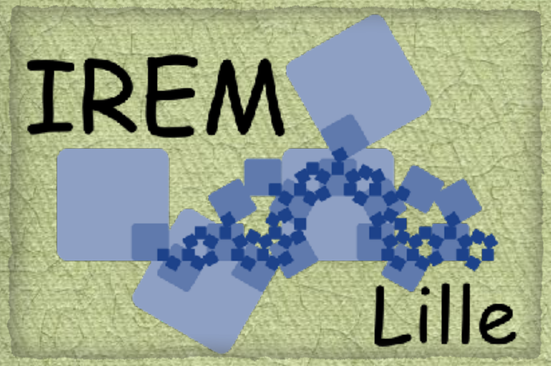


Groupe Info

IREM de Lille

Asli Grimaud

IA et Apprentissage Automatique

Mars - Avril 2025

asli.grimaud@ac-lille.fr

- Apprentissage supervisé
  - Algorithme des k-plus proches voisins
  - Algorithme ID3
- Apprentissage non supervisé
  - Classification hiérarchique ascendante
  - Algorithme des k-moyennes

Apprentissage supervisé

Technique d'apprentissage automatique où un algorithme apprend à partir d'un ensemble de données étiquetées, c'est-à-dire des données pour lesquelles les résultats attendus sont connus.

- Problèmes de classification (discret)

  L'algorithme apprend à assigner une catégorie à une entrée donnée

  Ex : spam ou non-spam ?
- Problèmes de régression (continue)

  L'algorithme apprend à prédire une valeur continue pour une entrée donnée

  Ex : Prédire le prix de vente d'une maison en fonction de ses caractéristiques

Ensemble d'apprentissage

$$Z = \left\{(x_i, y_i) \mid i \in [\![0, n-1]\!], x_i \in X, y_i \in Y\right\}$$

$x_i \in \mathbb{R}^d$: un individu de d mesures

$y_i \in \mathbb{N}$ (classification) ou $y_i \in \mathbb{R}^d$ (régression) : une étiquette

Ex : Un ensemble de mails déjà étiqueté spam ou non spam

Un ensemble de maisons avec des caractéristiques et le prix de vente

Jeu de données Iris *Fisher 1936*

Iris setosa

Iris versicolor

Iris virginica

`iris.data`

- 150 *individus*
- *quatre caractéristiques*
  - *longueur des sépales*
  - *largeur des sépales*
  - *longueur des pétales*
  - *largeur des pétales*

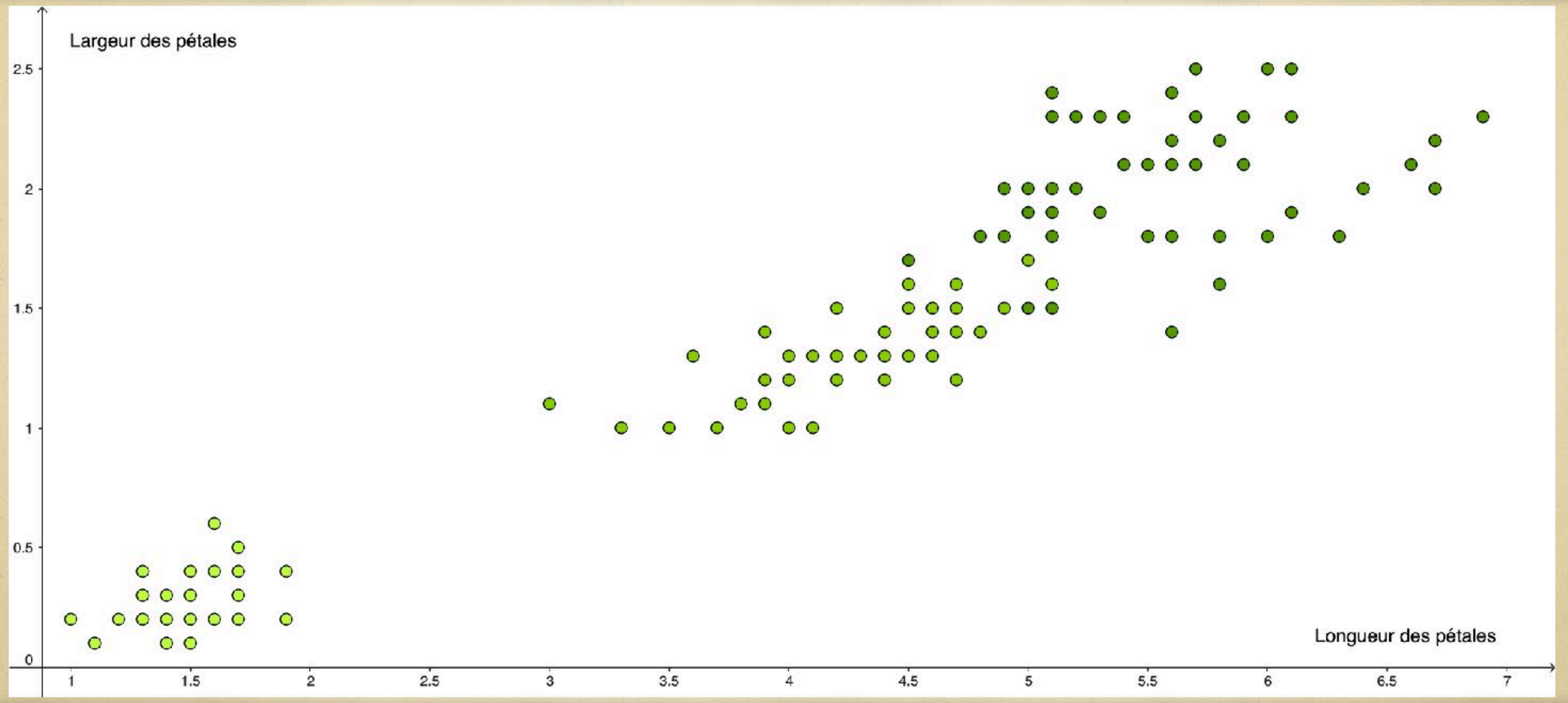
- $n = 150$
- $x_i = (x_{i0}, x_{i1}, x_{i2}, x_{i3}) \in \mathbb{R}^4$
- c_0 : *Iris setosa*

