



# IFSBM Module 11



**IFSBM**  
INSTITUT DE FORMATION  
SUPÉRIEURE BIOMÉDICALE

Méthodes de classification supervisée

Yoann Pradat

# Sommaire

---

## 1. Données et modélisation

1. Le jeu de données
2. L'estimation (model fitting)

## 2. Modèles de classification linéaires

1. Régression logistique binaire
2. Régression logistique multinomiale
3. Régression linéaire
4. Régression linéaire lasso

## 3. Modèles de classification non linéaires

1. Régression logistique
2. Arbre de décision CART
3. Forêts aléatoires

# 1.1 Le jeu de données

---

## Notations

$X : \Omega \mapsto \mathcal{X}$  : variable aléatoire (scalaire ou vectorielle)

$x \in \mathcal{X}$  : une observation de la variable aléatoire

### Exemples:

1. On lance un dé,  $\mathcal{X} = \{1, 2, \dots, 6\}$

# 1.1 Le jeu de données

---

## Notations

$X : \Omega \mapsto \mathcal{X}$  : variable aléatoire (scalaire ou vectorielle)

$x \in \mathcal{X}$  : une observation de la variable aléatoire

### Exemples:

1. On lance un dé,  $\mathcal{X} = \{1, 2, \dots, 6\}$

2. On mesure la longueur et la largeur de fleurs  $\mathcal{X} = \mathbb{R}^+ \times \mathbb{R}^+$

# 1.1 Le jeu de données

---

## Notations

$X : \Omega \mapsto \mathcal{X}$  : variable aléatoire (scalaire ou vectorielle)

$x \in \mathcal{X}$  : une observation de la variable aléatoire

### Exemples:

1. On lance un dé,  $\mathcal{X} = \{1, 2, \dots, 6\}$

2. On mesure la longueur et la largeur de fleurs  $\mathcal{X} = \mathbb{R}^+ \times \mathbb{R}^+$

3. On mesure le niveau d'expression de gènes relatif à 1M (K gènes)

$$\mathcal{X} = [0, 1M]^K = [0, 1M] \times [0, 1M] \times \dots \times [0, 1M]$$

# 1.1 Le jeu de données

## Notations

$X : \Omega \mapsto \mathcal{X}$  : variable aléatoire (scalaire ou vectorielle)

$x \in \mathcal{X}$  : une observation de la variable aléatoire

### Exemples:

1. On lance un dé,  $\mathcal{X} = \{1, 2, \dots, 6\}$

2. On mesure la longueur et la largeur de fleurs  $\mathcal{X} = \mathbb{R}^+ \times \mathbb{R}^+$

3. On mesure le niveau d'expression de gènes relatif à 1M (K gènes)

$$\mathcal{X} = [0, 1M]^K = [0, 1M] \times [0, 1M] \times \dots \times [0, 1M]$$

$n \in \mathbb{N}^*$ : nombre d'observations (=individus, échantillons)

$p \in \mathbb{N}^*$ : nombre de variables (=covariables, prédicteurs, features)

$x_{1:n} = (x_1, \dots, x_n)$ : ensemble d'observations (=échantillon, =dataset)

# 1.1 Le jeu de données

---

Exemple:

1. On réalise un essai clinique de réponse à un traitement sur 100 patients

$X = (\text{Age, Poids, Nb Tx antérieurs,}$   
 $\text{Exp gene 1, Exp gene 2, Mutation gene 1,}$   
 $\text{Temps avant rechute, Meilleure réponse})$

# 1.1 Le jeu de données

---

Exemple:

1. On réalise un essai clinique de réponse à un traitement sur 100 patients

$X = (\text{Age, Poids, Nb Tx antérieurs,}$   
 $\text{Exp gene 1, Exp gene 2, Mutation gene 1,}$   
 $\text{Temps avant rechute, Meilleure réponse})$

$x_1 = (45.2, 78.2, 3, 1032, 258, 1, 85, \text{PR})$

$x_2 = (81, 63, 6, 589, 903, 0, 390, \text{SD})$

...

# 1.1 Le jeu de données

Exemple:

1. On réalise un essai clinique de réponse à un traitement sur 100 patients

$X = (\text{Age, Poids, Nb Tx antérieurs,}$   
 $\text{Exp gene 1, Exp gene 2, Mutation gene 1,}$   
 $\text{Temps avant rechute, Meilleure réponse})$

$x_1 = (45.2, 78.2, 3, 1032, 258, 1, 85, \text{PR})$   
 $x_2 = (81, 63, 6, 589, 903, 0, 390, \text{SD})$   
...

$\left. \begin{array}{l} \\ \\ \end{array} \right\} (x_1, x_2, \dots, x_{100})$   
 $= \text{jeu de données}$

Objectifs

1. **Proposer un modèle** mathématique pour modéliser une variable en fonction d'autres. Exemple Meilleure réponse vs (Exp gene 1, Exp gene 2, Age)  
Modèle = Régression logistique multinomiale

# 1.1 Le jeu de données

Exemple:

1. On réalise un essai clinique de réponse à un traitement sur 100 patients

$X = (\text{Age, Poids, Nb Tx antérieurs,}$   
 $\text{Exp gene 1, Exp gene 2, Mutation gene 1,}$   
 $\text{Temps avant rechute, Meilleure réponse})$

$x_1 = (45.2, 78.2, 3, 1032, 258, 1, 85, \text{PR})$   
 $x_2 = (81, 63, 6, 589, 903, 0, 390, \text{SD})$   
...

$\left. \begin{array}{l} \\ \\ \end{array} \right\} (x_1, x_2, \dots, x_{100})$   
 $= \text{jeu de données}$

Objectifs

1. **Proposer un modèle** mathématique pour modéliser une variable en fonction d'autres. Exemple Meilleure réponse vs (Exp gene 1, Exp gene 2, Age)  
Modèle = Régression logistique multinomiale
2. **Estimer les paramètres** du modèle à partir de  $(x_1, x_2, \dots, x_{100})$

# 1.2 L'estimation

---

## Exemple:

1. On veut prédire le volume de la tumeur à partir de l'expression de 5000 gènes

->  $(G^1, \dots, G^{5000})$  profil d'expression et  $V = \text{volume}$

$x_{1:n} = (g^1, \dots, g^{5000}, v)_{1:n}$  observations de  $X = (G^1, \dots, G^{5000}, V)$

# 1.2 L'estimation

Exemple:

1. On veut predire le volume de la tumeur à partir de l'expression de 5000 gènes

->  $(G^1, \dots, G^{5000})$  profil d'expression et  $V = \text{volume}$

$x_{1:n} = (g^1, \dots, g^{5000}, v)_{1:n}$  observations de  $X = (G^1, \dots, G^{5000}, V)$

-> on prend comme modèle un **modèle de regression linéaire**

$$V = f_\theta(G^1, G^{5000})$$

$$= \theta_1 G^1 + \dots + \theta_{5000} G^{5000}$$

Schématique, pas  
mathématiquement exact

# 1.2 L'estimation

Exemple:

1. On veut predire le volume de la tumeur à partir de l'expression de 5000 gènes

->  $(G^1, \dots, G^{5000})$  profil d'expression et  $V = \text{volume}$

$x_{1:n} = (g^1, \dots, g^{5000}, v)_{1:n}$  observations de  $X = (G^1, \dots, G^{5000}, V)$

-> on prend comme modèle un **modèle de regression linéaire**

$$V = f_\theta(G^1, G^{5000})$$

$$= \theta_1 G^1 + \dots + \theta_{5000} G^{5000}$$

Schématique, pas  
mathématiquement exact

Fonction objectif (=coût) Ca veut dire quoi "coller au mieux aux données" ?

Mesure de distance entre l'observation  $v_i$  et la prediction  $\hat{v}_i = f_\theta(g_i^1, \dots, g_i^{5000})$  ?

# 1.2 L'estimation

Exemple:

1. On veut predire le volume de la tumeur à partir de l'expression de 5000 gènes

->  $(G^1, \dots, G^{5000})$  profil d'expression et  $V = \text{volume}$

$x_{1:n} = (g^1, \dots, g^{5000}, v)_{1:n}$  observations de  $X = (G^1, \dots, G^{5000}, V)$

-> on prend comme modèle un **modèle de regression linéaire**

$$V = f_\theta(G^1, G^{5000})$$

$$= \theta_1 G^1 + \dots + \theta_{5000} G^{5000}$$

Schématique, pas  
mathématiquement exact

Fonction objectif (=coût) Ca veut dire quoi "coller au mieux aux données" ?

Mesure de distance entre l'observation  $v_i$  et la prediction  $\hat{v}_i = f_\theta(g_i^1, \dots, g_i^{5000})$  ?

$$d(v_i, \hat{v}_i) = (v_i - \hat{v}_i) ?$$

# 1.2 L'estimation

Exemple:

1. On veut predire le volume de la tumeur à partir de l'expression de 5000 gènes

->  $(G^1, \dots, G^{5000})$  profil d'expression et  $V = \text{volume}$

$x_{1:n} = (g^1, \dots, g^{5000}, v)_{1:n}$  observations de  $X = (G^1, \dots, G^{5000}, V)$

-> on prend comme modèle un **modèle de regression linéaire**

$$V = f_\theta(G^1, G^{5000})$$

$$= \theta_1 G^1 + \dots + \theta_{5000} G^{5000}$$

Schématique, pas  
mathématiquement exact

Fonction objectif (=coût) Ca veut dire quoi "coller au mieux aux données" ?

Mesure de distance entre l'observation  $v_i$  et la prediction  $\hat{v}_i = f_\theta(g_i^1, \dots, g_i^{5000})$  ?

$$d(v_i, \hat{v}_i) = (v_i - \hat{v}_i) ? \quad (v_i - \hat{v}_i)^2 ?$$

# 1.2 L'estimation

Exemple:

1. On veut predire le volume de la tumeur à partir de l'expression de 5000 gènes

->  $(G^1, \dots, G^{5000})$  profil d'expression et  $V = \text{volume}$

$x_{1:n} = (g^1, \dots, g^{5000}, v)_{1:n}$  observations de  $X = (G^1, \dots, G^{5000}, V)$

-> on prend comme modèle un **modèle de regression linéaire**

$$V = f_\theta(G^1, G^{5000})$$

$$= \theta_1 G^1 + \dots + \theta_{5000} G^{5000}$$

Schématique, pas  
mathématiquement exact

Fonction objectif (=coût) Ca veut dire quoi "coller au mieux aux données" ?

Mesure de distance entre l'observation  $v_i$  et la prediction  $\hat{v}_i = f_\theta(g_i^1, \dots, g_i^{5000})$  ?

$$d(v_i, \hat{v}_i) = (v_i - \hat{v}_i) ? \quad (v_i - \hat{v}_i)^2 ? \quad (v_i - \hat{v}_i)^4 ?$$

# 1.2 L'estimation

Exemple:

1. On veut predire le volume de la tumeur à partir de l'expression de 5000 gènes

->  $(G^1, \dots, G^{5000})$  profil d'expression et  $V = \text{volume}$

$x_{1:n} = (g^1, \dots, g^{5000}, v)_{1:n}$  observations de  $X = (G^1, \dots, G^{5000}, V)$

-> on prend comme modèle un **modèle de regression linéaire**

$$V = f_\theta(G^1, G^{5000})$$

$$= \theta_1 G^1 + \dots + \theta_{5000} G^{5000}$$

Schématique, pas  
mathématiquement exact

Fonction objectif (=coût) Ca veut dire quoi "coller au mieux aux données" ?

Mesure de distance entre l'observation  $v_i$  et la prediction  $\hat{v}_i = f_\theta(g_i^1, \dots, g_i^{5000})$  ?

$$d(v_i, \hat{v}_i) = (v_i - \hat{v}_i) ? \quad (v_i - \hat{v}_i)^2 ? \quad (v_i - \hat{v}_i)^4 ? \quad e^{|v_i - \hat{v}_i|} - 1 ?$$

# 1.2 L'estimation

Exemple:

1. On veut predire le volume de la tumeur à partir de l'expression de 5000 gènes

->  $(G^1, \dots, G^{5000})$  profil d'expression et  $V = \text{volume}$

$x_{1:n} = (g^1, \dots, g^{5000}, v)_{1:n}$  observations de  $X = (G^1, \dots, G^{5000}, V)$

-> on prend comme modèle un **modèle de regression linéaire**

$$V = f_\theta(G^1, G^{5000})$$

$$= \theta_1 G^1 + \dots + \theta_{5000} G^{5000}$$

Schématique, pas  
mathématiquement exact

Fonction objectif (=coût) Ca veut dire quoi "coller au mieux aux données" ?

Mesure de distance entre l'observation  $v_i$  et la prediction  $\hat{v}_i = f_\theta(g_i^1, \dots, g_i^{5000})$  ?

$d(v_i, \hat{v}_i) = (v_i - \hat{v}_i)$  ?     $(v_i - \hat{v}_i)^2$  ?     $\sum_n (v_i - \hat{v}_i)^4$  ?     $e^{|v_i - \hat{v}_i|} - 1$  ?

Estimation     $\hat{\theta} = \operatorname{argmin}_{\theta \in \Theta} \sum_{i=1}^n d(v_i, \hat{v}_i)$

# 1.2 L'estimation

---

Quels modèles ? Le modèle peut avoir une interpretation probabiliste ou non.

-> si interpretation probabiliste

**fonction objectif = - maximum de vraisemblance**

Exemples: régression linéaire, régression logistique, classification bayésienne naïve

# 1.2 L'estimation

---

Quels modèles ? Le modèle peut avoir une interpretation probabiliste ou non.

-> si interpretation probabiliste

fonction objectif = **- maximum de vraisemblance**

Exemples: régression linéaire, régression logistique, classification bayésienne naïve

-> si pas d'interpretation probabiliste

fonction objectif = **a la main**

Exemples: machine à vecteur de support (SVM), forêt aléatoire, réseaux de neurones

# 1.2 L'estimation

---

Quels modèles ? Le modèle peut avoir une interpretation probabiliste ou non.

