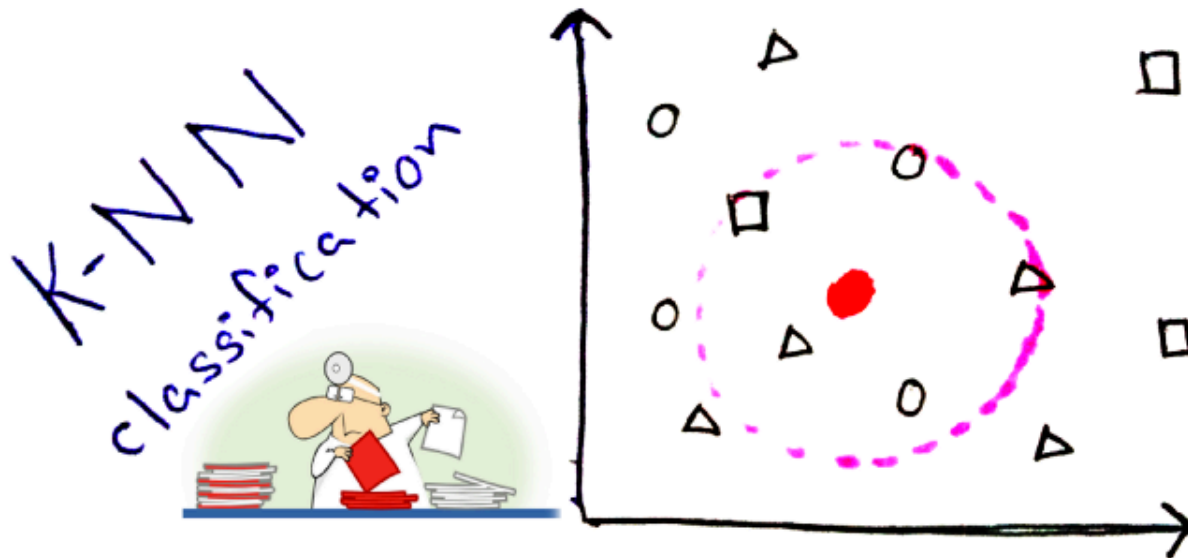
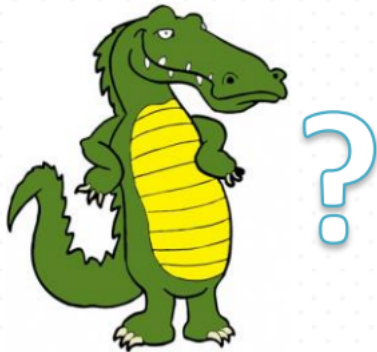
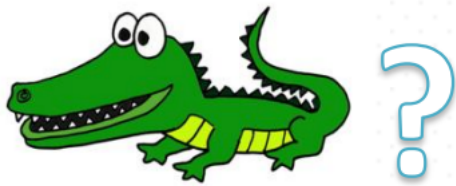