  c_1 : *Iris versicolor*

  c_2 : *Iris virginica*

- *problème de classification : étant donné un individu (4 caractéristiques), prédire à quelle classe il appartient*
- *problème de régression : étant donné 3 caractéristiques, prédire la valeur de la dernière*

Algorithme des k-plus proches voisins



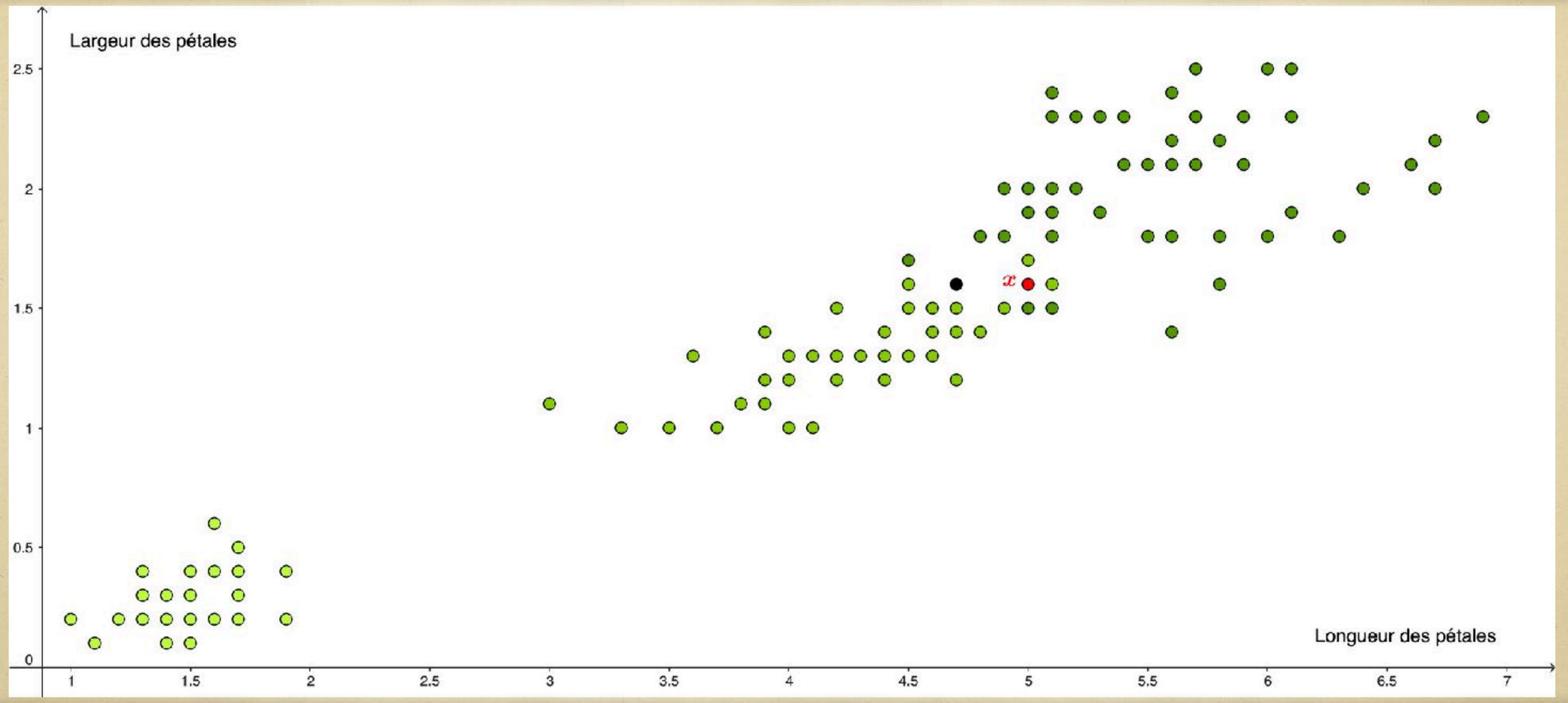
Ensemble d'apprentissage, projection sur 2D

pas de phase d'apprentissage

Iris setosa *Iris versicolor* *Iris virginica*

Algorithme des k-plus proches voisins

classification



$$x = (6{,}3;\ 3{,}3;\ 5{,}0;\ 1{,}6)$$

Algorithme des k-plus proches voisins

classification

Calcul de distances entre les individus et $x = (x_0, x_1, x_2, x_3)$

- distance de Manhattan
- distance de Minkowski
- distance de Chbyshev
- distance de Hamming
- distance euclidienne
- ...