-> si interpretation probabiliste (=modèle statistique)

fonction objectif = **- maximum de vraisemblance**

Exemples: régression linéaire, régression logistique, classification bayésienne naïve

-> si pas d'interpretation probabiliste (=modèle statistique?)

fonction objectif = **a la main**

Exemples: machine à vecteur de support (SVM), forêt aléatoire, réseaux de neurones

Les réseaux de neurones sont-ils des modèles statistiques ?

<https://ai.stackexchange.com/questions/10289/are-neural-networks-statistical-models/18580#18580>

Cf aussi sur la définition des modèles statistiques <https://www.stat.uchicago.edu/~pmcc/pubs/AOS023.pdf>

# 1.2 L'estimation

---

-> si interpretation probabiliste (=modèle statistique)

Modèle statistique = ensemble de lois (=mesures) de probabilité  $\mathbb{P}$  sur l'espace des observations  $\mathcal{X}$ . Si proba paramétriques, alors on parle de modèle paramétrique. Enfin si les proba ont une densité  $p$  alors le modèle s'écrit

$$\mathcal{M}_\Theta = \{p_\theta | \theta \in \Theta\}$$

# 1.2 L'estimation

-> si interprétation probabiliste (=modèle statistique)

Modèle statistique = ensemble de lois (=mesures) de probabilité  $\mathbb{P}$  sur l'espace des observations  $\mathcal{X}$ . Si proba paramétriques, alors on parle de modèle paramétrique. Enfin si les proba ont une densité  $p$  alors le modèle s'écrit

$$\mathcal{M}_\Theta = \{p_\theta | \theta \in \Theta\}$$

Quand on modélise, on fait l'hypothèse qu'il existe  $\theta^*$  tel que  $\mathbb{P}_X = \mathbb{P}_{\theta^*}$

Modéliser = calculer  $\hat{\theta}$  en “espérant” que  $\hat{\theta} \approx \theta^*$

## 1.2 L'estimation

Modèle statistique  $\mathcal{M}_\Theta = \{p_\theta | \theta \in \Theta\}$  Il existe  $\theta^*$  tel que  $\mathbb{P}_X = \mathbb{P}_{\theta^*}$

Pour le modèle candidat  $\mathbb{P}_\theta$

$p_\theta(x_i)$  = vraisemblance échantillon i

$\prod_{i=1}^n p_\theta(x_i)$  = vraisemblance jeu de données

# 1.2 L'estimation

Modèle statistique  $\mathcal{M}_\Theta = \{p_\theta | \theta \in \Theta\}$  Il existe  $\theta^*$  tel que  $\mathbb{P}_X = \mathbb{P}_{\theta^*}$

Pour le modèle candidat  $\mathbb{P}_\theta$

$p_\theta(x_i)$  = vraisemblance échantillon i

$\prod_{i=1}^n p_\theta(x_i)$  = vraisemblance jeu de données

$$L(\theta) = - \prod_{i=1}^n p_\theta(x_i)$$

Fonction de coût à minimiser



# 1.2 L'estimation

Modèle statistique  $\mathcal{M}_\Theta = \{p_\theta | \theta \in \Theta\}$  Il existe  $\theta^*$  tel que  $\mathbb{P}_X = \mathbb{P}_{\theta^*}$

Pour le modèle candidat  $\mathbb{P}_\theta$

$p_\theta(x_i)$  = vraisemblance échantillon i

$\prod_{i=1}^n p_\theta(x_i)$  = vraisemblance jeu de données

$L(\theta) = - \prod_{i=1}^n p_\theta(x_i)$  équivalent à

$$\ell(\theta) = - \sum_{i=1}^n \log p_\theta(x_i)$$

Fonction de coût à minimiser = - log de la vraisemblance

# Sommaire

---

## 1. Données et modélisation

1. Le jeu de données
2. L'estimation (model fitting)

## 2. Modèles de classification linéaires

1. Régression logistique binaire
2. Régression logistique multinomiale
3. Régression linéaire
4. Régression linéaire lasso

## 3. Modèles de classification non linéaires

1. Régression logistique
2. Arbre de décision CART
3. Forêts aléatoires

# 2.1 Régression logistique binaire

Modèle statistique de regression Prédire  $Y$  à partir de  $X$

Jeu de données  $x_{1:n}, y_{1:n}$

Un modèle de regression est une famille de lois de probabilités *conditionnelles*

$$\mathcal{M}_\Theta = \{\mathbb{P}_{Y|X=x}^\theta | \theta \in \Theta, x \in \mathcal{X}\}$$

# 2.1 Régression logistique binaire

Modèle statistique de regression Prédire  $Y$  à partir de  $X$

Jeu de données  $x_{1:n}, y_{1:n}$

Un modèle de regression est une famille de lois de probabilités *conditionnelles*

$$\mathcal{M}_\Theta = \{\mathbb{P}_{Y|X=x}^\theta | \theta \in \Theta, x \in \mathcal{X}\}$$

Regression logistique  $Y \in \{0, 1\}, \quad X \in \mathbb{R}^p$

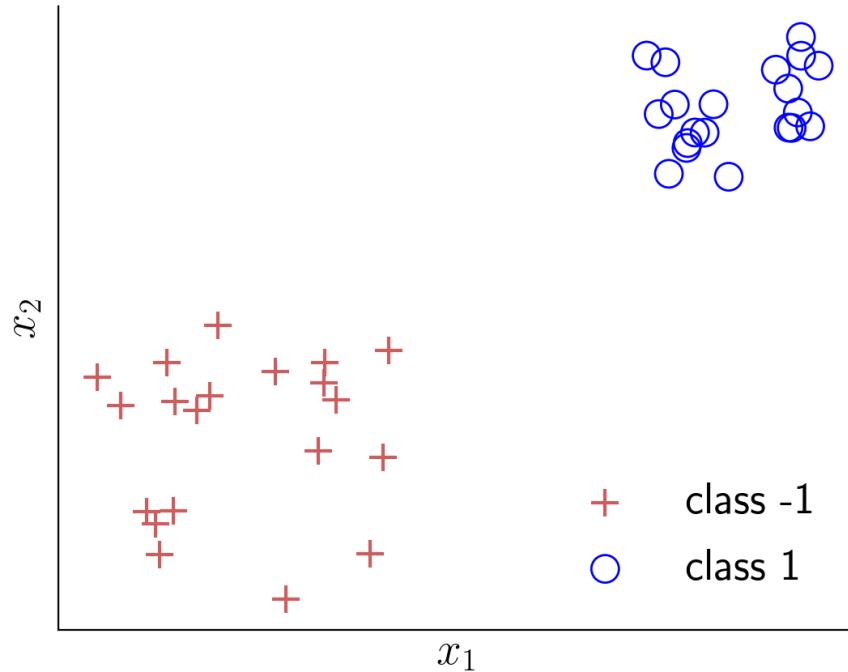
$$\mathbb{P}_{Y|X=x}^\theta = \text{Binomial}(\sigma(x^\top \theta)) \quad \sigma(u) = \frac{1}{1 + e^{-u}}$$

# 2.1 Régression logistique binaire

Regression logistique  $Y \in \{0, 1\}$ ,

Exemple:  $p=2$ ,  $X = (X^1, X^2)$

$$X \in \mathbb{R}^p \quad \mathbb{P}_{Y|X=x}^\theta = \text{Binomial}(\sigma(x^\top \theta))$$



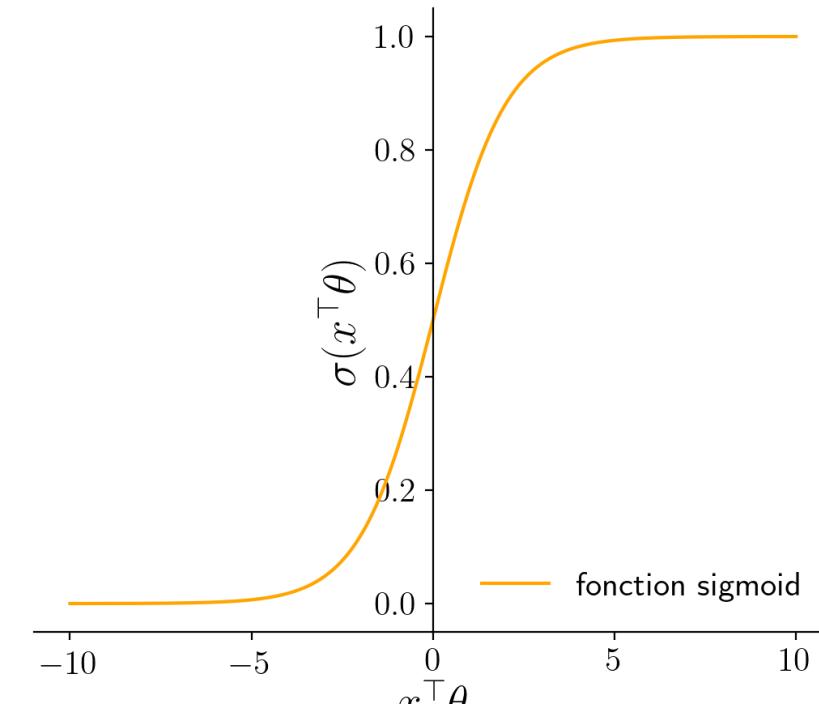
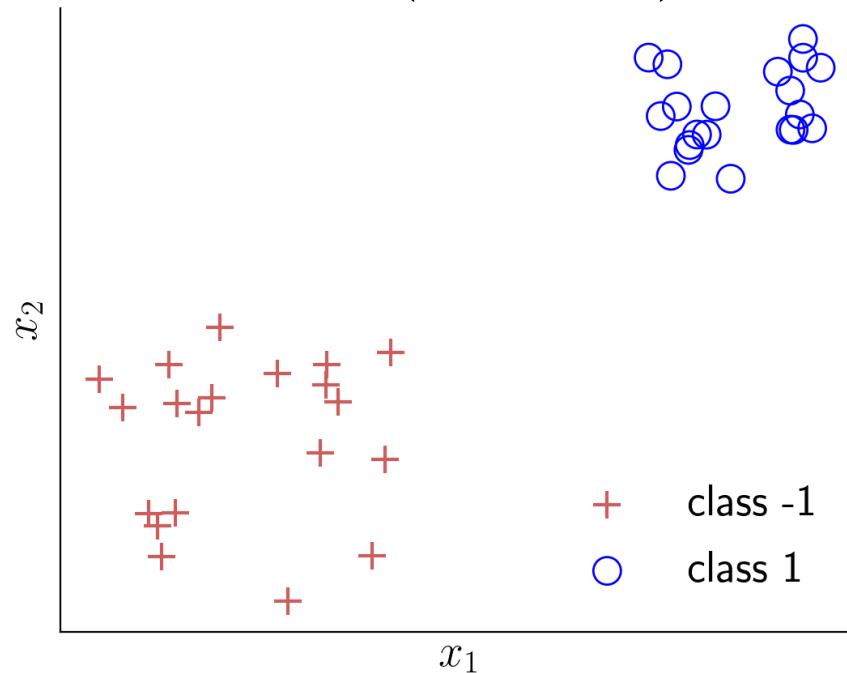
# 2.1 Régression logistique binaire

Regression logistique  $Y \in \{0, 1\}$ ,

$X \in \mathbb{R}^p$

$\mathbb{P}_Y^{\theta}|_{X=x} = \text{Binomial}(\sigma(x^\top \theta))$

Exemple:  $p=2$ ,  $X = (X^1, X^2)$



Pour  $X = x$ , prédition =  $\begin{cases} 1 & \text{si } \sigma(x^\top \theta) > 0.5, \text{i.e } x^\top \theta > 0 \\ 0 & \text{si } x^\top \theta < 0 \end{cases}$

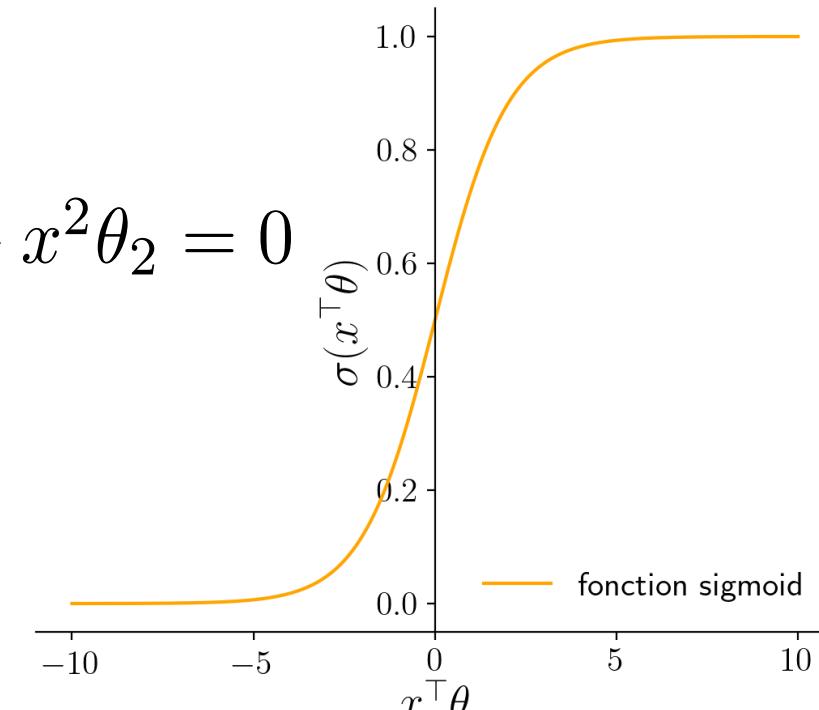
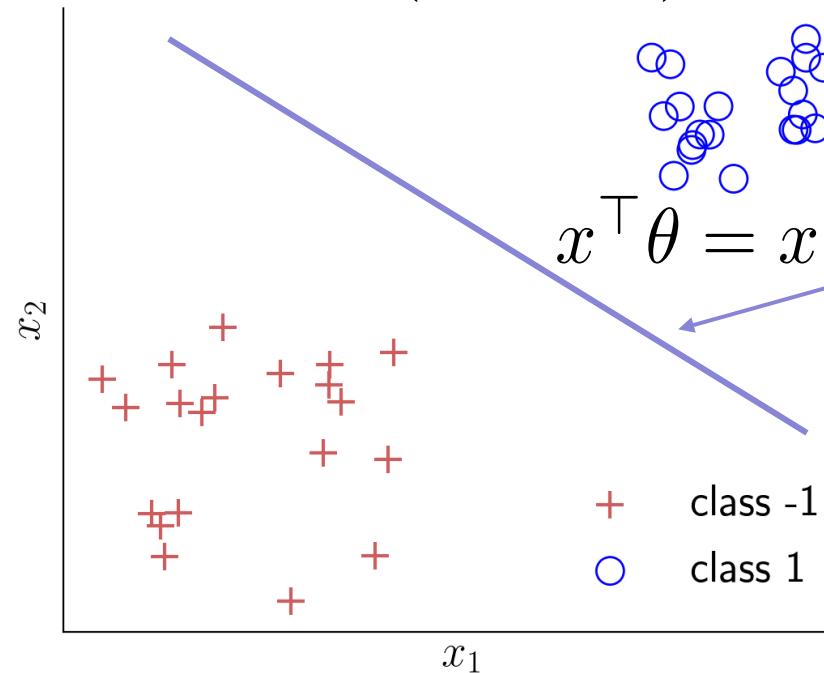
# 2.1 Régression logistique binaire