Les K plus proches voisins



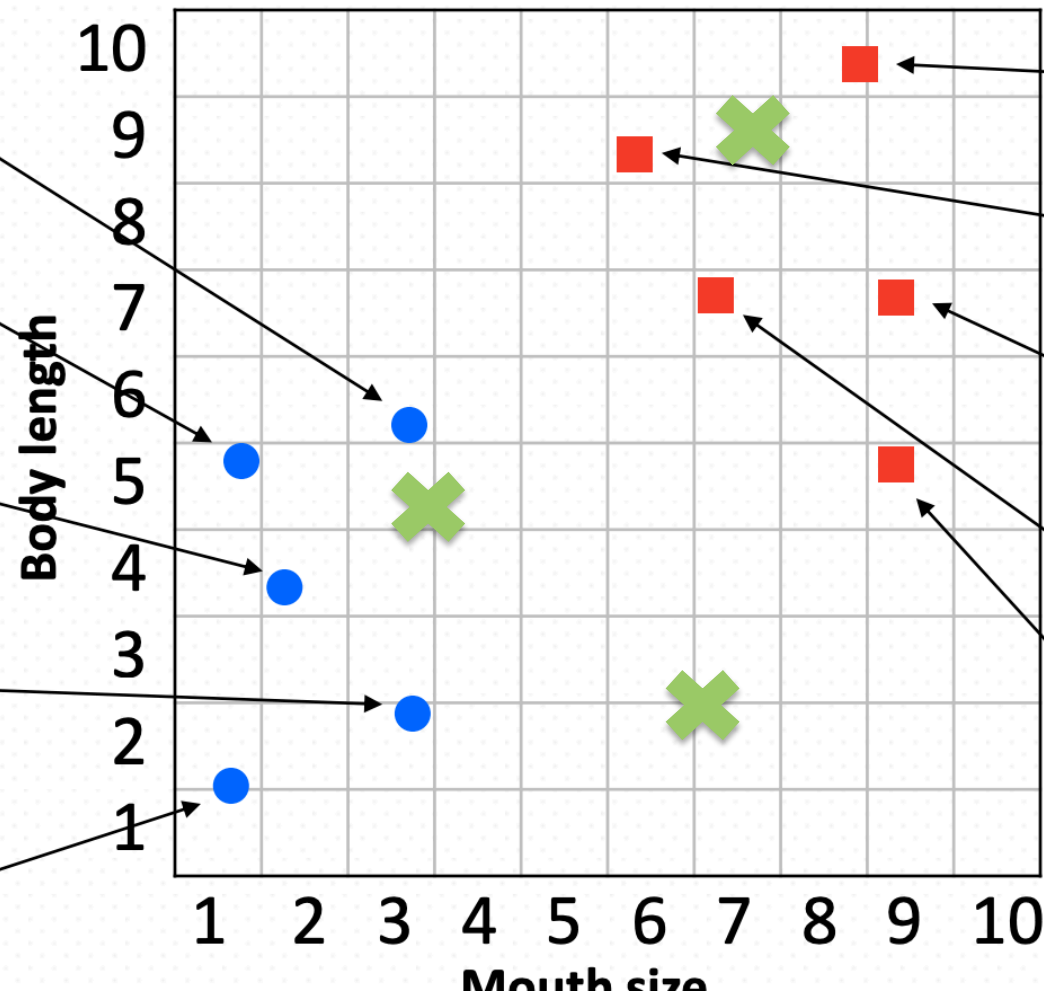
Exemple introductif



Exemple introductif

Alligators

Crocodiles



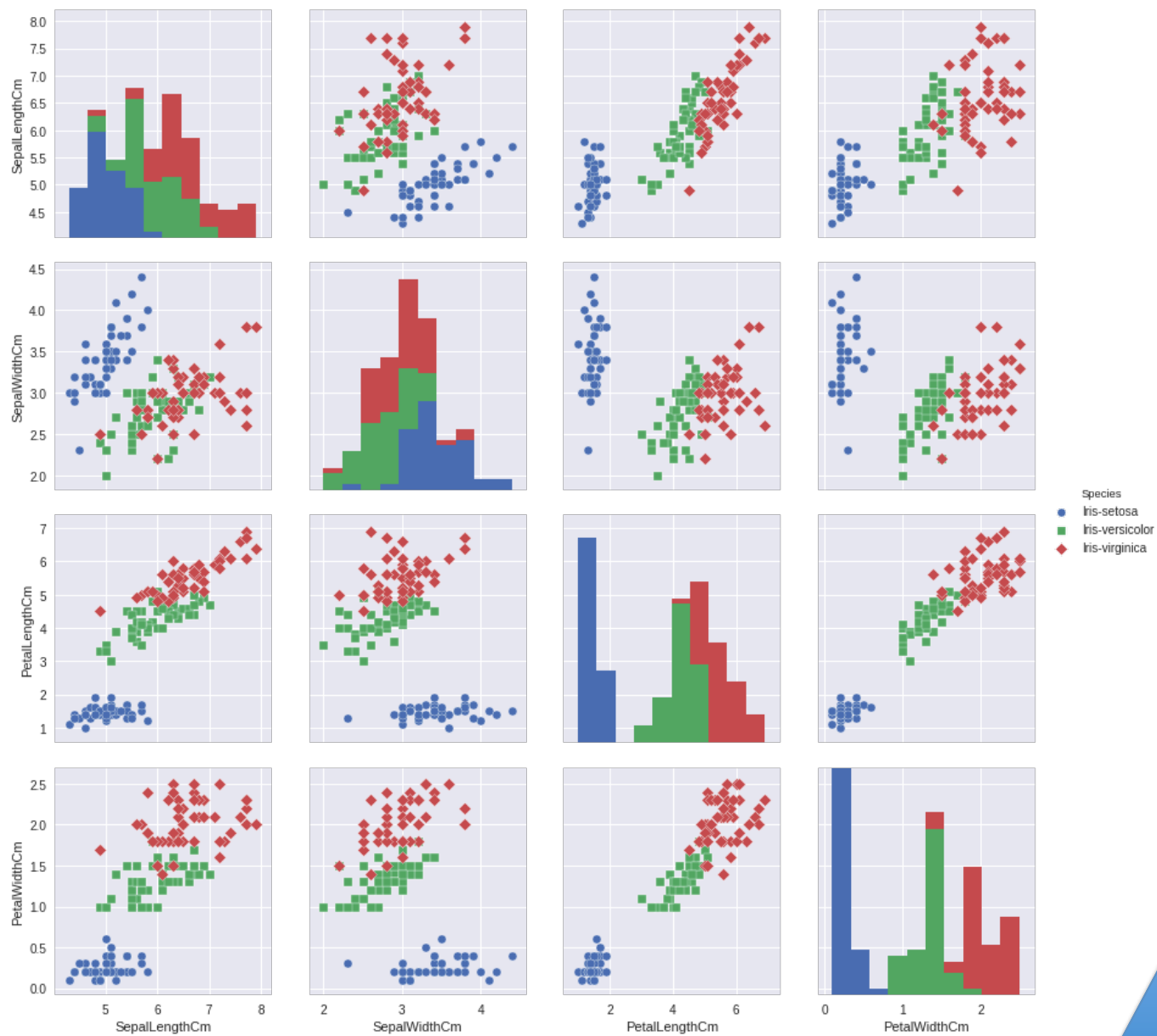
Données plus complexe



	Sepal length	Sepal width	Petal length	Petal width	Type
1	5.1	3.5	1.4	0.2	Iris setosa
2	4.9	3.0	1.4	0.2	Iris setosa
...					
51	7.0	3.2	4.7	1.4	Iris versicolor
52	6.4	3.2	4.5	1.5	Iris versicolor
...					
101	6.3	3.3	6.0	2.5	Iris virginica
102	5.8	2.7	5.1	1.9	Iris virginica
...					



Quelle iris est-ce ?





Pokémons

1. name: The English name of the Pokemon
2. japanese_name: The Original Japanese name of the Pokemon
3. pokedex_number: The entry number of the Pokemon in the National Pokedex
4. percentage_male: The percentage of the species that are male. Blank if the Pokemon is genderless.
5. type1: The Primary Type of the Pokemon
6. type2: The Secondary Type of the Pokemon
7. classification: The Classification of the Pokemon as described by the Sun and Moon Pokedex
8. height_m: Height of the Pokemon in metres
9. weight_kg: The Weight of the Pokemon in kilograms
10. capture_rate: Capture Rate of the Pokemon
11. base_egg_steps: The number of steps required to hatch an egg of the Pokemon
12. abilities: A stringified list of abilities that the Pokemon is capable of having
13. experience_growth: The Experience Growth of the Pokemon
14. base_happiness: Base Happiness of the Pokemon
15. against_?: Eighteen features that denote the amount of damage taken against an attack of a particular type
16. hp: The Base HP of the Pokemon
17. attack: The Base Attack of the Pokemon
18. defense: The Base Defense of the Pokemon
19. sp_attack: The Base Special Attack of the Pokemon
20. sp_defense: The Base Special Defense of the Pokemon
21. speed: The Base Speed of the Pokemon
22. generation: The numbered generation which the Pokemon was first introduced
23. is_legendary: Denotes if the Pokemon is legendary.