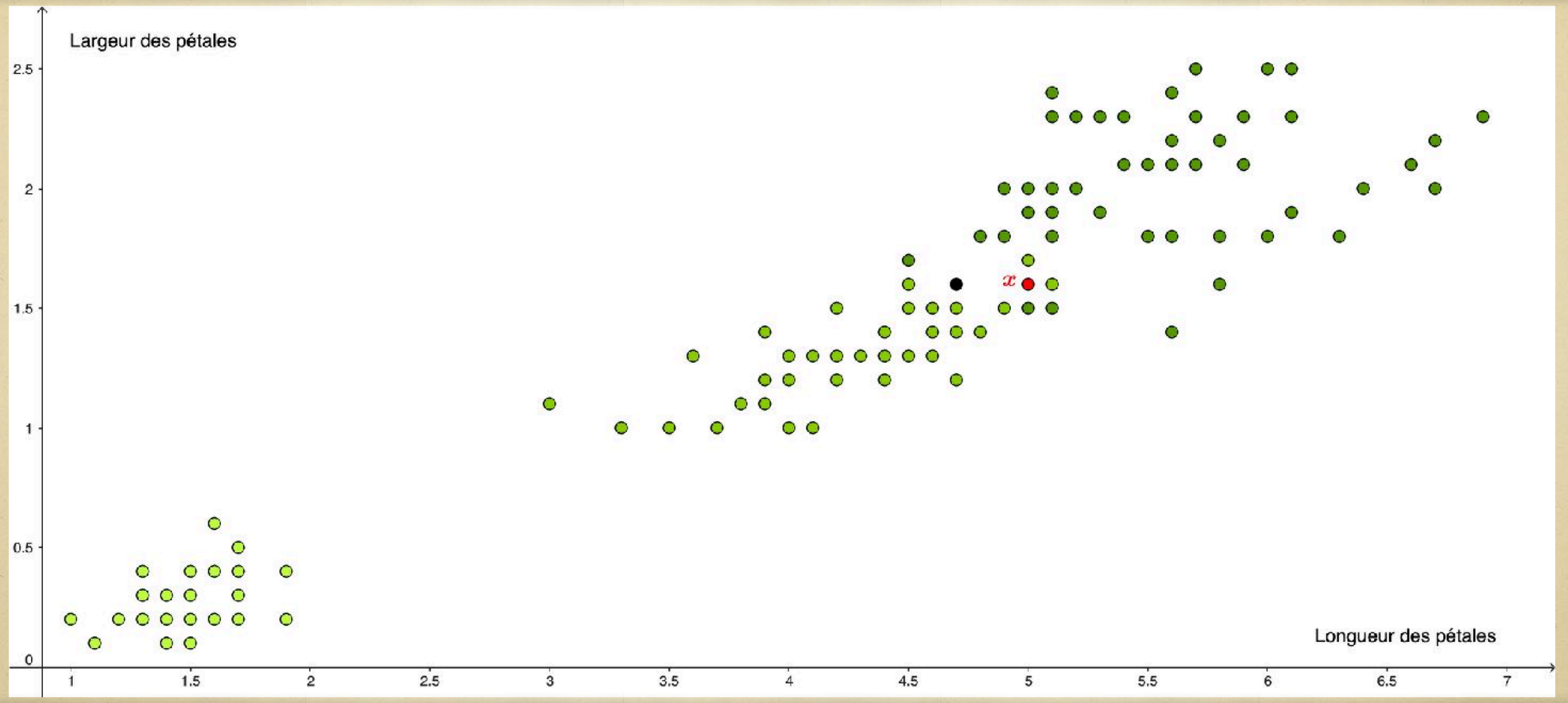
choix : distance euclidienne

$$\forall i \in [\![0, n-1]\!], \delta(x, x_i) = \sqrt{\sum_{j=0}^{d-1} \left(x_j - x_{ij}\right)^2}$$

et s'intéresser aux k plus proches...

Algorithme des k-plus proches voisins

classification

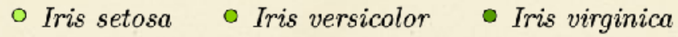


$x = (6{,}3; 3{,}3; 5{,}0; 1{,}6)$, $k = 1$

Algorithme des k-plus proches voisins

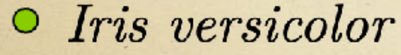
classification

classe de $x = (6{,}3; 3{,}3; 5{,}0; 1{,}6)$?