Regression logistique  $Y \in \{0, 1\}$ ,

$X \in \mathbb{R}^p$

$\mathbb{P}_Y^{\theta}|_{X=x} = \text{Binomial}(\sigma(x^{\top} \theta))$

Exemple:  $p=2$ ,  $X = (X^1, X^2)$



Pour  $X = x$ , prédition =  $\begin{cases} 1 & \text{si } \sigma(x^{\top} \theta) > 0.5, \text{i.e } x^{\top} \theta > 0 \\ 0 & \text{si } x^{\top} \theta < 0 \end{cases}$

# 2.1 Régression logistique binaire

Regression logistique  $Y \in \{0, 1\}, X \in \mathbb{R}^p$

$$\mathbb{P}_{Y|X=x}^\theta = \text{Binomial}(\sigma(x^\top \theta))$$

$$\sigma(u) = \frac{1}{1 + e^{-u}}$$

Comment s'écrit la vraisemblance de  $y_i|x_i$  ?

$$p_\theta(\cdot|x_i) = \begin{cases} 1 & \text{avec probabilité } \sigma(x_i^\top \theta) \\ 0 & \text{avec probabilité } 1 - \sigma(x_i^\top \theta) \end{cases}$$

# 2.1 Régression logistique binaire

Regression logistique  $Y \in \{0, 1\}, X \in \mathbb{R}^p$

$$\mathbb{P}_{Y|X=x}^\theta = \text{Binomial}(\sigma(x^\top \theta))$$

$$\sigma(u) = \frac{1}{1 + e^{-u}}$$

Comment s'écrit la vraisemblance de  $y_i|x_i$  ?

$$p_\theta(\cdot|x_i) = \begin{cases} 1 & \text{avec probabilité } \sigma(x_i^\top \theta) \\ 0 & \text{avec probabilité } 1 - \sigma(x_i^\top \theta) \end{cases}$$

$$p_\theta(y_i|x_i) = \sigma(\theta^\top x_i)^{y_i} (1 - \sigma(\theta^\top x_i))^{1-y_i}$$

# 2.1 Régression logistique binaire

Regression logistique  $Y \in \{0, 1\}, X \in \mathbb{R}^p$

$$\mathbb{P}_{Y|X=x}^\theta = \text{Binomial}(\sigma(x^\top \theta))$$

$$\sigma(u) = \frac{1}{1 + e^{-u}}$$

Comment s'écrit la vraisemblance de  $y_i|x_i$  ?

$$p_\theta(\cdot|x_i) = \begin{cases} 1 & \text{avec probabilité } \sigma(x_i^\top \theta) \\ 0 & \text{avec probabilité } 1 - \sigma(x_i^\top \theta) \end{cases}$$

$$p_\theta(y_i|x_i) = \sigma(\theta^\top x_i)^{y_i} (1 - \sigma(\theta^\top x_i))^{1-y_i}$$

**Fonction de coût**

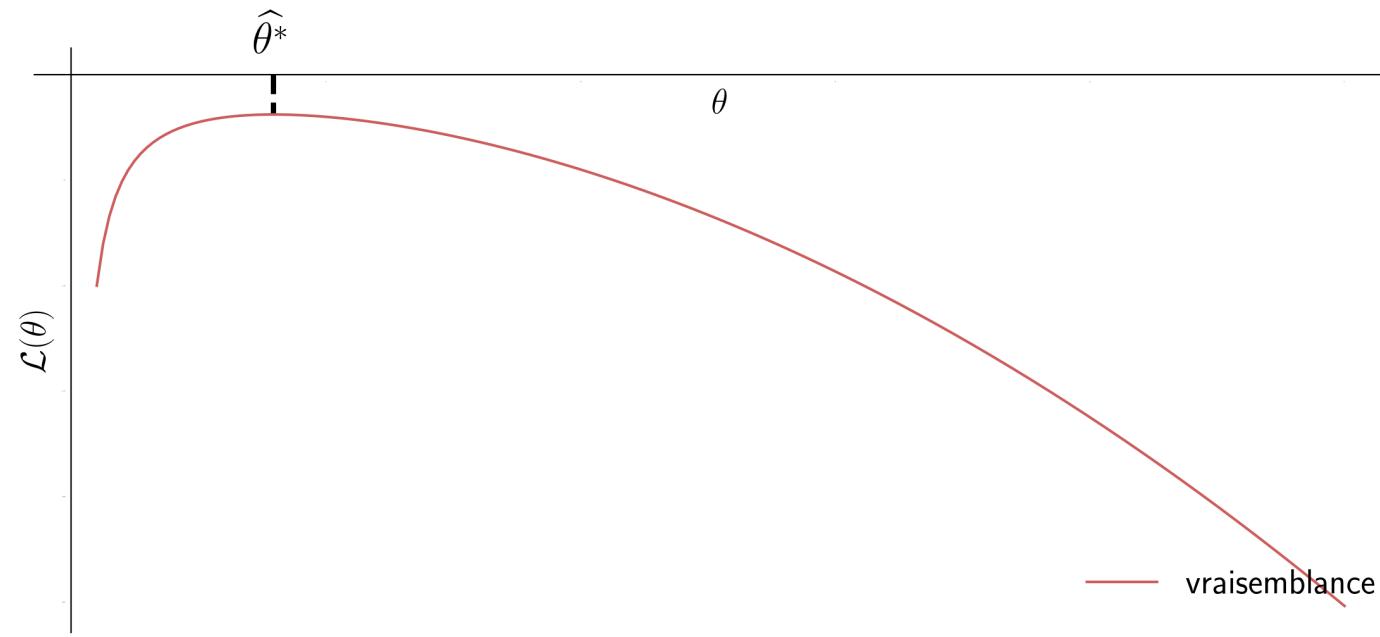
$$\ell(\theta; x_{1:n}, y_{1:n}) = - \sum_{i=1}^n y_i \log \sigma(\theta^\top x_i) + (1 - y_i) \log(1 - \sigma(\theta^\top x_i))$$

# 2.1 Régression logistique binaire

Regression logistique

**Fonction de coût**

$$\ell(\theta; x_{1:n}, y_{1:n}) = - \sum_{i=1}^n y_i \log \sigma(\theta^\top x_i) + (1 - y_i) \log(1 - \sigma(\theta^\top x_i))$$



Concave !

# 2.1 Régression logistique binaire

Regression logistique

**Fonction de coût**

$$\ell(\theta; x_{1:n}, y_{1:n}) = - \sum_{i=1}^n y_i \log \sigma(\theta^\top x_i) + (1 - y_i) \log(1 - \sigma(\theta^\top x_i))$$

Minimisation par descente de gradient

$$\ell(\theta^{t+1}) = \ell(\theta^t) + (\theta^{t+1} - \theta^t) \nabla \ell(\theta^t) + o(\theta^{t+1} - \theta^t)$$

(dvp de Taylor ordre 1)

# 2.1 Régression logistique binaire

Regression logistique

**Fonction de coût**

$$\ell(\theta; x_{1:n}, y_{1:n}) = - \sum_{i=1}^n y_i \log \sigma(\theta^\top x_i) + (1 - y_i) \log(1 - \sigma(\theta^\top x_i))$$

Minimisation par descente de gradient

$$\ell(\theta^{t+1}) = \ell(\theta^t) + (\theta^{t+1} - \theta^t) \nabla \ell(\theta^t) + o(\theta^{t+1} - \theta^t)$$

(dvp de Taylor ordre 1)

Idée. Choisir  $\theta^{t+1}$  tel que  $(\theta^{t+1} - \theta^t) = -\nabla \ell(\theta^t)$

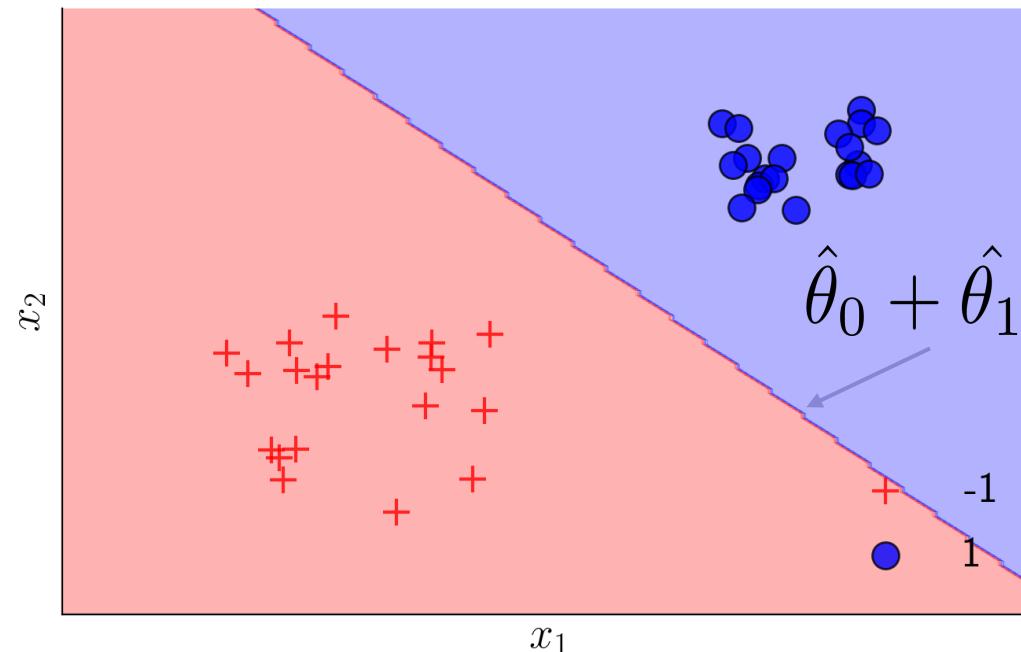
# 2.1 Régression logistique binaire

Regression logistique

**Fonction de coût**

$$\ell(\theta; x_{1:n}, y_{1:n}) = - \sum_{i=1}^n y_i \log \sigma(\theta^\top x_i) + (1 - y_i) \log(1 - \sigma(\theta^\top x_i))$$

Minimisation par descente de gradient



$$\hat{\theta}_0 + \hat{\theta}_1 x^1 + \hat{\theta}_2 x^2 = 0.5$$

## 2.2 Régression logistique multinomiale

Regression logistique multinomiale  $Y \in \{1, 2, \dots, K\}$ ,  $X \in \mathbb{R}^p$

$$\mathbb{P}_{Y|X=x}^{\Theta} = \text{Multinomial}(\sigma(\Theta^T x))$$

$$\sigma(z) = \frac{1}{\sum_{k=1}^K e^{z_k}} \begin{bmatrix} e^{z_1} \\ \vdots \\ e^{z_K} \end{bmatrix}$$

$$\Theta = \begin{bmatrix} \theta_{1,1} & \cdots & \theta_{1,p} \\ \vdots & \ddots & \vdots \\ \theta_{K,1} & \cdots & \theta_{K,p} \end{bmatrix}$$

## 2.2 Régression logistique multinomiale

Regression logistique multinomiale  $Y \in \{1, 2, \dots, K\}$ ,  $X \in \mathbb{R}^p$

$$\mathbb{P}_{Y|X=x}^\Theta = \text{Multinomial}(\sigma(\Theta^\top x))$$

$$\sigma(z) = \frac{1}{\sum_{k=1}^K e^{z_k}} \begin{bmatrix} e^{z_1} \\ \vdots \\ e^{z_K} \end{bmatrix}$$

$$\Theta = \begin{bmatrix} \theta_{1,1} & \cdots & \theta_{1,p} \\ \vdots & \ddots & \vdots \\ \theta_{K,1} & \cdots & \theta_{K,p} \end{bmatrix}$$

$$\begin{bmatrix} \mathbb{P}_{Y|X=x}^\Theta(Y=1) \\ \vdots \\ \mathbb{P}_{Y|X=x}^\Theta(Y=K) \end{bmatrix} = \frac{1}{\sum_{k=1}^K e^{(\Theta^\top x)_k}} \begin{bmatrix} e^{(\Theta^\top x)_1} \\ \vdots \\ e^{(\Theta^\top x)_K} \end{bmatrix}$$

## 2.2 Régression logistique multinomiale

Regression logistique multinomiale  $Y \in \{1, 2, \dots, K\}$ ,  $X \in \mathbb{R}^p$

Comment s'écrit la vraisemblance de  $y_i | x_i$  ?