Est il légendaire ??

Type de variable

- Variables qualitatives ou catégorielles.
 - Ex.: couleur des yeux, type d'engrais, méthode d'enseignement, catégorie grammaticale...
 - Deux types: *nominal* ou *ordinal*.
 - On appelle “niveaux” ou “modalités” les valeurs que peuvent prendre une variable qualitative.
- Variables quantitatives ou numériques
 - Elles peuvent être *discrètes* (à valeurs dans les entiers; exemple: comptage) ou *continues* (à valeurs dans les réels).
 - Deux types: *intervalle* (seule la différence à un sens, ex: heure) ou *ratio* (le rapport à un sens, ex: vitesse).
 - Ex.: taille, production en maïs, temps de réaction...
- Les procédures statistiques diffèrent en fonction des types des variables.

Variables

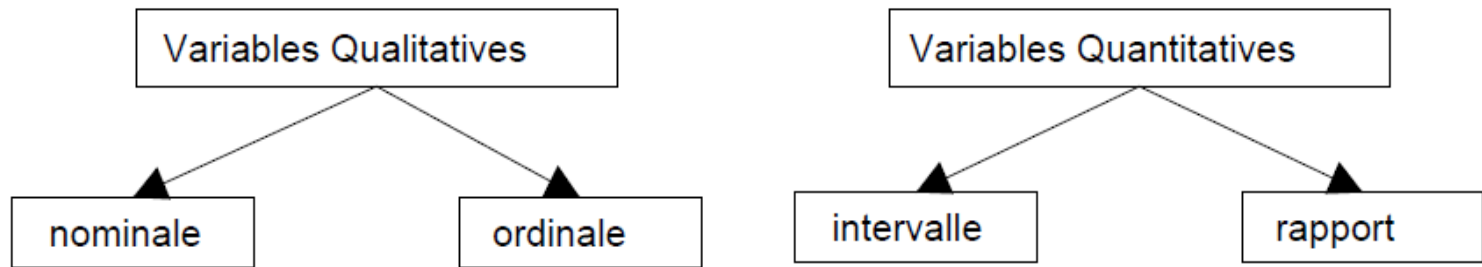
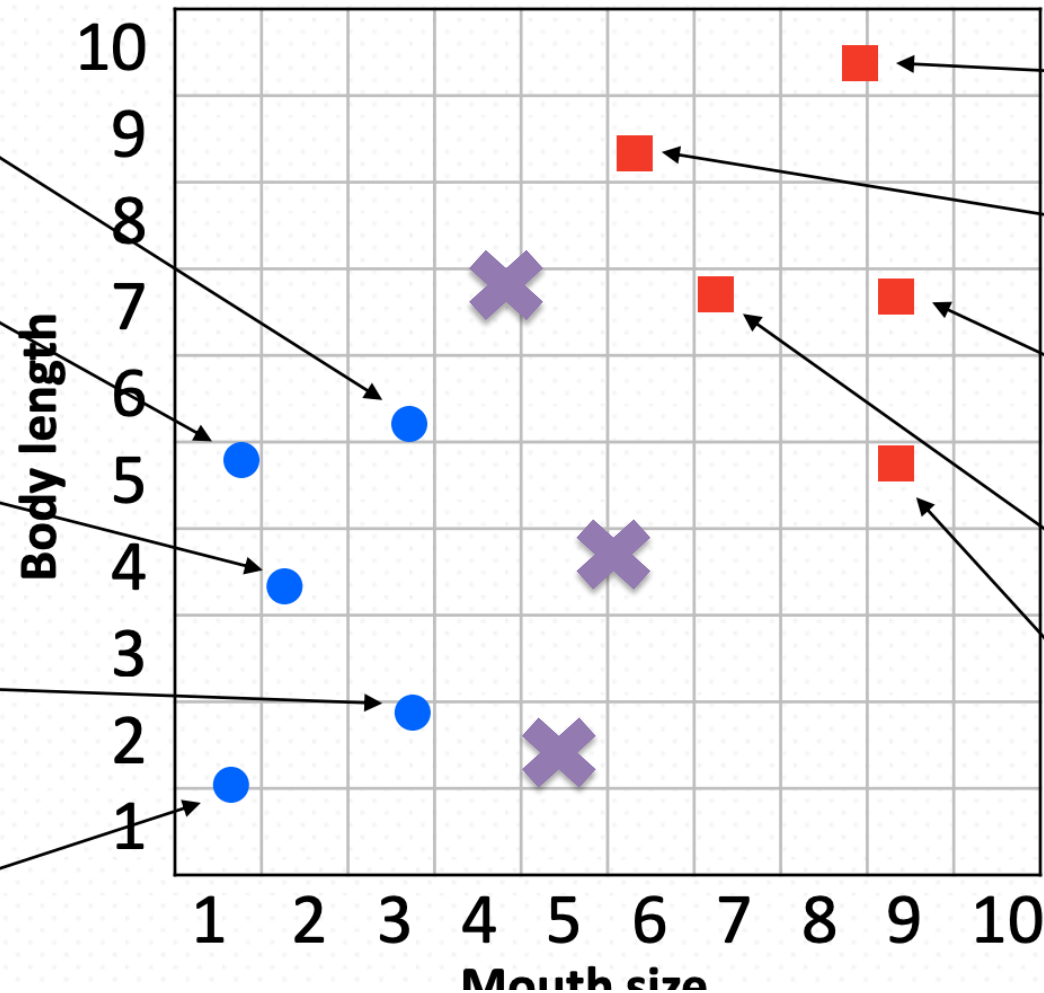


Fig. 1.1: Les deux grandes classes de variables.

Exemple introductif

Alligators


Crocodiles



Classification

- Elle permet de **prédire** si un élément est membre d'un groupe ou d'une catégorie donnée.
- **Classes**
 - Identification de groupes avec des profils particuliers
 - Possibilité de décider de l'appartenance d'une entité à une classe
- Caractéristiques
 - **Apprentissage supervisé** : classes connues à l'avance
 - Pb : qualité de la classification (taux d'erreur)
 - Ex : établir un diagnostic (si erreur !!!)