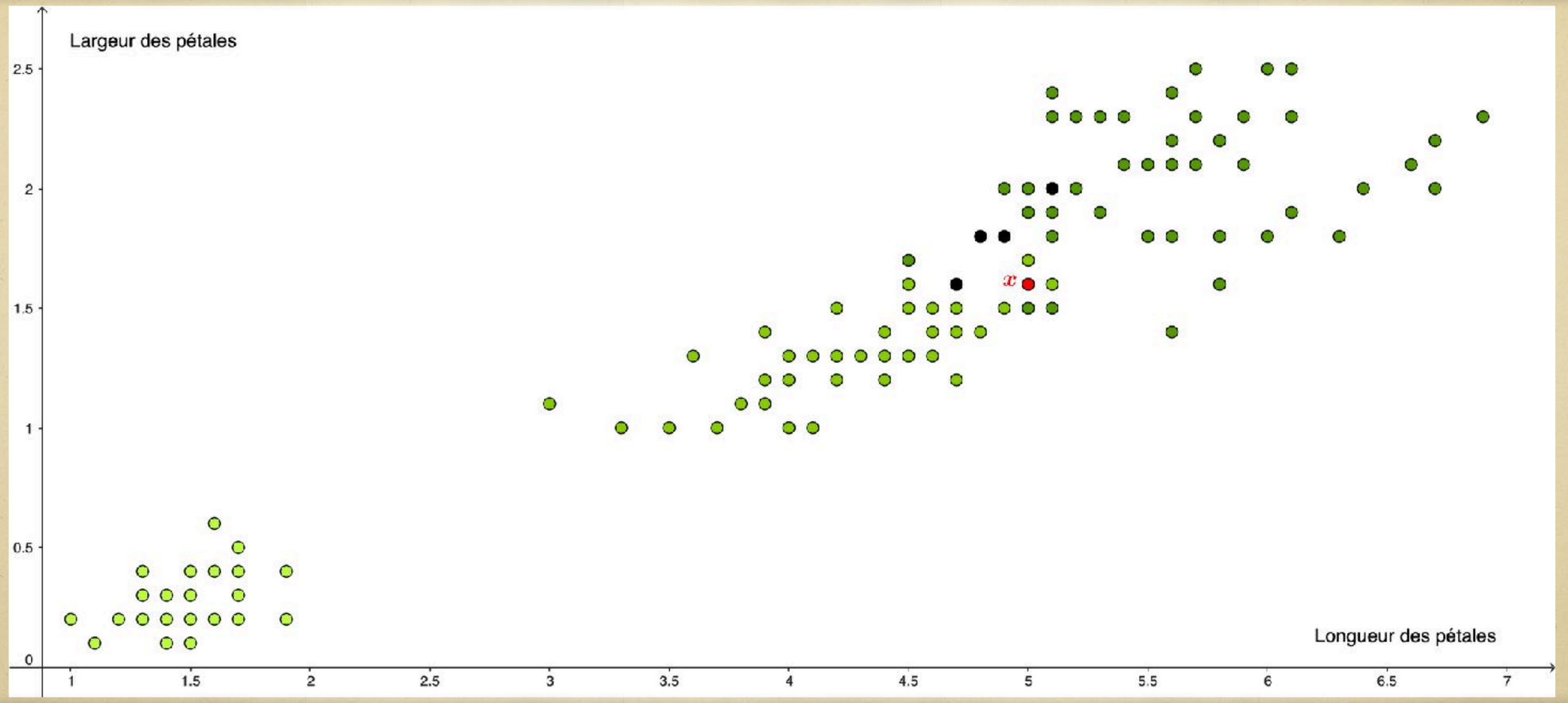
$k = 1$

- classe du voisin le plus proche
- modèle facile à interpréter
- sensibilité aux bruits
- mauvaise généralisation



Algorithme des k-plus proches voisins

classification

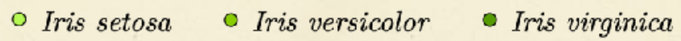


$x = (6{,}3; 3{,}3; 5{,}0; 1{,}6)$, $k = 4$

Algorithme des k-plus proches voisins

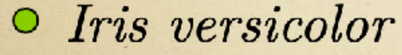
classification

classe de $x = (6{,}3; 3{,}3; 5{,}0; 1{,}6)$?



$k = 1$

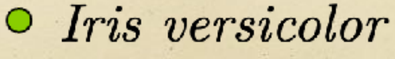
- classe du voisin le plus proche
- modèle facile à interpréter
- sensibilité aux bruits
- mauvaise généralisation