$p_{\Theta}(\cdot | x_i) = k$  avec probabilité  $\sigma(\Theta^\top x)_k$

$$p_{\Theta}(y_i | x_i) = \prod_{k=1}^K \sigma(\Theta^\top x_i)_k^{1_{y_i=k}}$$

## 2.2 Régression logistique multinomiale

Régression logistique multinomiale  $Y \in \{1, 2, \dots, K\}$ ,  $X \in \mathbb{R}^p$

Comment s'écrit la vraisemblance de  $y_i | x_i$  ?

$$p_{\Theta}(\cdot | x_i) = k \text{ avec probabilité } \sigma(\Theta^\top x)_k$$

$$p_{\Theta}(y_i | x_i) = \prod_{k=1}^K \sigma(\Theta^\top x_i)_k^{1_{y_i=k}}$$

**Fonction de coût**

$$\ell(\Theta; x_{1:n}, y_{1:n}) = - \sum_{i=1}^n \sum_{k=1}^K \mathbf{1}_{y_i=k} \log \sigma(\Theta^\top x_i)_k$$

## 2.2 Régression logistique multinomiale

Estimation par descente de gradient  $(\Theta^{t+1} - \Theta^t) = -\nabla \ell(\Theta^t)$

## 2.2 Régression logistique multinomiale

Estimation par descente de gradient  $(\Theta^{t+1} - \Theta^t) = -\nabla \ell(\Theta^t)$

NOTE 1: modèle surparamétré, car

$$\text{si } \Theta \leftarrow \Theta - \psi = \begin{bmatrix} \Theta_{1,:} - \psi \\ \vdots \\ \Theta_{K,:} - \psi \end{bmatrix}$$

## 2.2 Régression logistique multinomiale

Estimation par descente de gradient  $(\Theta^{t+1} - \Theta^t) = -\nabla \ell(\Theta^t)$

NOTE 1: modèle surparamétré, car

si  $\Theta \leftarrow \Theta - \psi = \begin{bmatrix} \Theta_{1,:} - \psi \\ \vdots \\ \Theta_{K,:} - \psi \end{bmatrix}$  alors  $\begin{bmatrix} \mathbb{P}_{Y|X=x}^\Theta(Y=1) \\ \vdots \\ \mathbb{P}_{Y|X=x}^\Theta(Y=K) \end{bmatrix} = \begin{bmatrix} \mathbb{P}_{Y|X=x}^{\Theta-\psi}(Y=1) \\ \vdots \\ \mathbb{P}_{Y|X=x}^{\Theta-\psi}(Y=K) \end{bmatrix}$

## 2.2 Régression logistique multinomiale

Estimation par descente de gradient  $(\Theta^{t+1} - \Theta^t) = -\nabla \ell(\Theta^t)$

NOTE 1: modèle surparamétré, car

si  $\Theta \leftarrow \Theta - \psi = \begin{bmatrix} \Theta_{1,:} - \psi \\ \vdots \\ \Theta_{K,:} - \psi \end{bmatrix}$  alors  $\begin{bmatrix} \mathbb{P}_{Y|X=x}^\Theta(Y=1) \\ \vdots \\ \mathbb{P}_{Y|X=x}^\Theta(Y=K) \end{bmatrix} = \begin{bmatrix} \mathbb{P}_{Y|X=x}^{\Theta-\psi}(Y=1) \\ \vdots \\ \mathbb{P}_{Y|X=x}^{\Theta-\psi}(Y=K) \end{bmatrix}$

On fixe donc  $\Theta_{K,:} = 1$

## 2.2 Régression logistique multinomiale

Estimation par descente de gradient  $(\Theta^{t+1} - \Theta^t) = -\nabla \ell(\Theta^t)$

NOTE 1: modèle surparamétré, car

NOTE 2: généralisation de la régression logistique binaire

## 2.2 Régression logistique multinomiale

Estimation par descente de gradient  $(\Theta^{t+1} - \Theta^t) = -\nabla \ell(\Theta^t)$

NOTE 1: modèle surparamétré, car

NOTE 2: généralisation de la régression logistique binaire

NOTE 3: donne le même modèle de classification que K modèles de regression logistique binaire combinés en stratégie multiclasse one-vs-rest

$f_{\theta^1}^1 = \text{Reg. log. binaire } Y = 1 \text{ vs } Y \neq 1$

...

$f_{\theta^K}^K = \text{Reg. log. binaire } Y = K \text{ vs } Y \neq K$

# 2.3 Régression linéaire

Modèle statistique de regression Prédire  $Y$  à partir de  $X$

Jeu de données  $x_{1:n}, y_{1:n}$

Un modèle de regression est une famille de lois de probabilités *conditionnelles*

$$\mathcal{M}_\Theta = \{\mathbb{P}_{Y|X=x}^\theta | \theta \in \Theta, x \in \mathcal{X}\}$$

Regression linéaire  $Y \in \mathbb{R}, X \in \mathbb{R}^p$

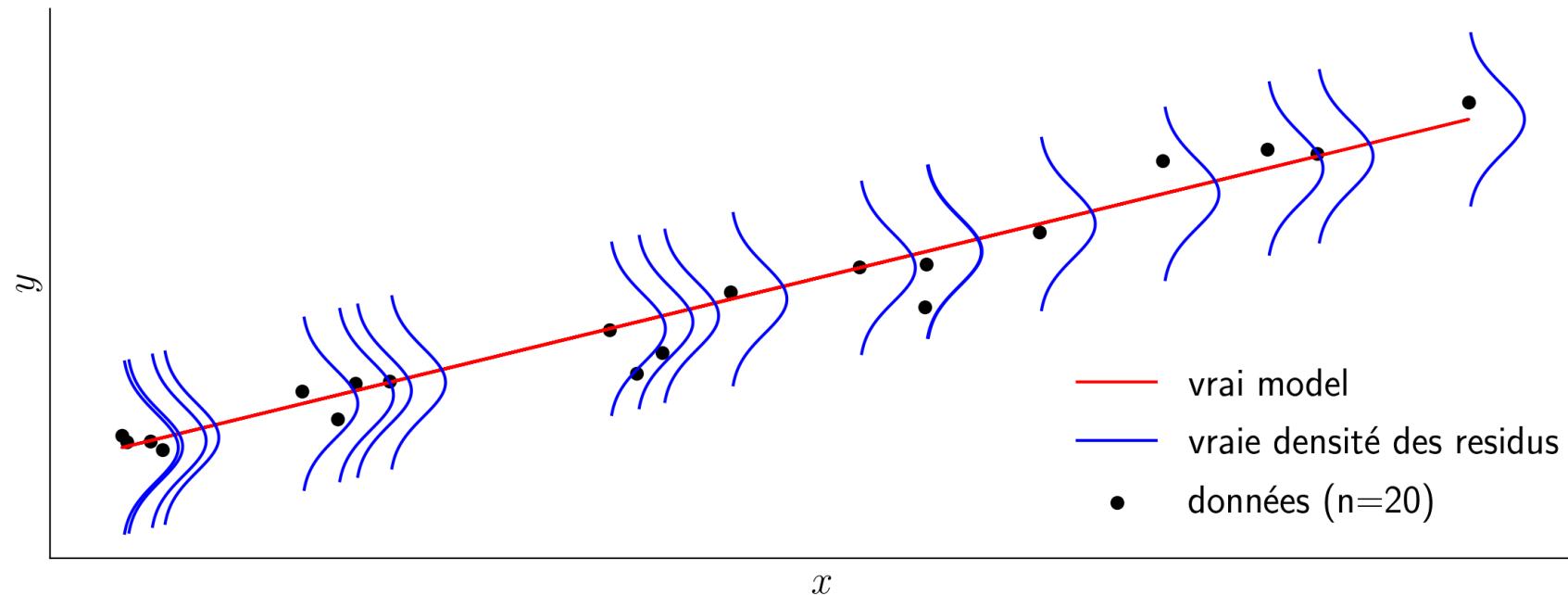
$$\mathbb{P}_{Y|X=x}^\theta = \mathcal{N}(\beta_0 + x^\top \beta, \sigma^2) \quad \theta = (\beta_0, \beta_1, \dots, \beta_p, \sigma^2)$$

# 2.3 Régression linéaire

Regression linéaire

$$Y \in \mathbb{R}, \quad X \in \mathbb{R}^p$$

$$\mathbb{P}_{Y|X=x}^\theta = \mathcal{N}(\beta_0 + x^\top \beta, \sigma^2) \quad \theta = (\beta_0, \beta_1, \dots, \beta_p, \sigma^2)$$



# 2.3 Régression linéaire

## Estimation par maximum de vraisemblance

$$\begin{aligned}\mathcal{L}(\theta; \mathbf{x}_{1:n}, \mathbf{y}_{1:n}) &= \prod_{i=1}^n p_{Y_i|X_i=\mathbf{x}_i}(y_i) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{\sigma^2 2\pi}} e^{-\frac{1}{2\sigma^2} (y_i - \beta_0 - \mathbf{x}_i^\top \boldsymbol{\beta})^2}\end{aligned}$$

# 2.3 Régression linéaire

## Estimation par maximum de vraisemblance

$$\begin{aligned}\mathcal{L}(\theta; \mathbf{x}_{1:n}, \mathbf{y}_{1:n}) &= \prod_{i=1}^n p_{Y_i|X_i=\mathbf{x}_i}(y_i) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{\sigma^2 2\pi}} e^{-\frac{1}{2\sigma^2} (y_i - \beta_0 - \mathbf{x}_i^\top \boldsymbol{\beta})^2}\end{aligned}$$

On peut montrer que  $\theta \mapsto \mathcal{L}(\theta; \mathbf{x}_{1:n}, \mathbf{y}_{1:n})$  est concave.

$\theta \mapsto \ell(\theta) = -\log \mathcal{L}(\theta; \mathbf{x}_{1:n}, \mathbf{y}_{1:n})$  est convexe.

# 2.3 Régression linéaire

## Estimation par maximum de vraisemblance

$$\begin{aligned}\mathcal{L}(\theta; \mathbf{x}_{1:n}, \mathbf{y}_{1:n}) &= \prod_{i=1}^n p_{Y_i|X_i=\mathbf{x}_i}(y_i) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{\sigma^2 2\pi}} e^{-\frac{1}{2\sigma^2} (y_i - \beta_0 - \mathbf{x}_i^\top \beta)^2}\end{aligned}$$

On peut montrer que  $\theta \mapsto \mathcal{L}(\theta; \mathbf{x}_{1:n}, \mathbf{y}_{1:n})$  est concave.

$\theta \mapsto \ell(\theta) = -\log \mathcal{L}(\theta; \mathbf{x}_{1:n}, \mathbf{y}_{1:n})$  est convexe.

$$\ell(\theta; \mathbf{x}_{1:n}, \mathbf{y}_{1:n}) = \frac{n}{2} 2\pi + \frac{n}{2} \log(\sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \mathbf{x}_i^\top \beta)^2$$

# 2.3 Régression linéaire

## Estimation par maximum de vraisemblance

$$\begin{aligned}\mathcal{L}(\theta; \mathbf{x}_{1:n}, \mathbf{y}_{1:n}) &= \prod_{i=1}^n p_{Y_i|X_i=\mathbf{x}_i}(y_i) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{\sigma^2 2\pi}} e^{-\frac{1}{2\sigma^2} (y_i - \beta_0 - \mathbf{x}_i^\top \beta)^2}\end{aligned}$$

On peut montrer que  $\theta \mapsto \mathcal{L}(\theta; \mathbf{x}_{1:n}, \mathbf{y}_{1:n})$  est concave.

$\theta \mapsto \ell(\theta) = -\log \mathcal{L}(\theta; \mathbf{x}_{1:n}, \mathbf{y}_{1:n})$  est convexe.

Moindres carrés !

$$\ell(\theta; \mathbf{x}_{1:n}, \mathbf{y}_{1:n}) = \frac{n}{2} 2\pi + \frac{n}{2} \log(\sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \mathbf{x}_i^\top \beta)^2$$

# 2.3 Régression linéaire

## Estimation par maximum de vraisemblance

Le minimum est réalisé aux dérivées nulles, i.e

$$\frac{\partial \ell}{\partial \beta_0} = 0, \frac{\partial \ell}{\partial \beta_1} = 0, \dots, \frac{\partial \ell}{\partial \beta_p} = 0, \text{ et } \frac{\partial \ell}{\partial \sigma^2} = 0$$

# 2.3 Régression linéaire

## Estimation par maximum de vraisemblance

Le minimum est réalisé aux dérivées nulles, i.e

$$\frac{\partial \ell}{\partial \beta_0} = 0, \frac{\partial \ell}{\partial \beta_1} = 0, \dots, \frac{\partial \ell}{\partial \beta_p} = 0, \text{ et } \frac{\partial \ell}{\partial \sigma^2} = 0$$

Tous calculs faits

$$\hat{\beta}(\mathbf{X}_{1:n}, \mathbf{Y}_{1:n}) = (\tilde{\mathbf{X}}_{1:n}^\top \tilde{\mathbf{X}}_{1:n})^{-1} \tilde{\mathbf{X}}_{1:n}^\top \mathbf{Y}_{1:n} \quad (1)$$

Avec

$$\tilde{\mathbf{X}}_{1:n} = \begin{bmatrix} 1 & \cdots & 1 \\ x_{1,1} & \ddots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,p} \end{bmatrix}$$

## 2.4 Régression linéaire lasso

Observation: Les estimateurs précédents sans biais mais forte variance.

Idée: Diminuer la variance au prix de biais

$$\min_{\beta_0, \beta} = \frac{1}{2n} \sum_{i=1}^n (y_i - \beta_0 - \mathbf{x}_i^\top \beta)^2$$



$$\min_{\beta_0, \beta} = \frac{1}{2n} \sum_{i=1}^n (y_i - \beta_0 - \mathbf{x}_i^\top \beta)^2$$

$$\text{s.t } \sum_{i=0}^p |\beta_i| = \|\beta\|_1 \leq t$$

## 2.4 Régression linéaire lasso

Observation: Les estimateurs précédents sans biais mais forte variance.