Classification - Applications

- Comprendre les critères prépondérants pour l'achat d'un produit ou d'un service
 - Isoler les critères explicatifs d'un comportement d'achat
 - Analyse de risque: détecter les facteurs prédisant un comportement de non paiement
 - Détecter les causes de réclamation
- 
- A blue triangle graphic is located in the bottom right corner of the slide, pointing towards the top right.

Processus à deux étapes



- Etape 1 :
- Construction du modèle à partir de l'ensemble d'apprentissage (training set)
- Etape 2 :
- Utilisation du modèle : tester la précision du modèle et l'utiliser dans la classification de nouvelles données

Construction du modèle

- Chaque **instance** est supposée appartenir à une classe prédéfinie
- La classe d'une instance est déterminée par l'attribut "**classe**"
- L'ensemble des instances d'apprentissage est utilisé dans la construction du modèle
- Le **modèle** est représenté par des règles de classification, arbres de décision, formules mathématiques, ...



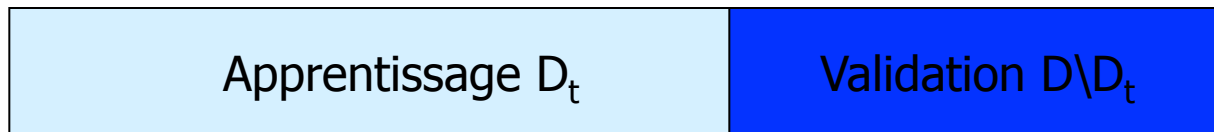
Utilisation du modèle



- Classification de nouvelles instances ou instances inconnues
- Estimer le taux d'erreur du modèle
 - la classe connue d'une instance test est comparée avec le résultat du modèle
 - Taux d'erreur = pourcentage de tests incorrectement classés par le modèle

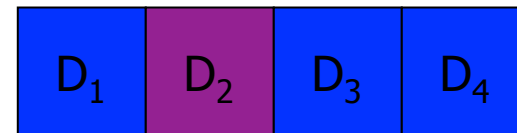
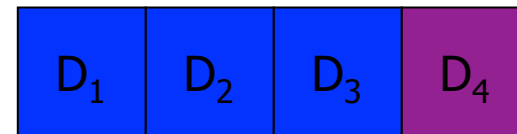
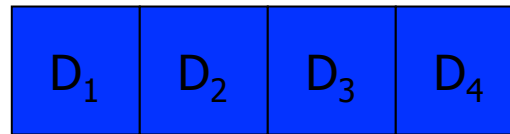
Validation de la Classification (accuracy)

- Estimation des taux d'erreurs :
- Partitionnement : apprentissage et test (ensemble de données important)
 - Utiliser 2 ensembles indépendents, e.g., ensemble d'apprentissage (2/3), ensemble test (1/3)



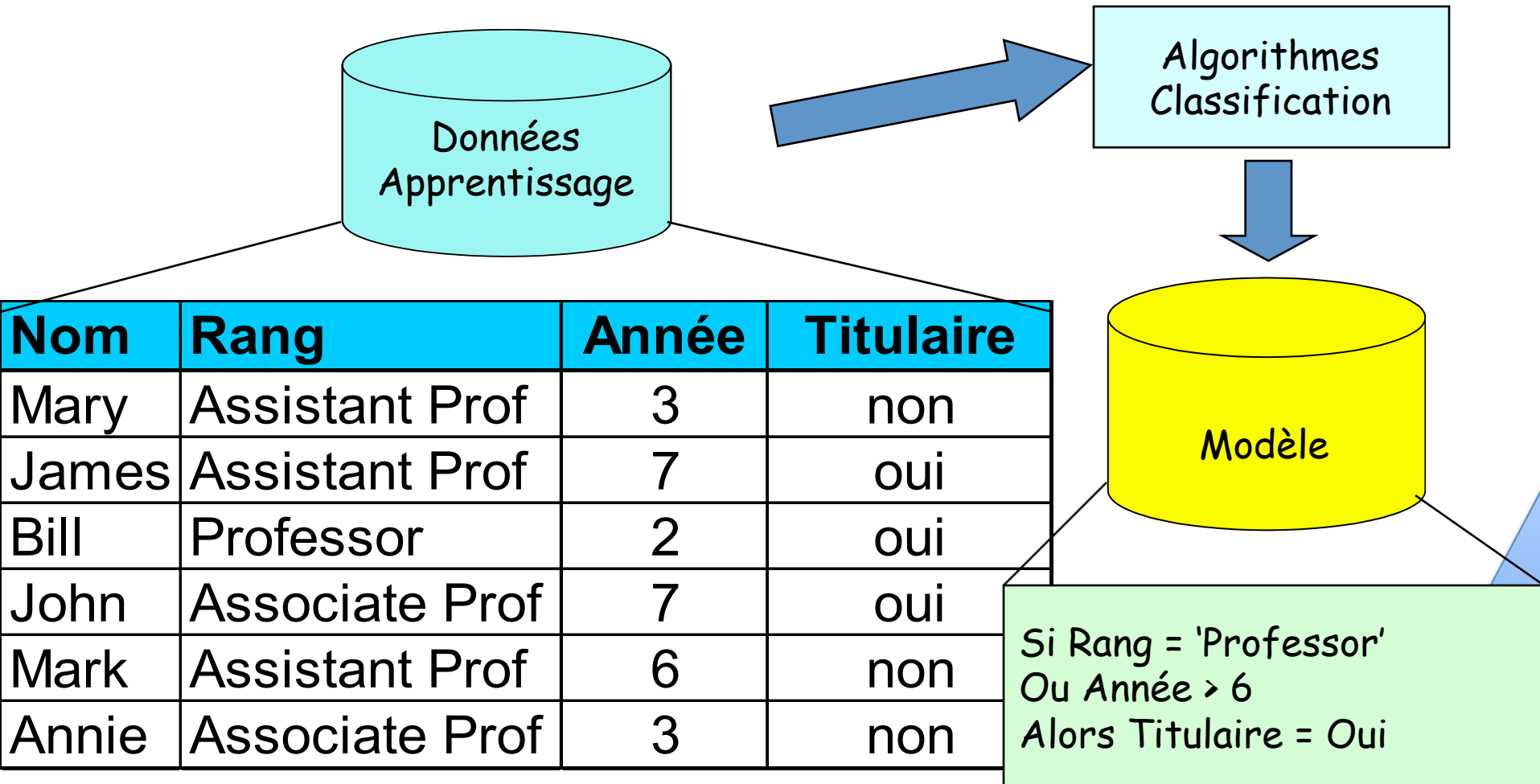
Validation de la Classification (accuracy)