$k = 4$

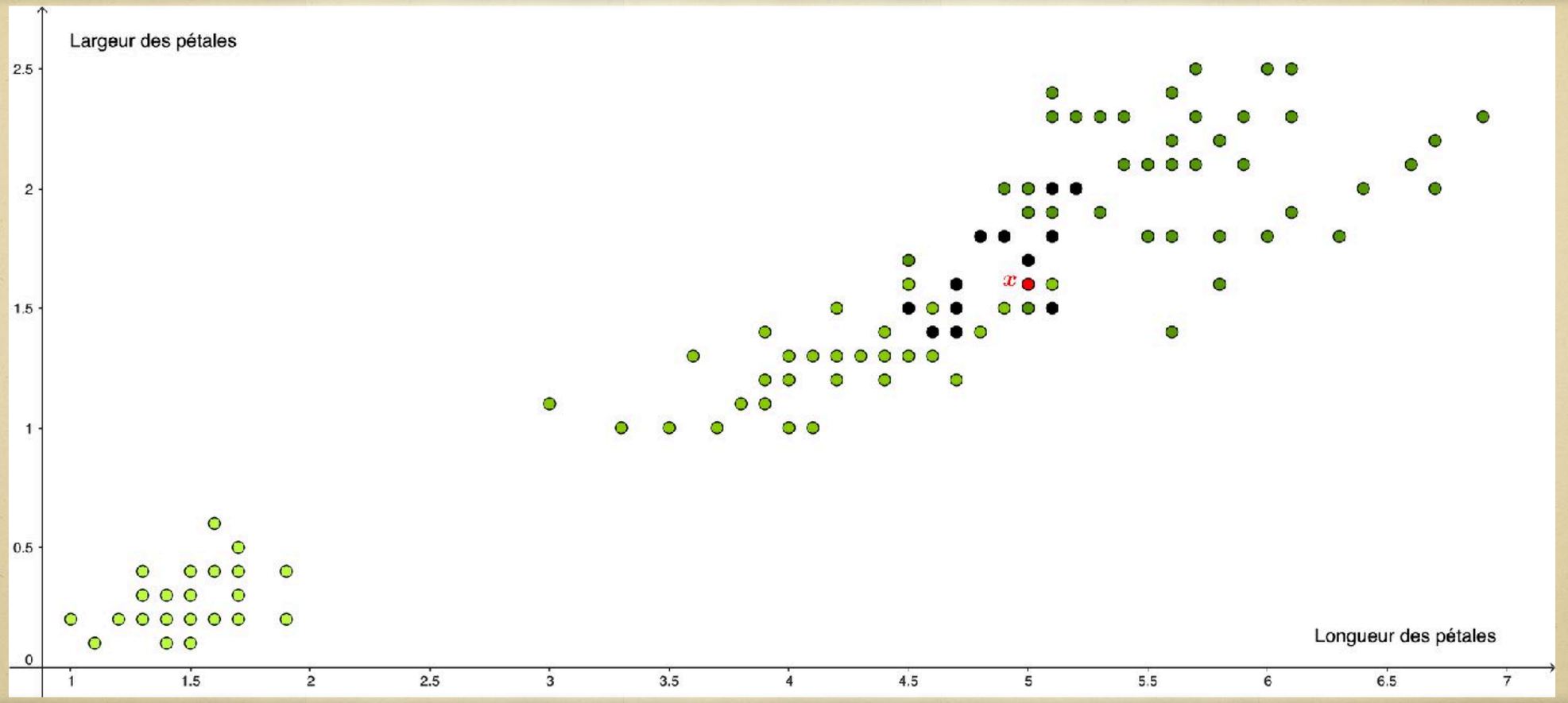
- 2 versicolor, 2 virginica
- ambiguïté de décision
- choix :
  hasard, somme des distances,
  pondération de vote, …

$$\frac{1}{2}, \frac{1}{4}, \frac{1}{6}, \frac{1}{12}$$



Algorithme des k-plus proches voisins

classification



$x = (6{,}3; 3{,}3; 5{,}0; 1{,}6)$, $k = 13$

Algorithme des k-plus proches voisins

classification

classe de $x = (6{,}3; 3{,}3; 5{,}0; 1{,}6)$ *?*

Iris setosa *Iris versicolor* *Iris virginica*

$k = 1$

- *classe du voisin le plus proche*
- *modèle facile à interpréter*
- *sensibilité aux bruits*
- *mauvaise généralisation*

Iris versicolor

$k = 4$

- *2 versicolor, 2 virginica*
- *ambiguïté de décision*
- *choix :*
  hasard, somme des distances,
  pondération de vote, …

$$\frac{1}{2}, \frac{1}{4}, \frac{1}{6}, \frac{1}{12}$$

Iris versicolor

$k = 13$

- *7 versicolor, 6 virginica*
- *choix : majorité*

Iris versicolor

Algorithme des k-plus proches voisins

et pourquoi $k = 1$, $k = 4$ *ou* $k = 13$ *???*

choix : validation croisée, expérimentation empirique, heuristique $(\sqrt{n})$, *...*

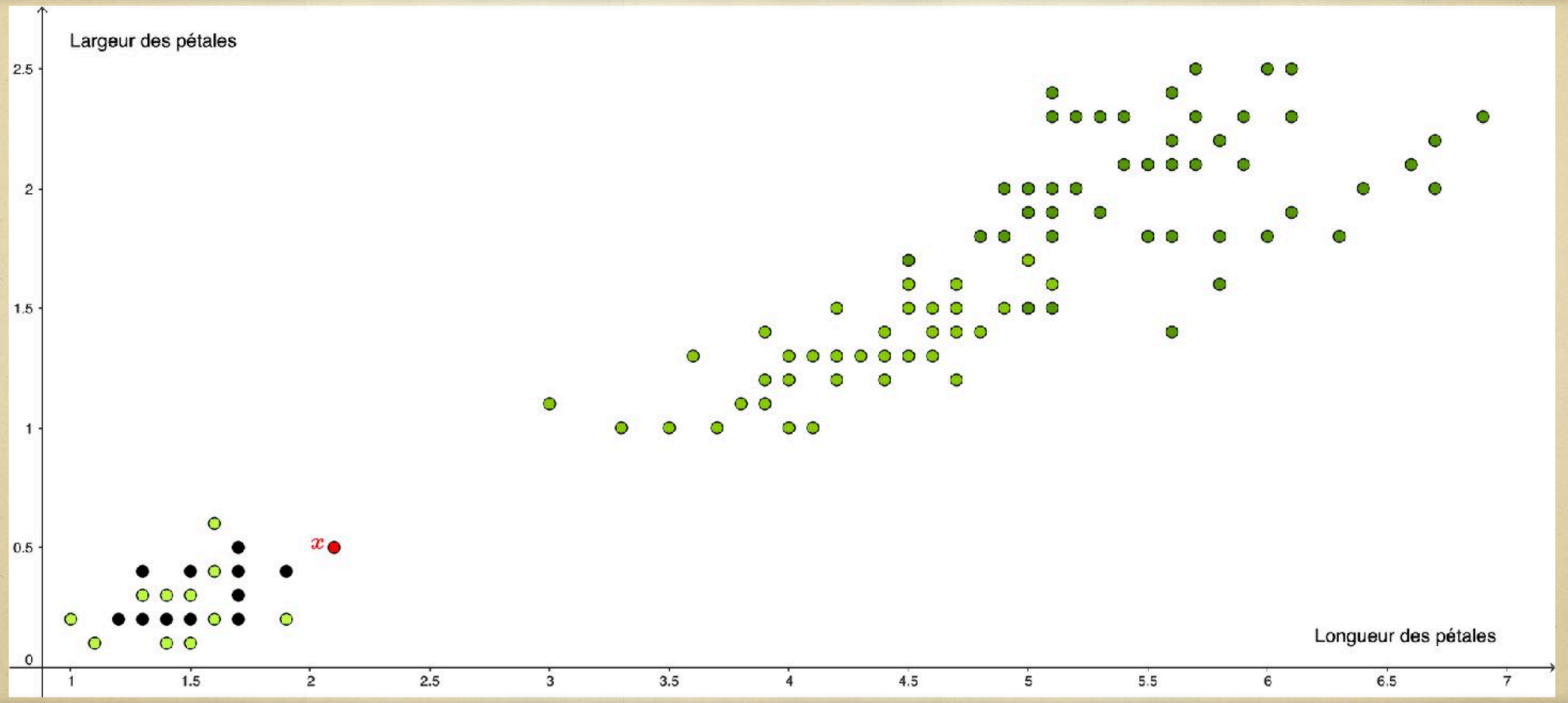
petit k : sensible au bruit, risque de sur-ajustement