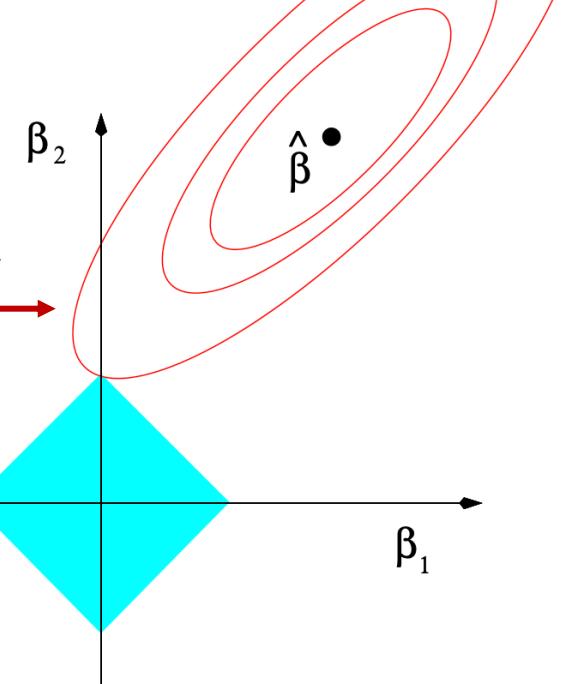
Idée: Diminuer la variance au prix de biais

$$\min_{\beta_0, \beta} = \frac{1}{2n} \sum_{i=1}^n (y_i - \beta_0 - \mathbf{x}_i^\top \beta)^2$$

$$\min_{\beta_0, \beta} = \frac{1}{2n} \sum_{i=1}^n (y_i - \beta_0 - \mathbf{x}_i^\top \beta)^2$$

$$\text{s.t } \sum_{i=0}^p |\beta_i| = \|\beta\|_1 \leq t$$

$\hat{\beta}$  Estimateur moindres carrés



## 2.4 Pénalisation lasso

Observation: Les estimateurs précédents sans biais mais forte variance.

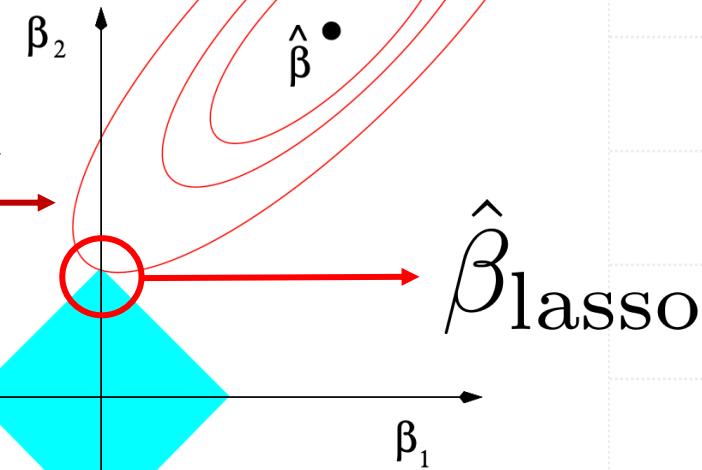
Idée: Diminuer la variance au prix de biais

$$\min_{\beta_0, \beta} = \frac{1}{2n} \sum_{i=1}^n (y_i - \beta_0 - \mathbf{x}_i^\top \beta)^2$$

$$\min_{\beta_0, \beta} = \frac{1}{2n} \sum_{i=1}^n (y_i - \beta_0 - \mathbf{x}_i^\top \beta)^2$$

$$\text{s.t } \sum_{i=0}^p |\beta_i| = \|\beta\|_1 \leq t$$

$\hat{\beta}$  Estimateur moindres carrés



$\hat{\beta}_{\text{lasso}}$

## 2.4 Régression linéaire lasso

$$\min_{\beta_0, \beta} = \frac{1}{2n} \sum_{i=1}^n (y_i - \beta_0 - \mathbf{x}_i^\top \beta)^2$$

$$\text{s.t } \sum_{i=0}^p |\beta_i| = \|\beta\|_1 \leq t$$

Dualité de Lagrange



$$\min_{\beta_0, \beta} = \frac{1}{2n} \sum_{i=1}^n (y_i - \beta_0 - \mathbf{x}_i^\top \beta)^2 + \lambda \|\beta\|_1$$

1:1

Question: Comment choisir  $t/\lambda$  ?  
Réponse: Par validation croisée!

## 2.4 Régression linéaire lasso

$$\min_{\beta_0, \beta} = \frac{1}{2n} \sum_{i=1}^n (y_i - \beta_0 - \mathbf{x}_i^\top \beta)^2$$

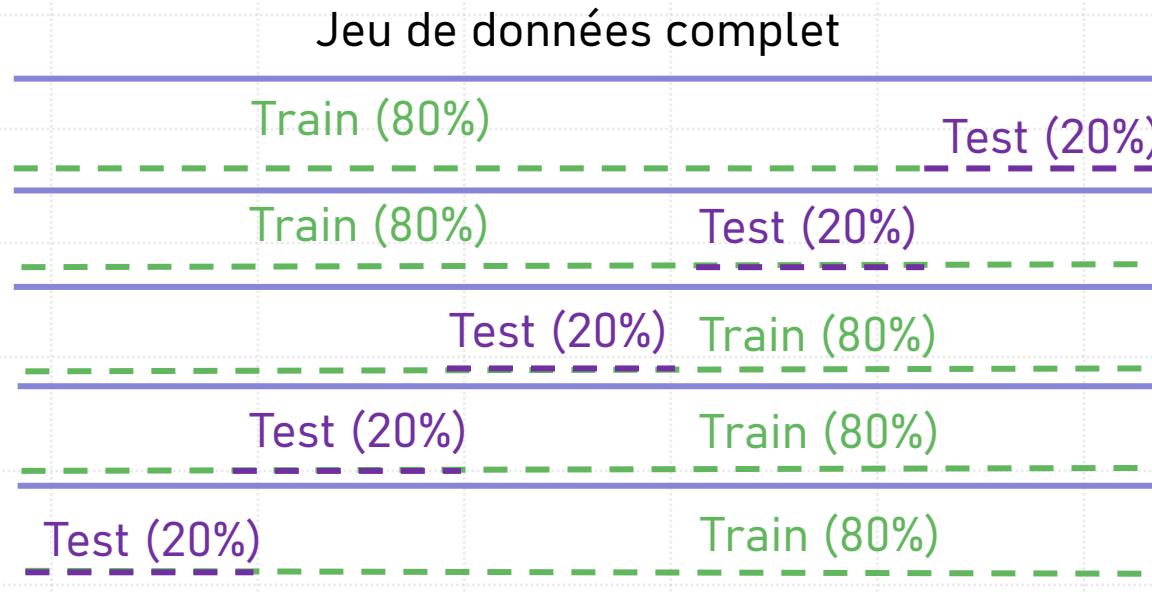
$$\text{s.t. } \sum_{i=0}^p |\beta_i| = \|\beta\|_1 \leq t$$

Dualité de Lagrange

$$\min_{\beta_0, \beta} = \frac{1}{2n} \sum_{i=1}^n (y_i - \beta_0 - \mathbf{x}_i^\top \beta)^2 + \lambda \|\beta\|_1$$

1:1

Question: Comment choisir  $t/\lambda$ ?  
Réponse: Par validation croisée!



# Sommaire

---

## 1. Données et modélisation

1. Le jeu de données
2. L'estimation (model fitting)

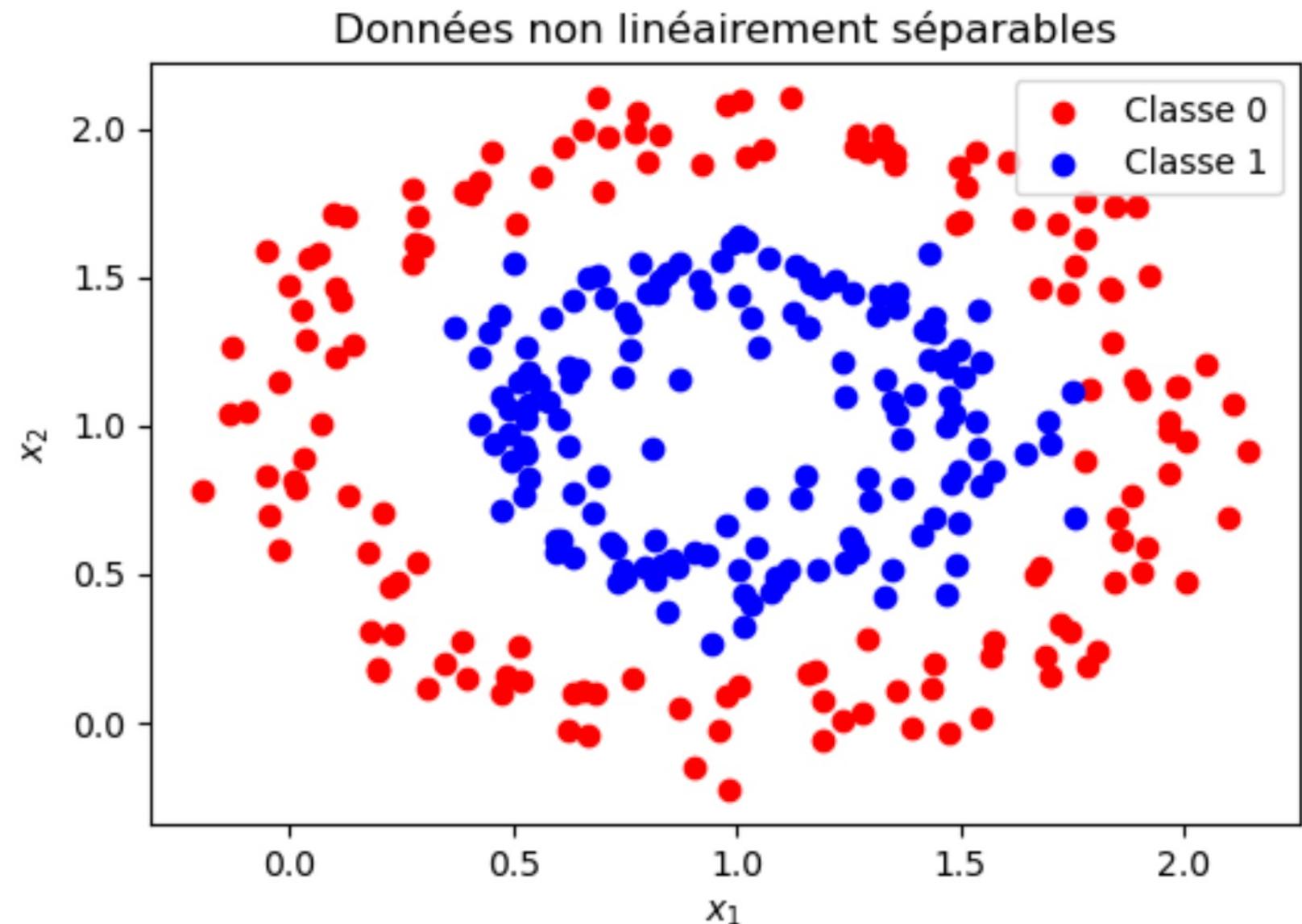
## 2. Modèles de classification linéaires

1. Régression logistique binaire
2. Régression logistique multinomiale
3. Régression linéaire
4. Régression linéaire lasso

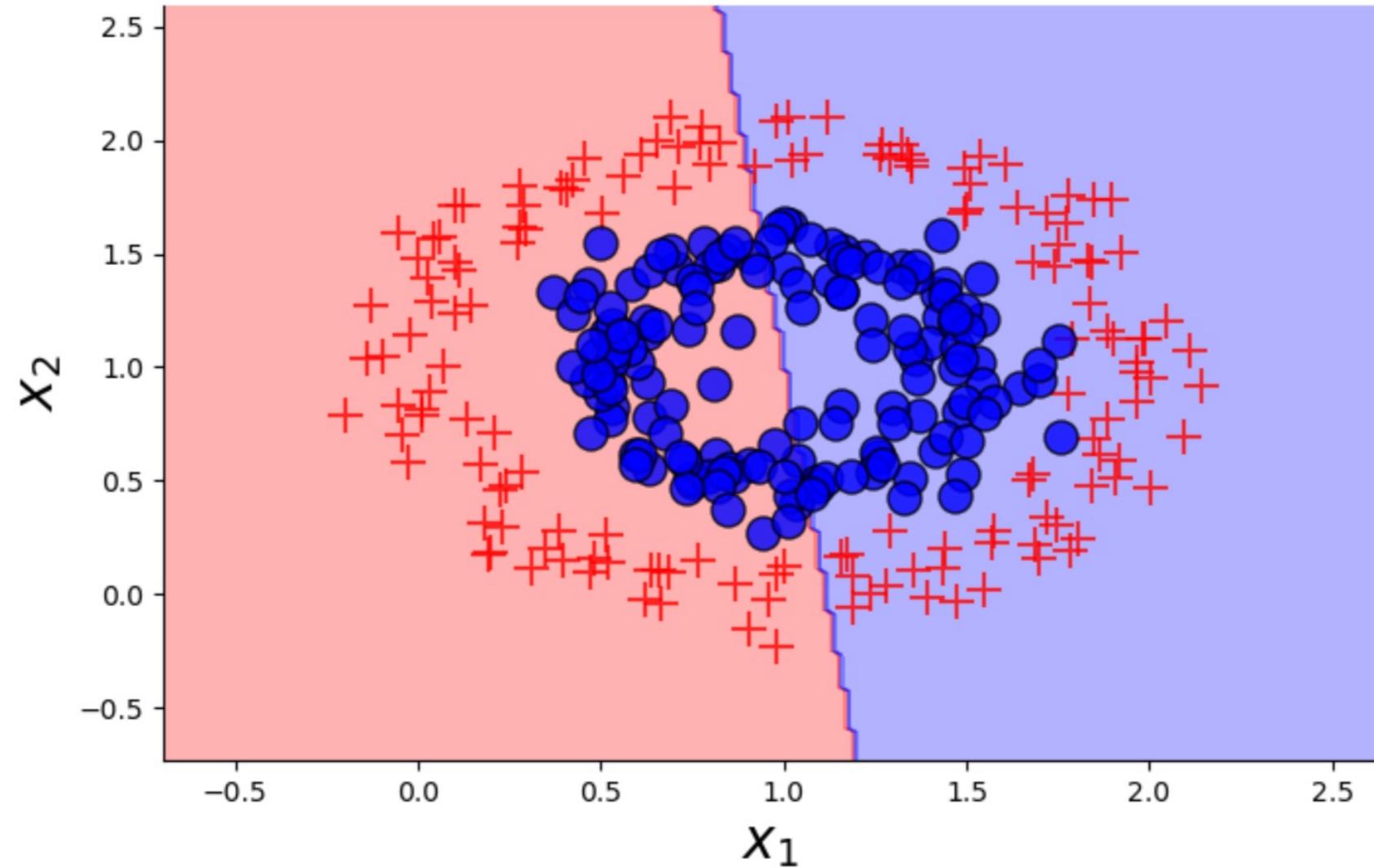
## 3. Modèles de classification non linéaires