- Validation croisée (ensemble de données modéré)
 - Diviser les données en k sous-ensembles
 - Utiliser $k-1$ sous-ensembles comme données d'apprentissage et un sous-ensemble comme données test

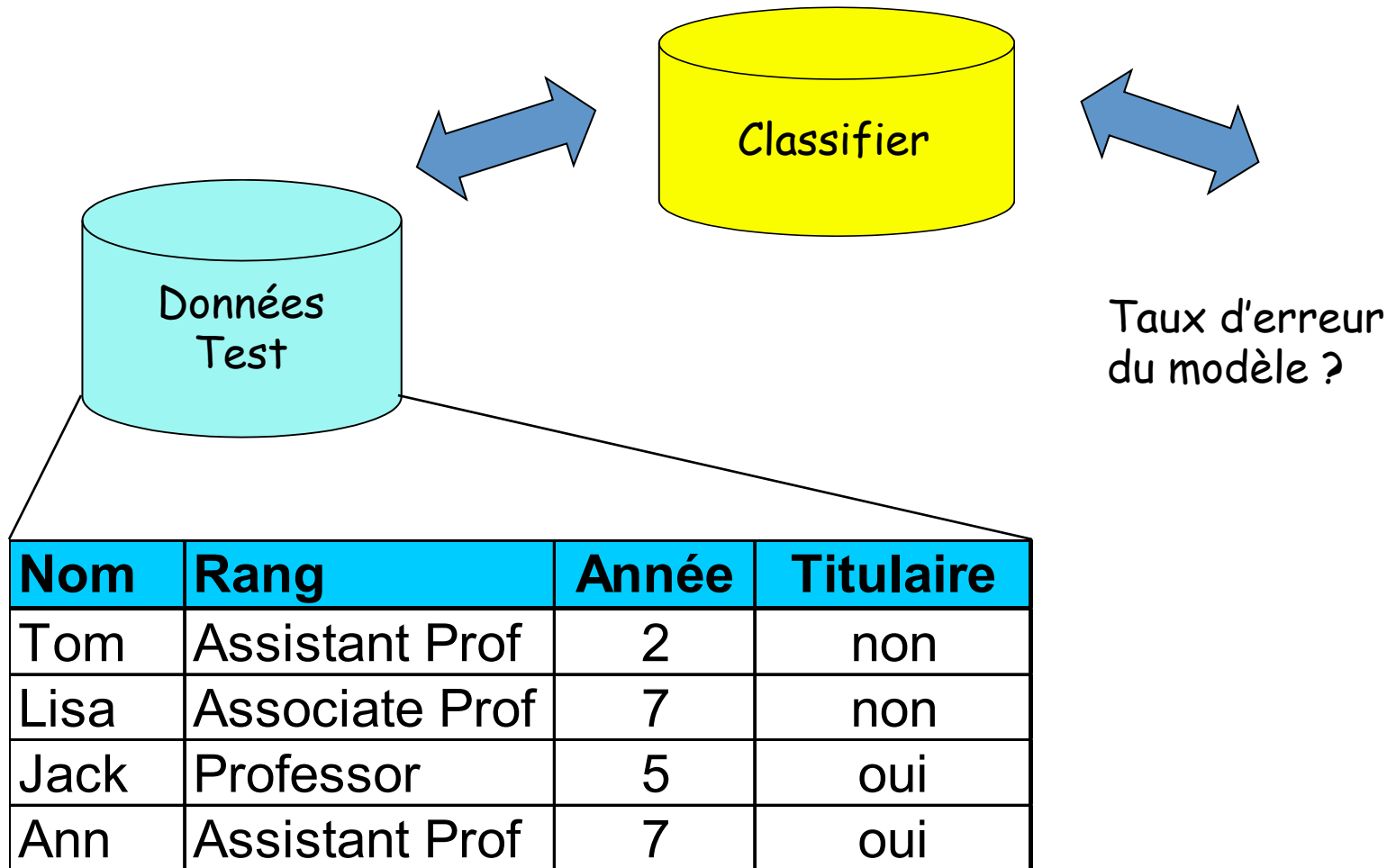


- Bootstrapping : n instances test aléatoires (ensemble de données réduit)