grand k : trop lisse, risque de sous-ajustement

un k qui équilibre des deux effets

Algorithme des k-plus proches voisins

classification



- *Iris setosa*

$y = (6{,}3; 3{,}3; 2{,}1; 0{,}5)$, $k = 13$

Algorithme des k-plus proches voisins

régression

Calcul de distances entre les individus et $x = (x_0, x_1, x_2, \times)$

$$\forall i \in [\![0, n-1]\!], \delta(x, x_i) = \sqrt{\sum_{j=0}^{d-2} \left(x_j - x_{ij}\right)^2}$$

choix : moyenne sur x_{i4} des k individus les plus proches

pour $x = (6{,}3; 3{,}3; 5{,}0; x_4)$, avec $k = 13$, on obtient $x_4 = 1{,}8$.

Algorithme des k-plus proches voisins

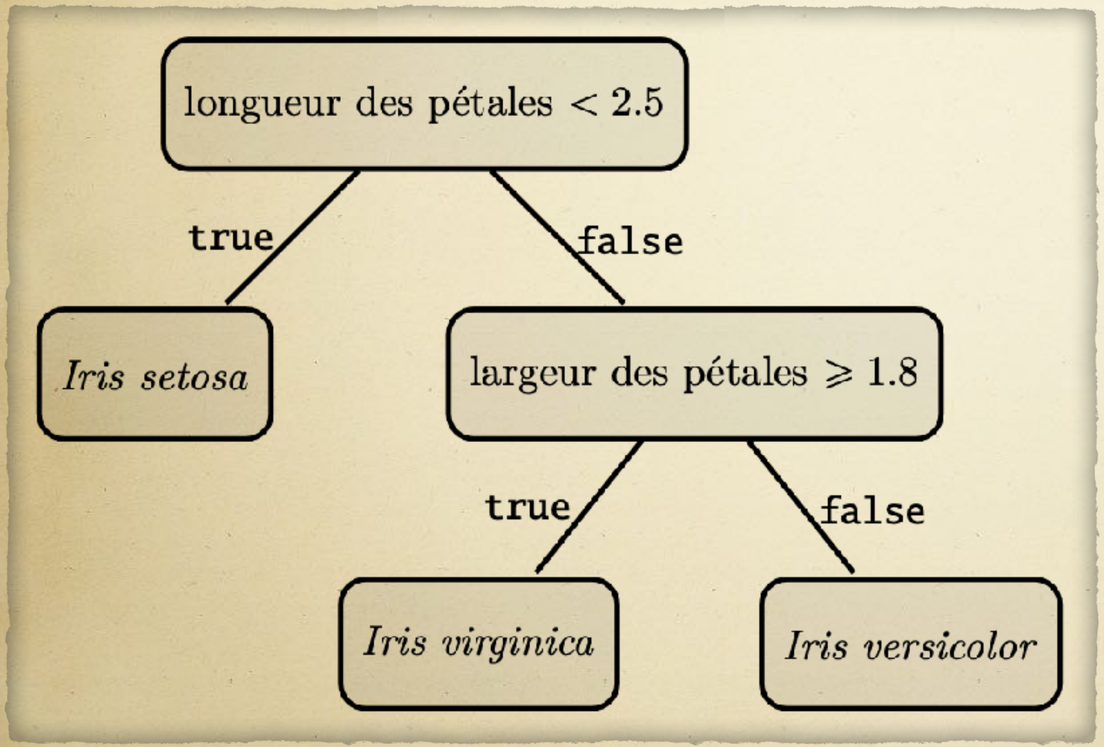
statistiques, grande quantité de données, choix, choix, choix…