1. Régression logistique
2. Arbre de décision CART
3. Forêts aléatoires

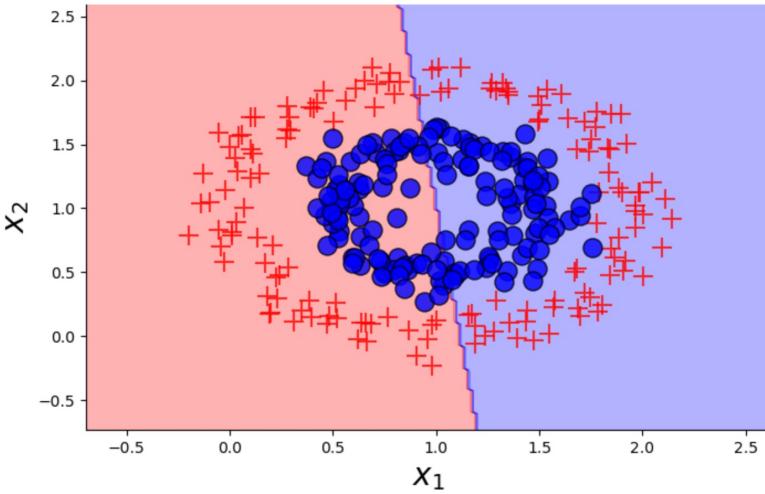
# 3.1 Régression logistique



# 3.1 Régression logistique



# 3.1 Régression logistique

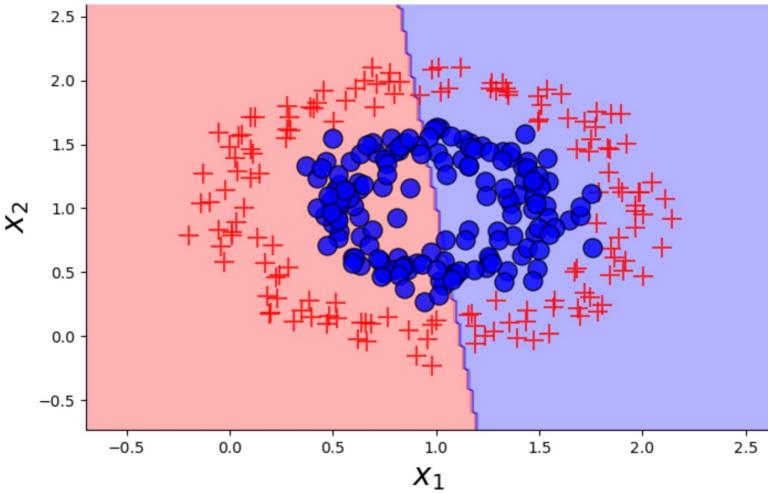


On ne trouvera pas de bonne frontière de séparation via  
issu d'un modèle

$$\hat{\theta}_0 + \hat{\theta}_1 x^1 + \hat{\theta}_2 x^2 = 0.5$$

$$p_{Y|X=x}^{\theta}(\hat{y}) = \theta_0 + \theta_1 x^1 + \theta_2 x^2$$

# 3.1 Régression logistique



On ne trouvera pas de bonne frontière de séparation via  
issu d'un modèle

$$\hat{\theta}_0 + \hat{\theta}_1 x^1 + \hat{\theta}_2 x^2 = 0.5$$

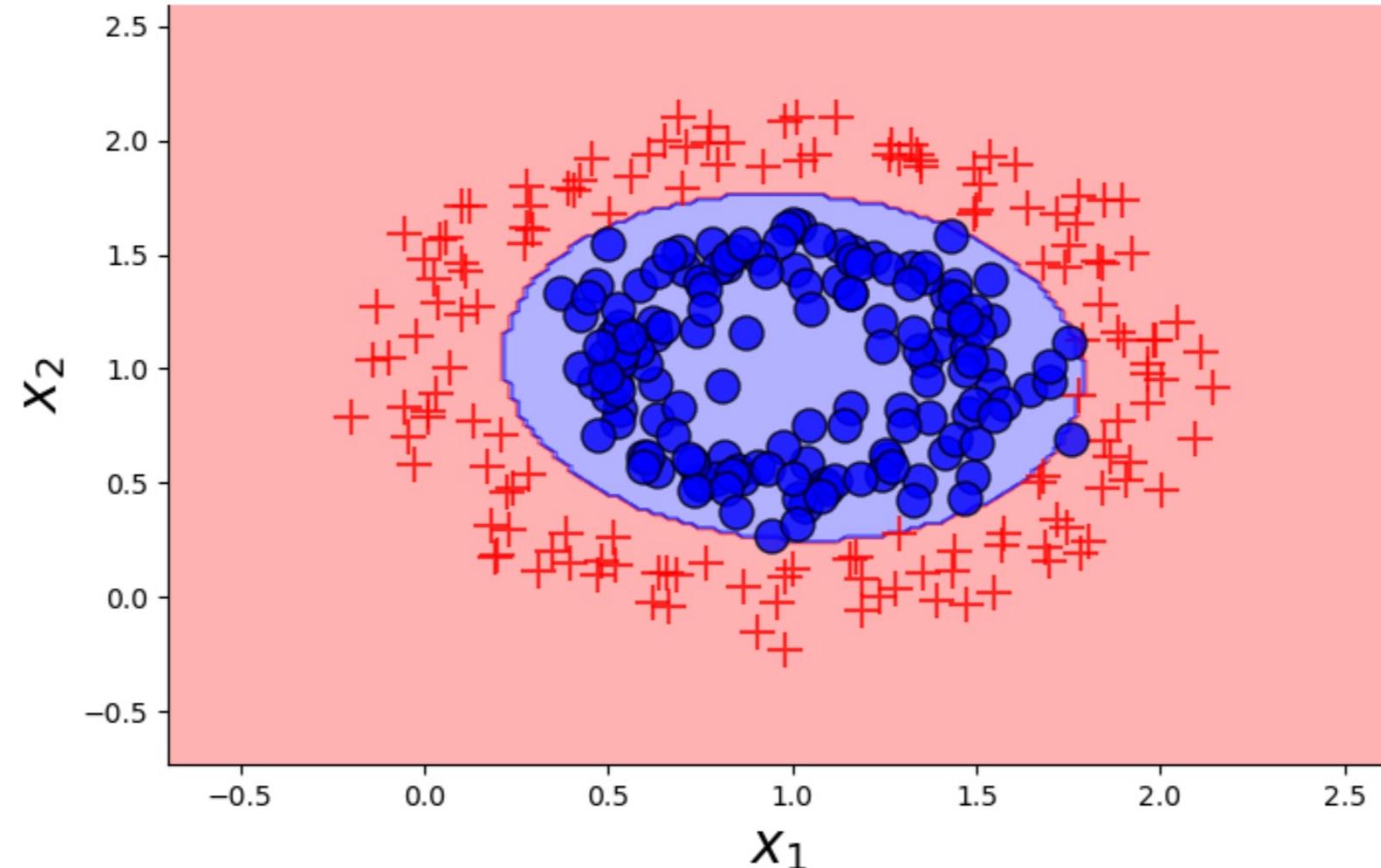
$$p_{Y|X=x}^\theta(\hat{y}) = \theta_0 + \theta_1 x^1 + \theta_2 x^2$$



Insérer de la non-linéarité dans les variables

$$p_{Y|X=x}^\theta(\hat{y}) = \theta_0 + \theta_1 x^1 + \theta_2 x^2 + \theta_3 (x^1)^2 + \theta_4 (x^2)^2$$

# 3.1 Régression logistique



! L'ingénierie de variables est très utile et potentiellement puissante

# 3.2 Arbre de décision

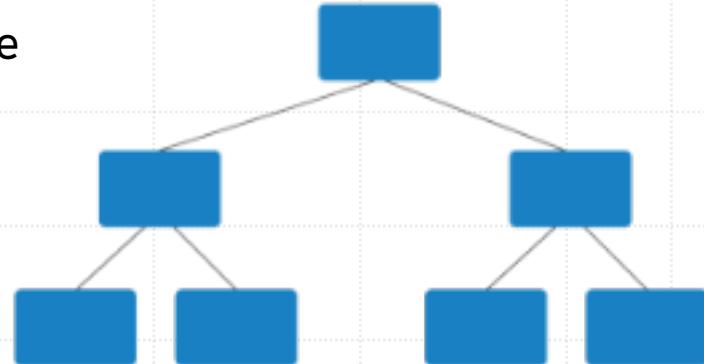
**L'intuition :** Découper l'espace des données en rectangles (ou hyper-rectangles) de plus en plus petits. L'objectif est de créer des zones où une seule classe est majoritaire ("pureté").

**L'algorithme:** Pour chaque nœud, l'algorithme cherche :

1. La variable  $x^j$  (ex:  $x^1$ )
2. Le seuil de coupure (ex: 0.8 ) ...qui minimise l'impureté du noeud

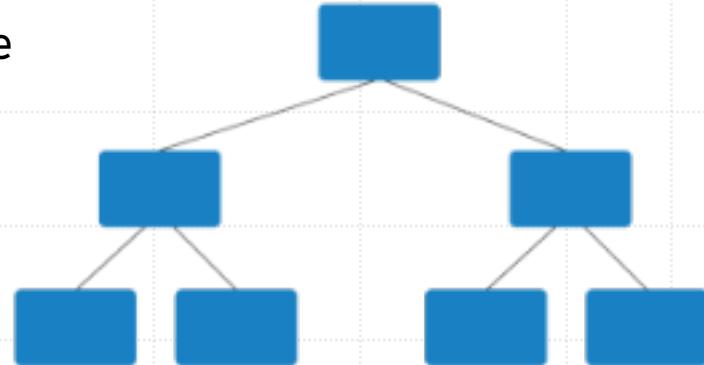
**Mesure de l'impureté (Gini) :**  $G = 1 - \sum_k p_k^2$

où  $p_k$  est la proportion de la classe  $k$  dans le nœud.



# 3.2 Arbre de décision

**L'intuition :** Découper l'espace des données en rectangles (ou hyper-rectangles) de plus en plus petits. L'objectif est de créer des zones où une seule classe est majoritaire ("pureté").



**L'algorithme:** Pour chaque nœud, l'algorithme cherche :

1. La variable  $x^j$  (ex:  $x^1$ )
2. Le seuil de coupure (ex: 0.8 ) ...qui minimise l'impureté du noeud

**Mesure de l'impureté (Gini) :** 
$$G = 1 - \sum_k p_k^2$$

où  $p_k$  est la proportion de la classe  $k$  dans le nœud.

**Divers hyperparamètres:**

1. **max\_depth** : La profondeur maximale de l'arbre.
2. **max\_features**: Nombre de variables explorées pour tenter de diviser un noeud
3. **min\_samples\_split** : Nb minimum d'échantillons requis pour autoriser une nouvelle division
4. **min\_samples\_leaf** : Nb nombre minimum d'échantillons qui doivent rester dans une feuille finale.

# 3.2 Arbre de décision

**L'intuition :** Découper l'espace des données en rectangles (ou hyper-rectangles) de plus en plus petits. L'objectif est de créer des zones où une seule classe est majoritaire ("pureté").

**L'algorithme:** Pour chaque nœud, l'algorithme cherche :

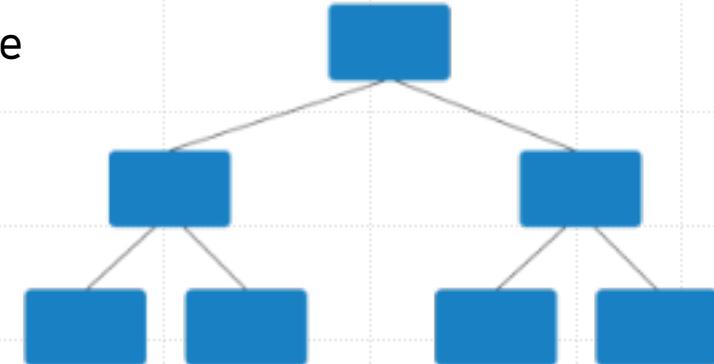
1. La variable  $x^j$  (ex:  $x^1$ )
2. Le seuil de coupure (ex: 0.8 ) ...qui minimise l'impureté du noeud

**Mesure de l'impureté (Gini) :**  $G = 1 - \sum_k p_k^2$

où  $p_k$  est la proportion de la classe  $k$  dans le nœud.

