifier que la formule est cohérente avec le cas où les datasets sont indépendants. Que se passe-t-il quand les datasets sont parfaitement corrélés ?