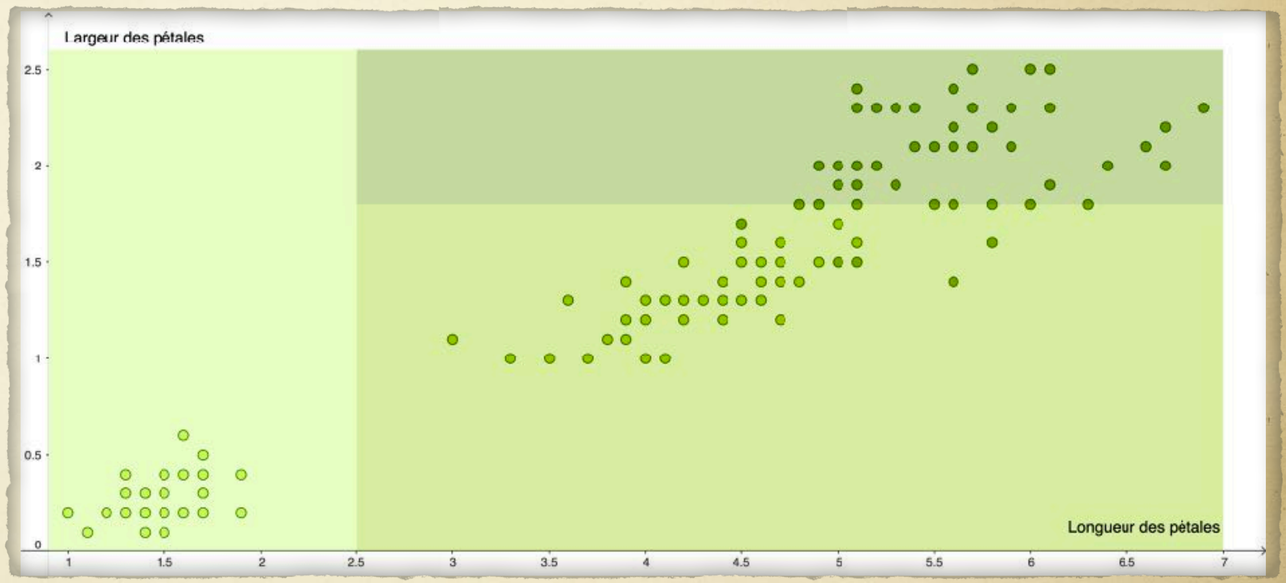
Malgré ces choix, les décisions ne sont pas mises en question par le programme…contrairement aux humaines qui ont une capacité de doute…

Algorithme ID3

Iterative Dichotomiser 3

Arbres de décision : phase d'apprentissage





- *ensemble d'apprentissage*
- *construction d'arbre de manière algorithmique*

Algorithme ID3

Algorithme glouton qui cherche à maximiser localement le gain d'information, mais il ne garantit pas d'obtenir un arbre de décision optimal à l'échelle globale.

Entropie de Shannon : Mesure de l'incertitude d'un système.

Soient $c_0, c_1, \ldots, c_{q-1}$ un ensemble de q classes et Z l'ensemble d'apprentissage.

$$H(Z) = -\sum_{i=0}^{q-1} p_i \log p_i$$

où $p_i \neq 0$ est la probabilité d'appartenance d'un individu à la classe c_i.

Algorithme ID3

Algorithme glouton qui cherche à maximiser localement le gain d'information, mais il ne garantit pas d'obtenir un arbre de décision optimal à l'échelle globale.

Gain d'information : Mesure la réduction de l'incertitude après avoir divisé un ensemble de données en fonction d'un attribut.

$$G(Z, a, s) = H(Z) - \left(\frac{|Z_{a \leq s}|}{|Z|} H(Z_{a \leq s}) + \frac{|Z_{a > s}|}{|Z|} H(Z_{a > s}) \right)$$