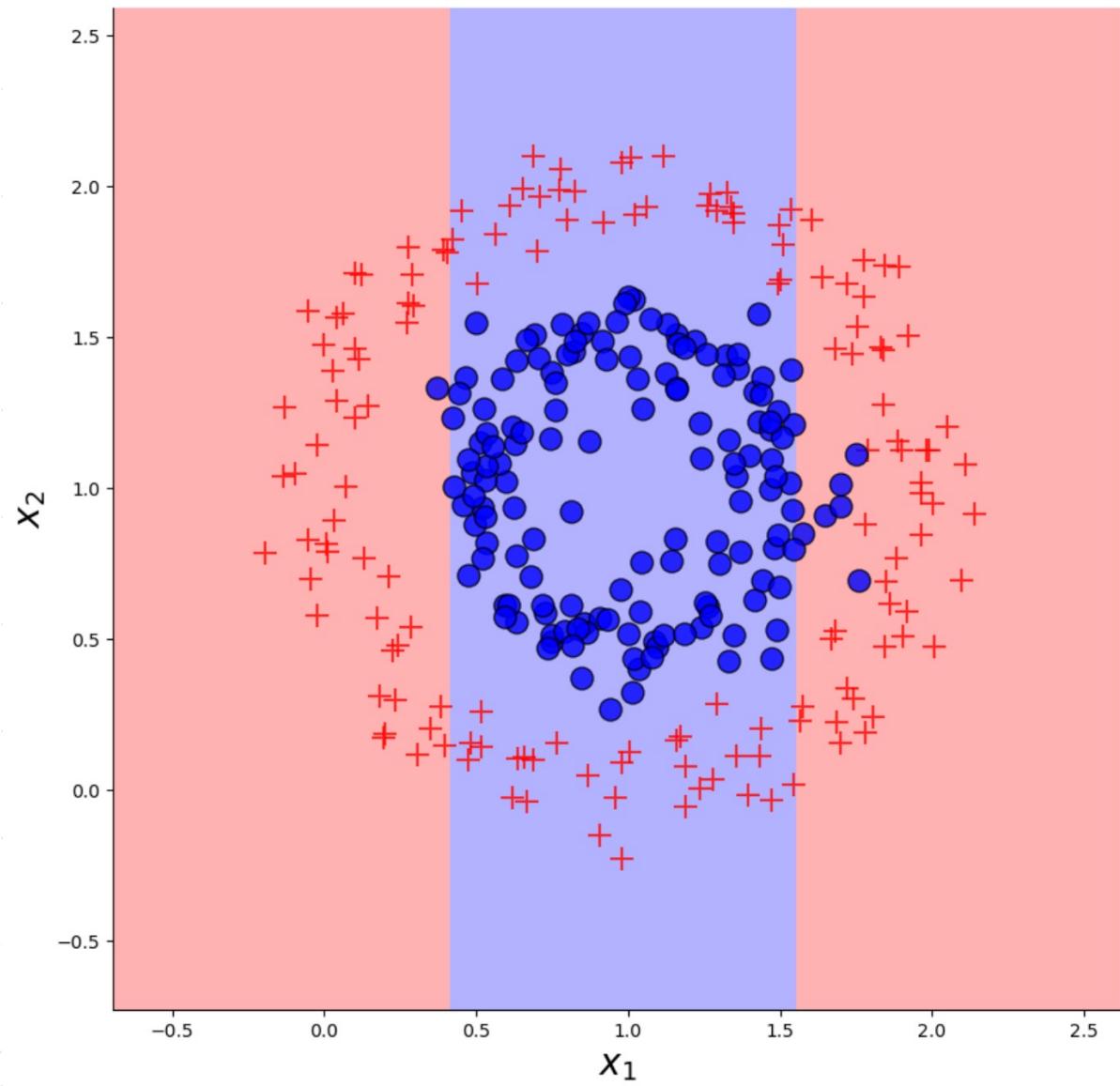
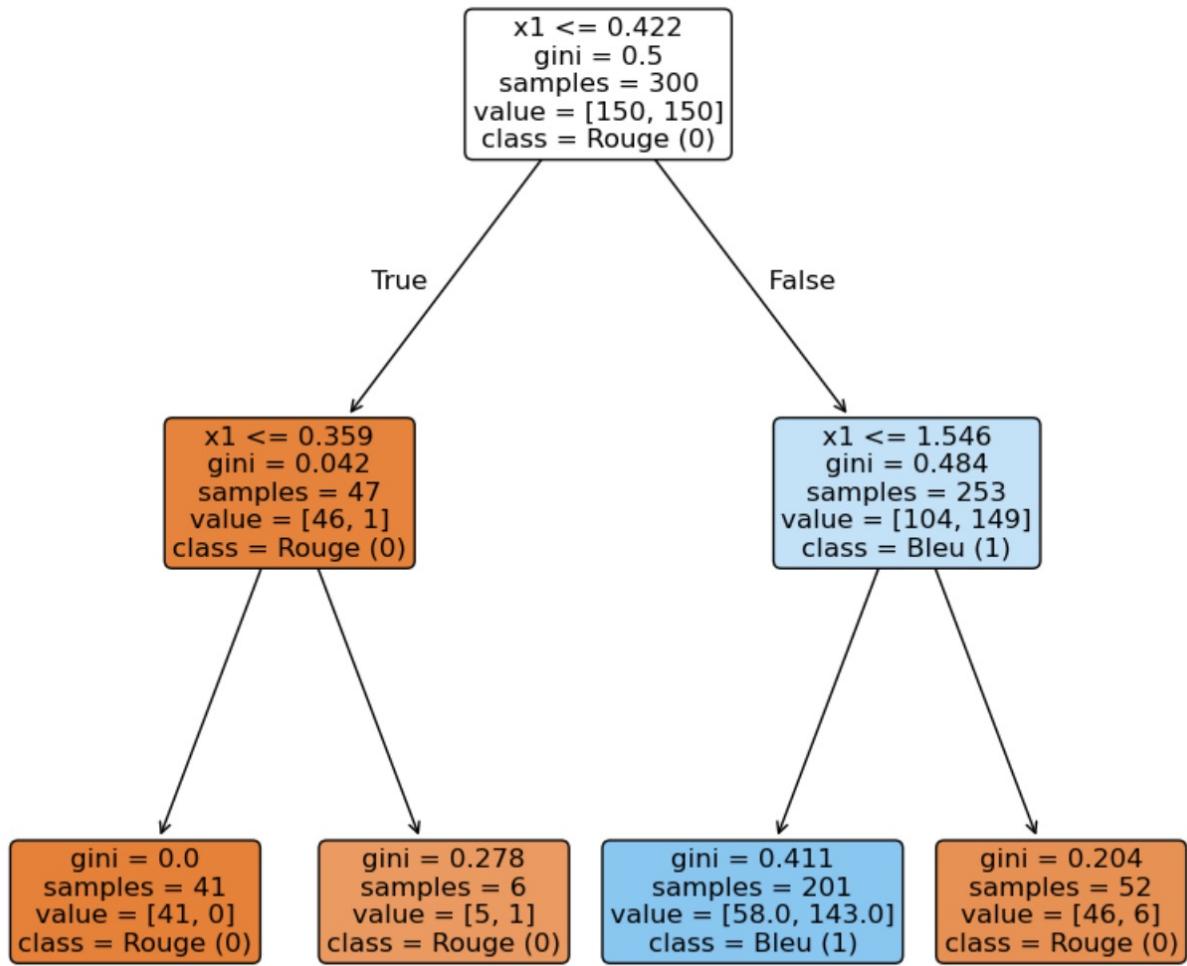
**Divers hyperparamètres:**

1. **max\_depth** : La profondeur maximale de l'arbre.
2. **max\_features**: Nombre de variables explorées pour tenter de diviser un noeud
3. **min\_samples\_split** : Nb minimum d'échantillons requis pour autoriser une nouvelle division
4. **min\_samples\_leaf** : Nb nombre minimum d'échantillons qui doivent rester dans une feuille finale.

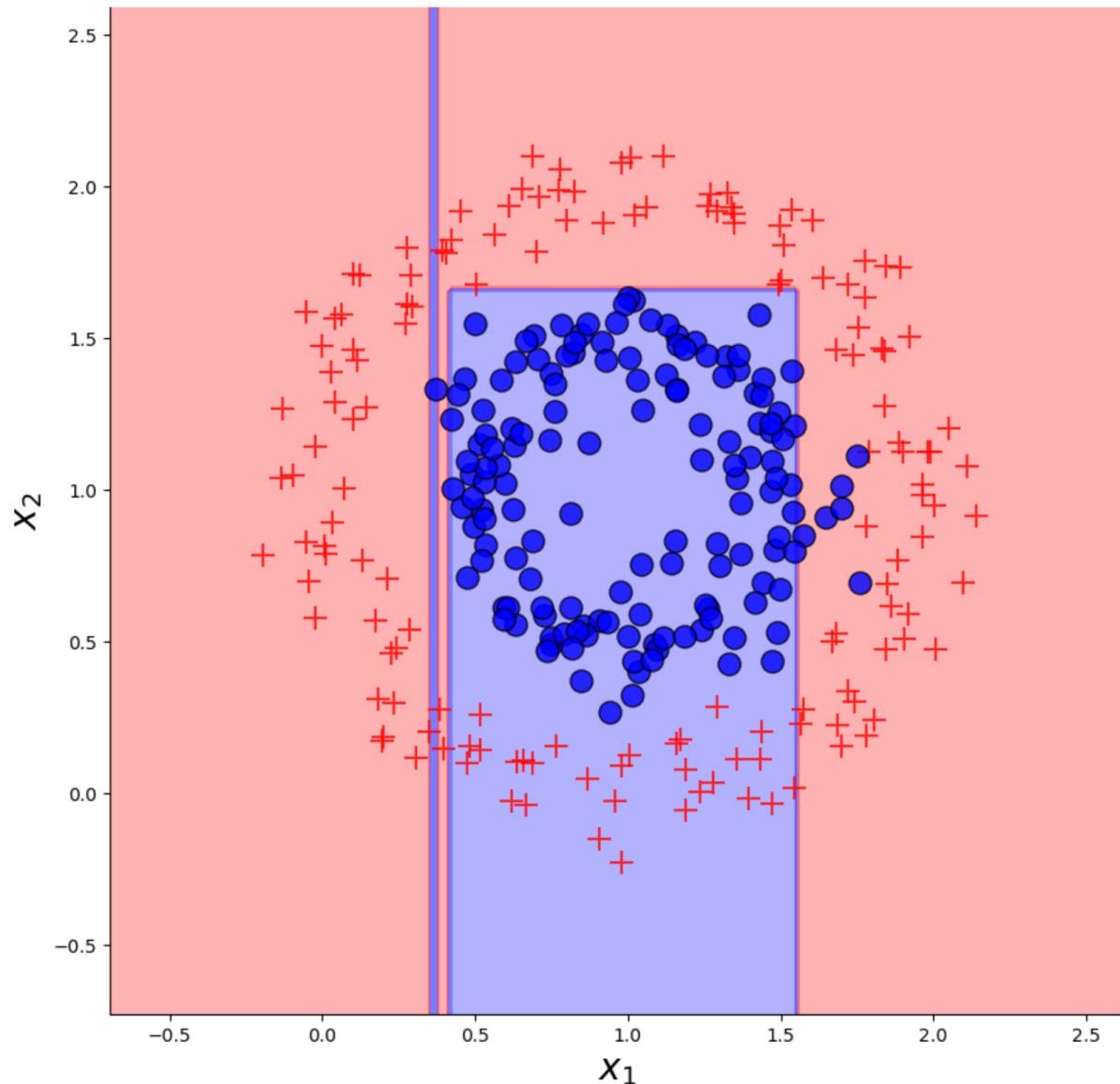
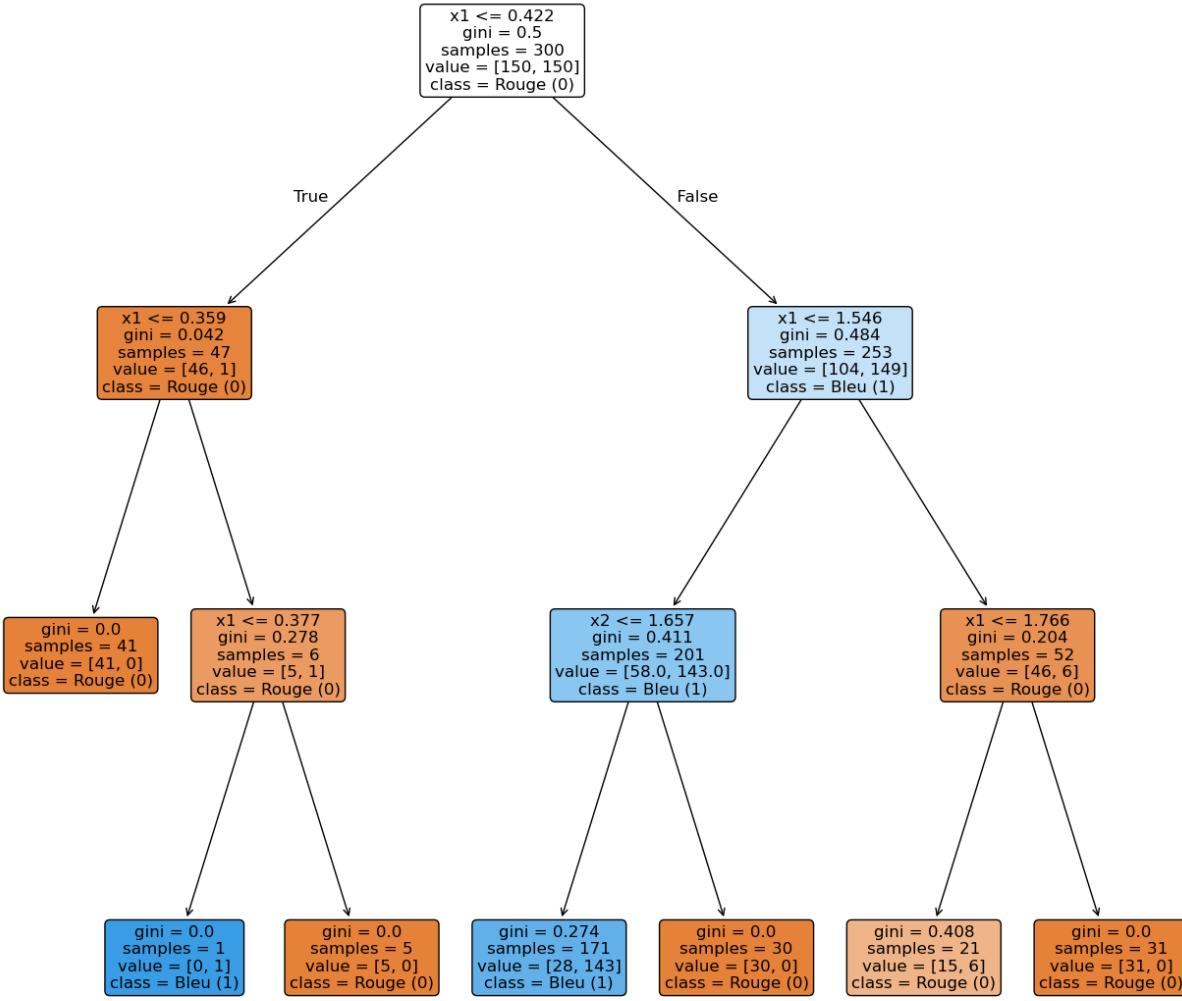