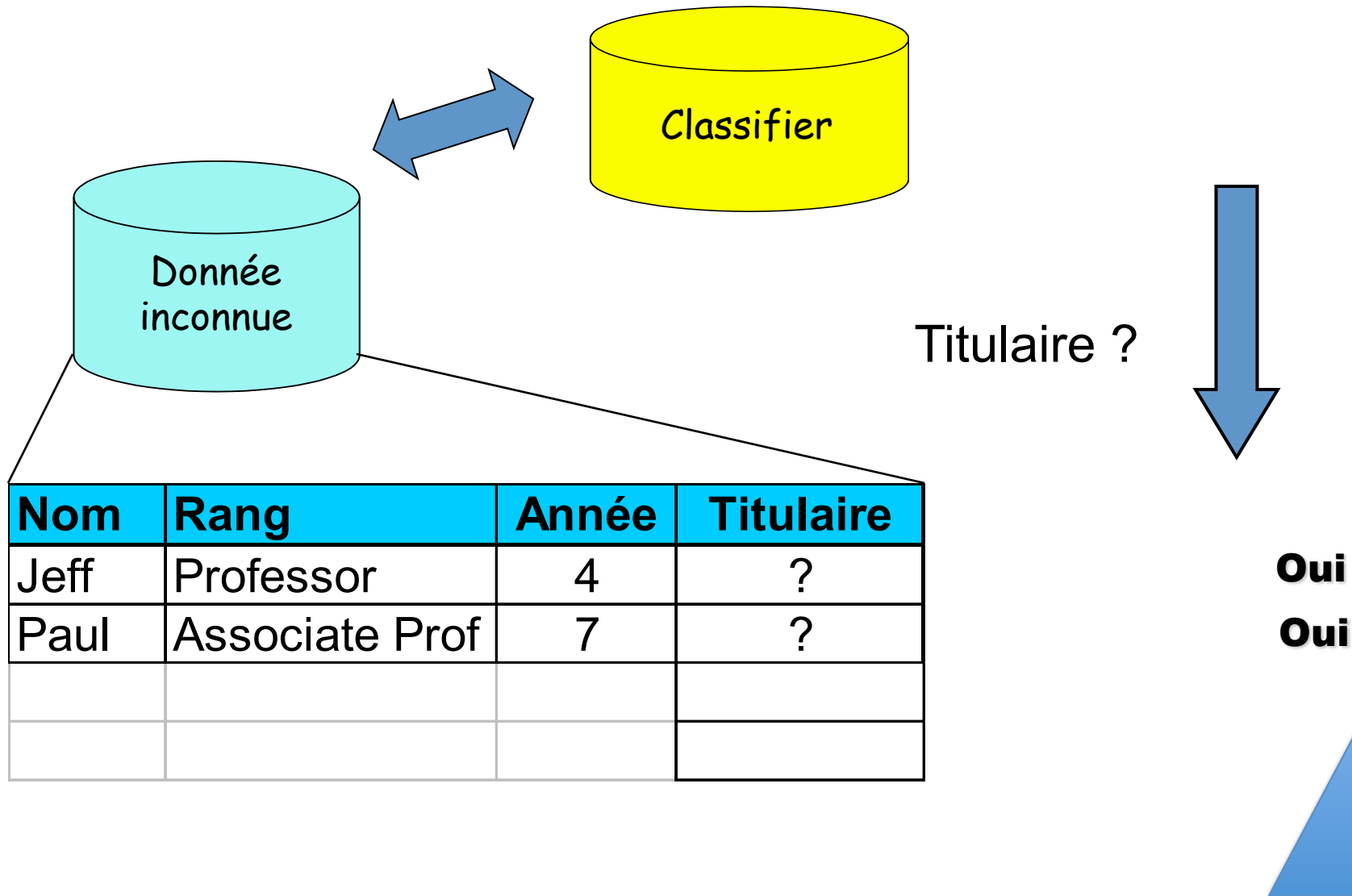
Exemple : Construction du modèle




Exemple : Utilisation du modèle



Exemple : Utilisation du modèle



Evaluation des méthodes de classification

- Taux d'erreur (Accuracy)
 - Temps d'exécution (construction, utilisation)
 - Robustesse (bruit, données manquantes,...)
 - Extensibilité
 - Interprétabilité
 - Simplicité
- 
- A blue triangle graphic is located in the bottom right corner of the slide, pointing towards the top right.

Méthodes de Classification

- Méthode K-NN (plus proche voisin)
- Arbres de décision
- Réseaux de neurones
- Classification bayésienne
- Caractéristiques
 - Apprentissage supervisé (classes connues)

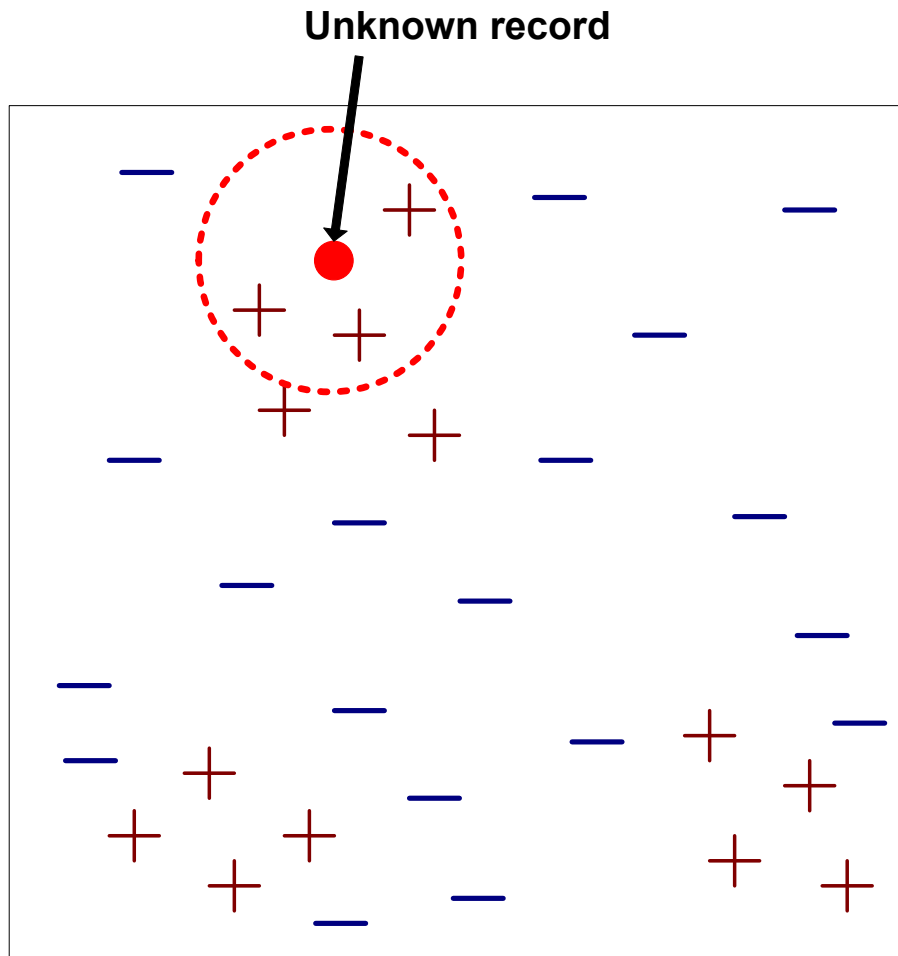
Dis moi qui sont tes amis, je te dirais qui tu es ...