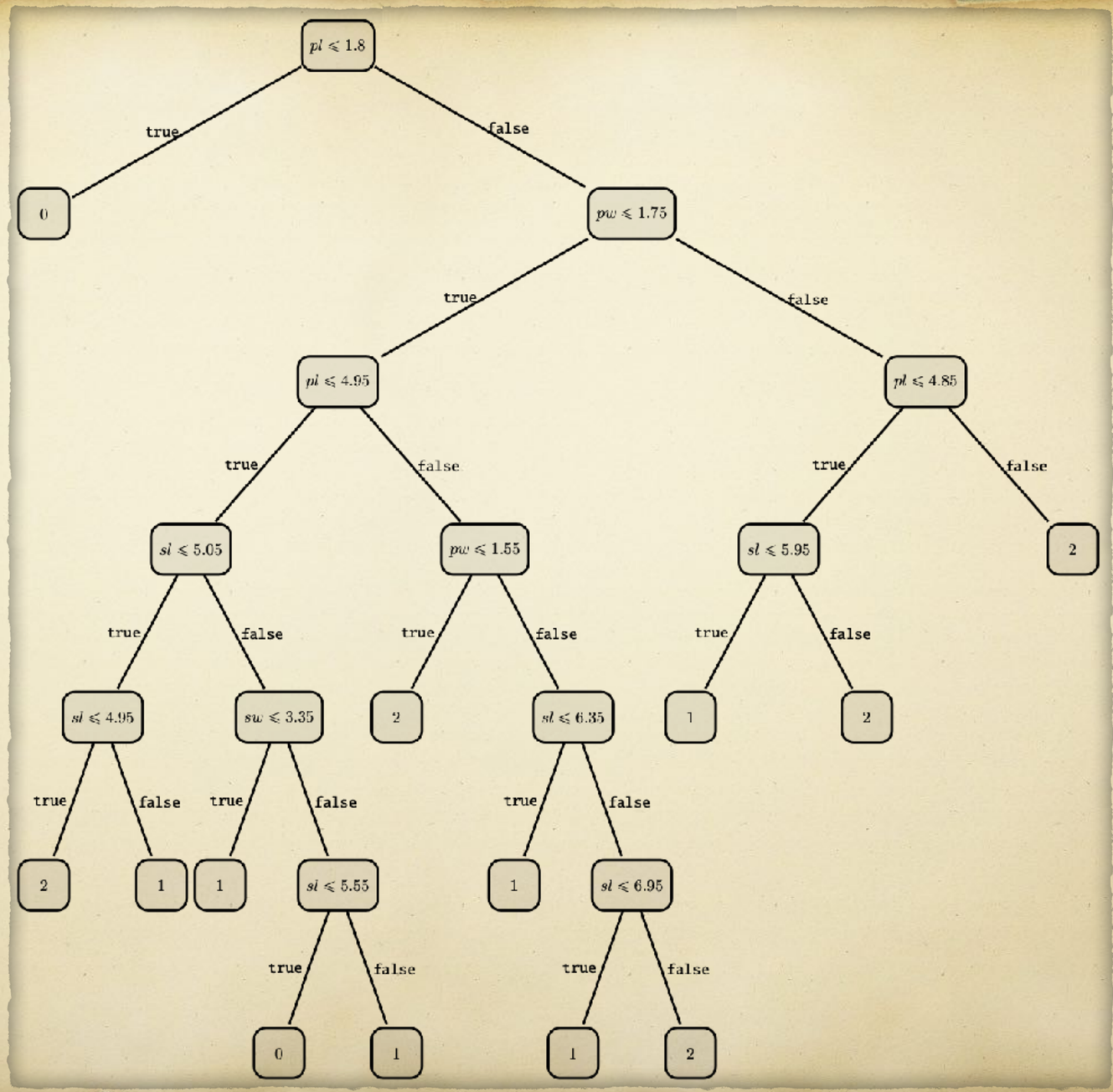
où $Z_{a \leq s} \subseteq Z$ et $Z_{a > s} \subseteq Z$.

Algorithme ID3

- *sl : sepal length*
- *sw : sepal width*
- *pl : petal length*
- *pw : petal width*

$x = (6{,}3; 3{,}3; 5{,}0; 1{,}6)$

- *Iris versicolor*

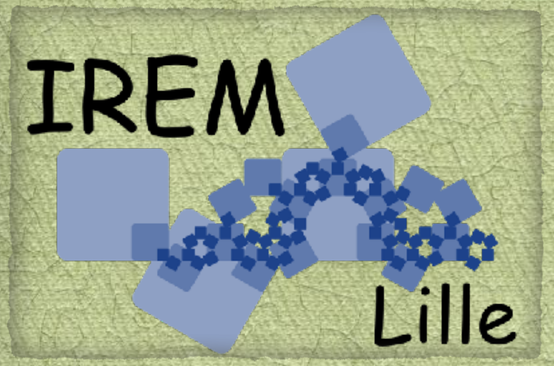


Algorithme ID3

Surapprentissage : un modèle apprend non seulement les structures sous-jacentes des données d'entraînement, mais aussi le bruit et les détails spécifiques de ces données. Il mémorise presque chaque exemple.

- *perfection sur l'ensemble d'apprentissage*
- *problèmes de généralisation*
- *modèle trop complexe, manque de données, absence de régularisation*

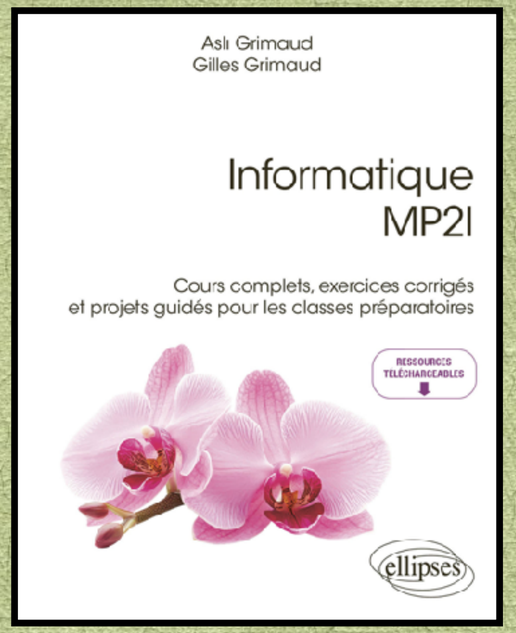
- *hauteur maximale de l'arbre*
- *un sous-ensemble de données contenant un nombre minimal d'individus*
- *l'indice de confiance suffisamment élevé*



Groupe Info

IREM de Lille

Asli Grimaud



ISBN : 9782340103849

IA et Apprentissage Automatique

MERCI

Mars - Avril 2025

asli.grimaud@ac-lille.fr