<https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html>

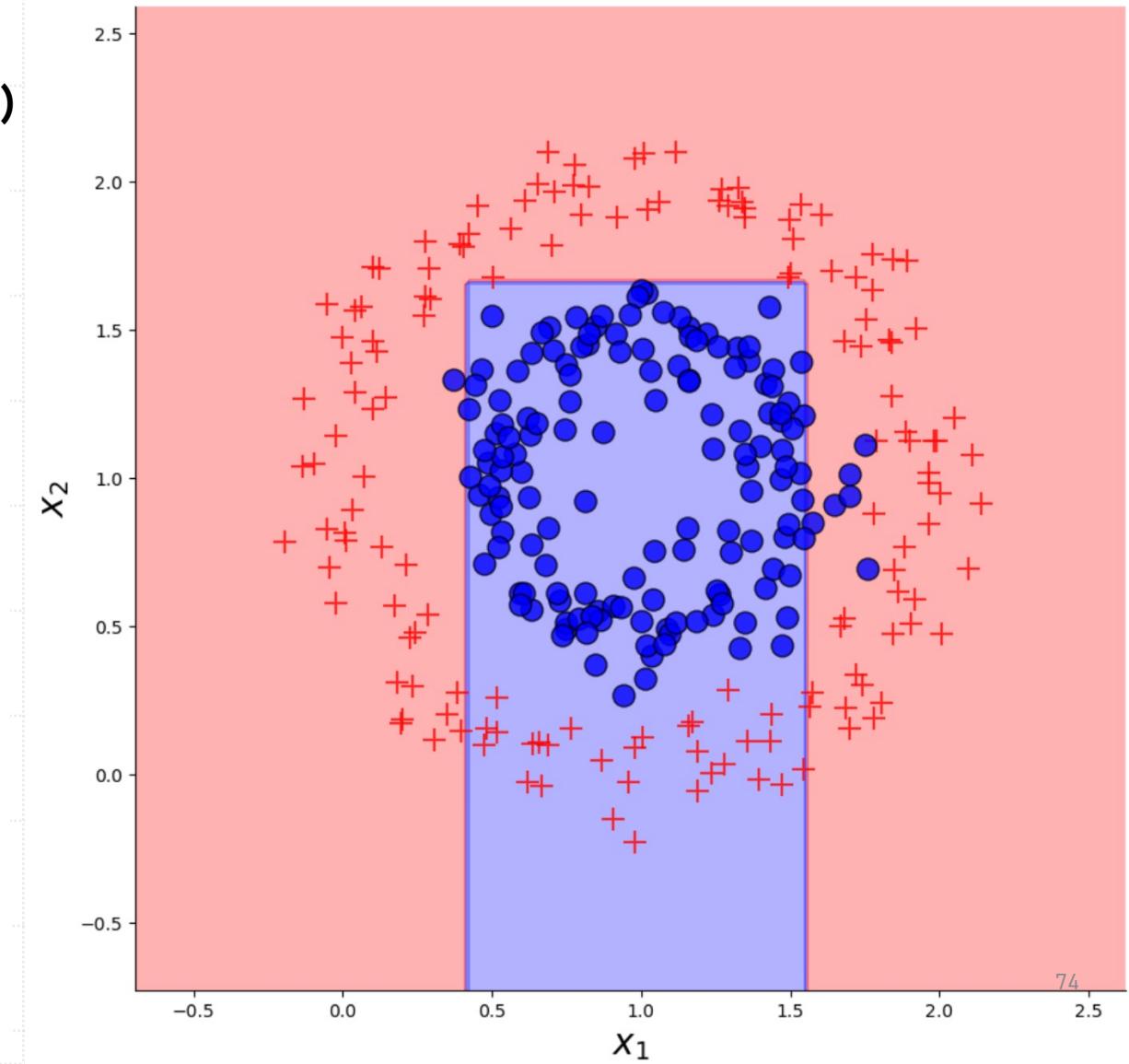
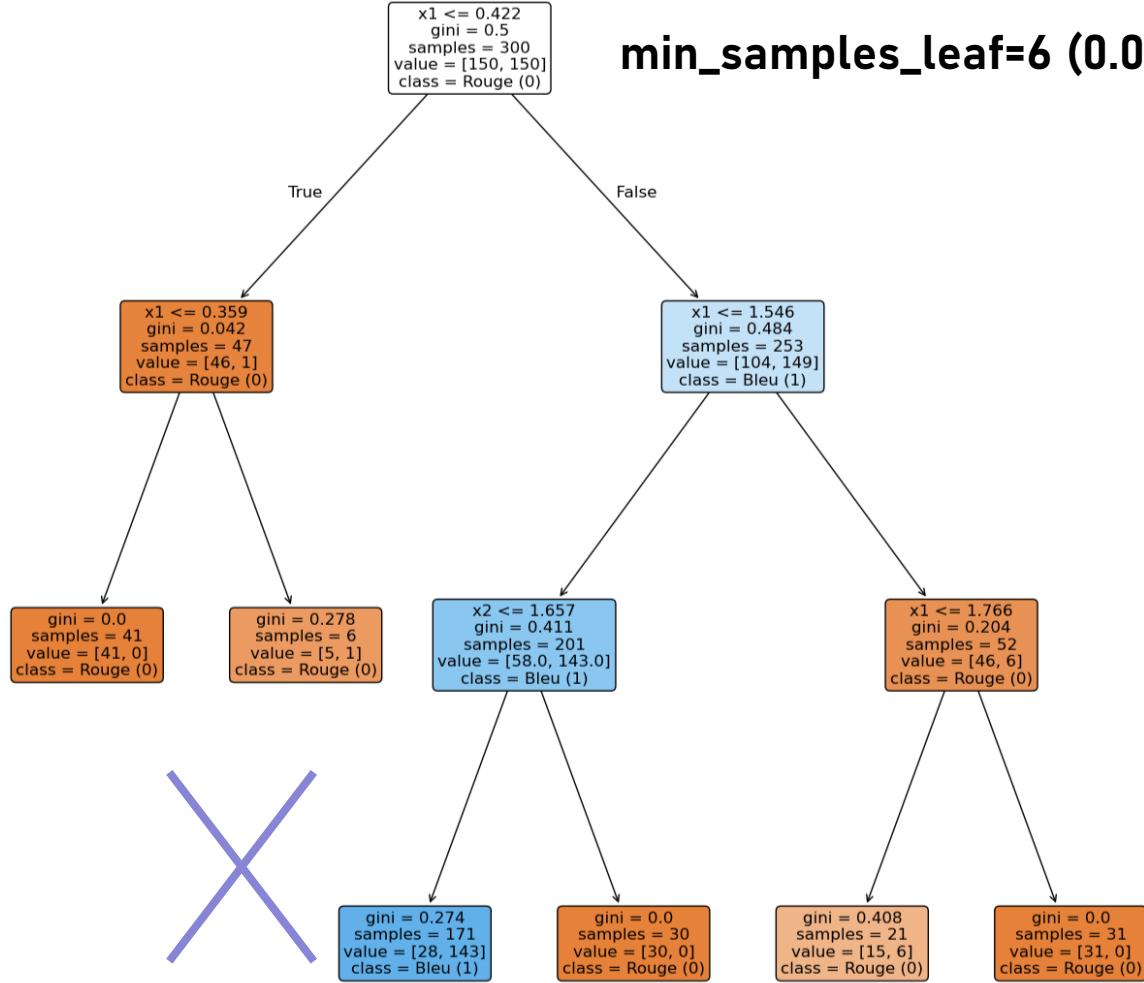
# 3.2 Arbre de décision



# 3.2 Arbre de décision



# 3.2 Arbre de décision



# 3.3 Forêt aléatoire

---

## 1. Le Principe : L'union fait la force (Méthode d'Ensemble)

**Concept** : Remplacer un arbre unique (instable et surapprentissage) par une multitude (ex: 100 ou 500).

**Prédiction** : Le modèle final agrège les décisions de tous les arbres.

*Classification* : Vote majoritaire (la classe la plus populaire l'emporte).

*Régression* : Moyenne des prédictions.

# 3.3 Forêt aléatoire

## 1. Le Principe : L'union fait la force (Méthode d'Ensemble)

**Concept** : Remplacer un arbre unique (instable et surapprentissage) par une multitude (ex: 100 ou 500).

**Prédiction** : Le modèle final agrège les décisions de tous les arbres.

*Classification* : Vote majoritaire (la classe la plus populaire l'emporte).

*Régression* : Moyenne des prédictions.

## 2. Les deux sources d'aléatoire (Pourquoi "Random" ?)

Arbres différents et indépendants (décorrélation), aléatoire à deux niveaux :

### A. Sur les Données (Bagging / Bootstrap) :

Chaque arbre est entraîné sur un **sous-échantillon différent** du jeu de données, tire avec remise.

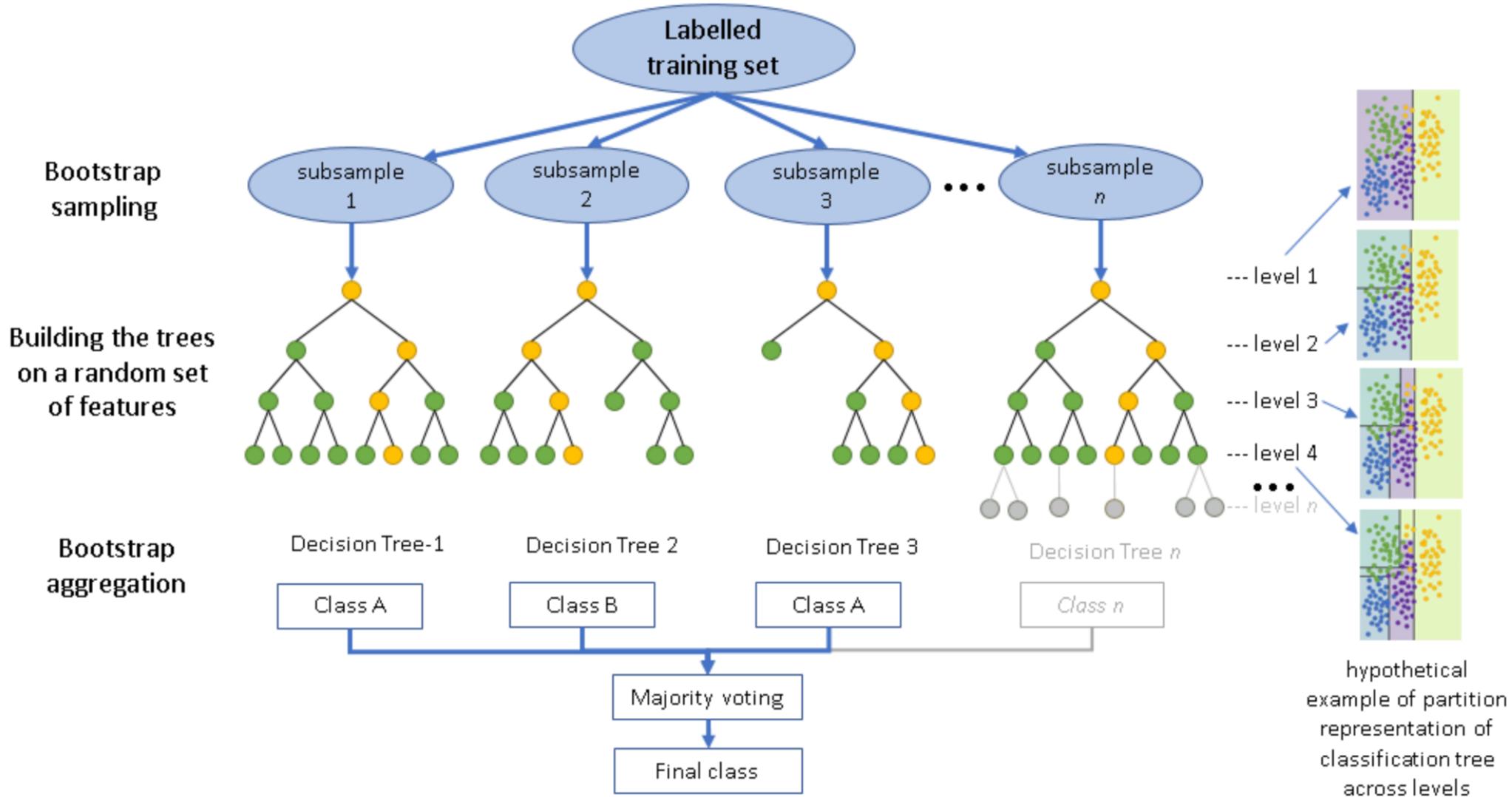
→ *Effet* : Chaque arbre apprend sur une vision légèrement différente de la réalité.

### B. Sur les Variables (Feature Sampling) :

À chaque nœud de chaque arbre, l'algorithme ne considère qu'un **sous-ensemble aléatoire de variables** pour trouver la meilleure division (ex: log2 ou racine carrée).

→ *Effet* : Force le modèle à explorer des combinaisons de variables variées, évite qu'une variable dominante la structure de tous les arbres.

# 3.3 Forêt aléatoire



### 3.3 Forêt aléatoire

```
rf_model = RandomForestClassifier(n_estimators=100, max_depth=3, random_state=42)
```