KNN

Méthode des plus proches voisins

- Méthode dédiée à la classification (k-NN : nearest Neighbors).
- **Méthode de raisonnement** à partir de cas : prendre des décisions en recherchant un ou des cas similaires déjà résolus.
- **Pas d'étape d'apprentissage** : construction d'un modèle à partir d'un échantillon d'apprentissage (réseaux de neurones, arbres de décision, ...).
- Modèle = échantillon d'apprentissage + fonction de distance + fonction de choix de la classe en fonction des classes des voisins les plus proches.

Nearest-Neighbor



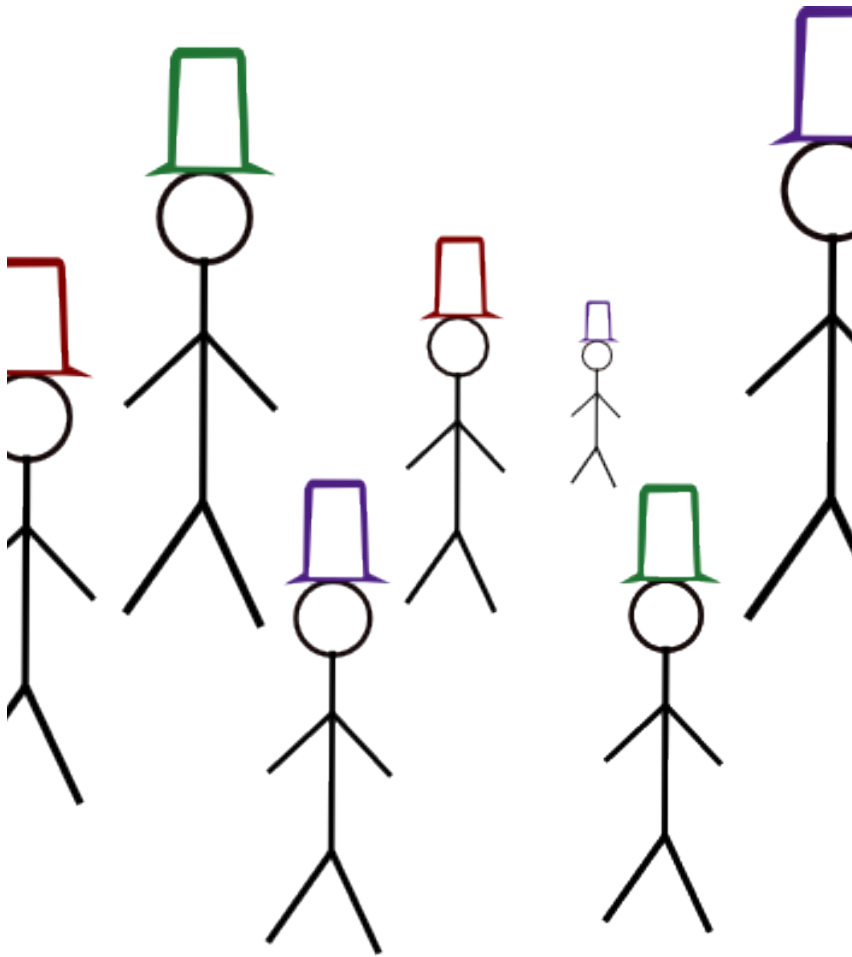
Algorithme kNN (K-nearest neighbors)

- Objectif : affecter une classe à une nouvelle instance
- donnée : un échantillon de m enregistrements classés $(x, c(x))$
- entrée : un enregistrement y
 - 1. Déterminer les k plus proches enregistrements de y
 - 2. combiner les classes de ces k exemples en une classe c
- sortie : la classe de y est $c(y)=c$

Qu'est ce qu'être proche ?

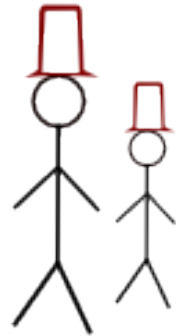
- Vocabulaire
- Mesure de dissimilarité (DM) : plus la mesure est faible plus les points sont similaires (\sim distance)
- Mesure de similarité (SM) : plus la mesure est grande, plus les points sont similaires
- $DM = borne - SM$

Mesure de la similarité

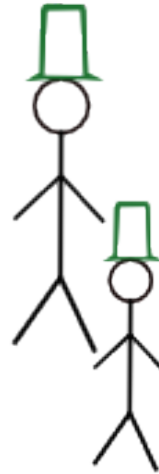


- Il n'y a pas de définition unique de la similarité entre objets
 - Différentes mesures de distances $d(x,y)$
- La définition de la similarité entre objets dépend de :
 - Le type des données considérées
 - Le type de similarité recherchée

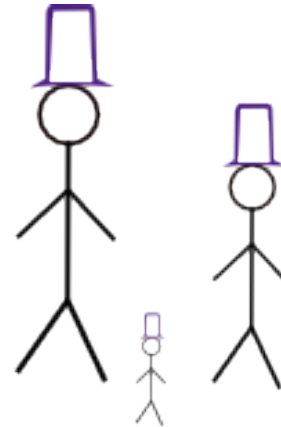
Mesure de similarité



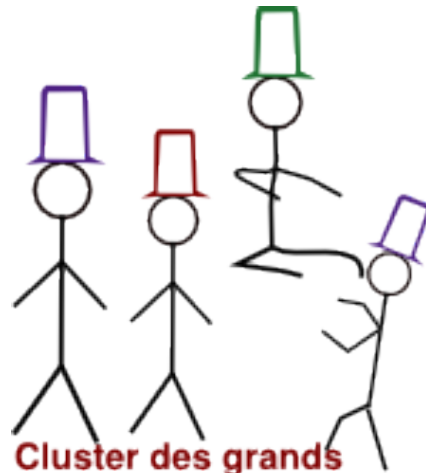
Cluster des rouges



Cluster des verts



Cluster des violets



Cluster des grands
(par la taille...)



Cluster des petits
et qui le vivent bien!

Distance

- Propriétés d'une distance :
 1. $d(x, y) \geq 0$
 2. $d(x, y) = 0$ iff $x = y$
 3. $d(x, y) = d(y, x)$
 4. $d(x, z) \leq d(x, y) + d(y, z)$
- Similarité : vérifie $s(i, j) = s(j, i)$, $s(i, j) \geq 0$;
 $s(i, i) \geq s(i, j)$

Distance – Données numériques

- Combiner les distances : Soient $x=(x_1,...,x_n)$ et $y=(y_1, ...,y_n)$
- Exemples numériques :

- Distance euclidienne :
$$d(x,y)=\sqrt{\sum_{i=1}^n (x_i-y_i)^2}$$

- Distance de Manhattan :
$$d(x,y)=\sum_{i=1}^n |x_i-y_i|$$

- Distance de Minkowski :
$$d(x,y)=\sqrt[q]{\sum_{i=1}^n |x_i-y_i|^q}$$

- $q=1$: distance de Manhattan.
- $q=2$: distance euclidienne

Distance données énumératives

- Champs discrets :
 - Données binaires : $d(0,0)=d(1,1)=0$,
 $d(0,1)=d(1,0)=1$
 - Donnée énumératives : distance nulle si les valeurs sont égales et 1 sinon.
 - Donnée énumératives ordonnées : idem. On peut définir une distance utilisant la relation d'ordre.

Distance – Données énumératives

- Généralisation des variables binaires, avec plus de 2 états, e.g., rouge, jaune, bleu, vert
- Méthode 1: correspondance simple
 - m : # de correspondances, p : # total de variables

$$d(i, j) = \frac{p - m}{p}$$

Variables Ordinales

- Une variable ordinale peut être discrète ou continue
- L'ordre peut être important, ex: classement
- Peuvent être traitées comme les variables intervalles
 - remplacer x_{if} par son rang $r_{if} \in \{1, \dots, M_f\}$
 - Remplacer le rang de chaque variable par une valeur dans $[0, 1]$ en remplaçant la variable f dans l'objet I par

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

- Utiliser une distance pour calculer la similarité

Variables Ordinales

- Formulaire de satisfaction
 - Att1 : Très satisfait, Satisfait, Neutre, Mécontent
 - Donc 4 valeurs, dont les rangs sont 1,2,3,4

Devient :

$(1-1)/(4-1), (2-1)/(4-1), (3-1)/(4-1), (4-1)/(4-1)$

Donc Valeurs : 0, $1/3$, $2/3$, $3/3$ (1)

Données mixtes

- Soit - transformation des variables numériques en variables catégorielles
- (découpage en intervalles -> pris comme modalités) $d^2(i, j) = \frac{1}{p} \sum_{k=1}^p \delta_k(i, j)$
- → distance/similarité sur tableau disjonctif
- transformation des variables catégorielles en variables numériques
- - utilisation de mesures "mixtes" » Normaliser !!!!
- Principe :



Données mixtes

- Normalisation d'un attribut

$$a_i = \frac{v_i - \min v_i}{\max v_i - \min v_i}$$

$$a_i = \frac{v_i - Avg(v_i)}{StDev(v_i)}$$

- Ou directement dans le calcul de la distance
Pour une variable numérique :

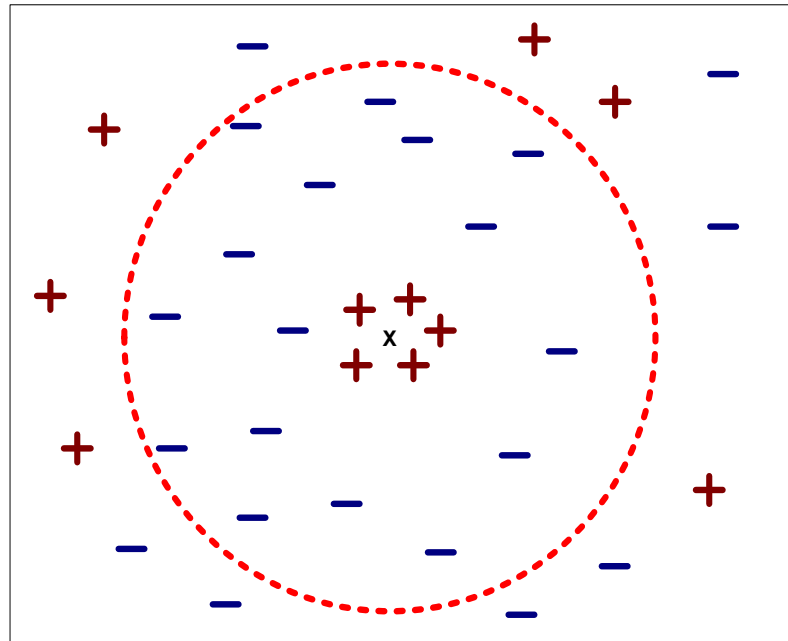
$$\delta_k(i, j) = \frac{(x_{ik} - x_{jk})}{(\max - \min)}$$

Distance – Données mixtes

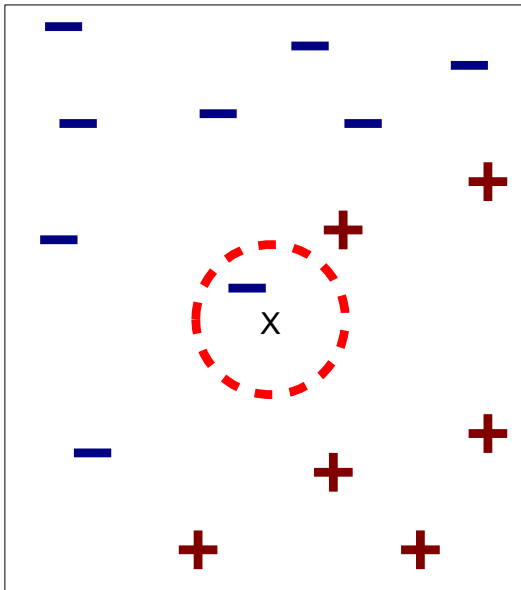
- Exemple : (Age, Propriétaire résidence principale, montant des mensualités en cours)
- $x=(30,1,1000)$, $y=(40,0,2200)$, $z=(45,1,4000)$
- $d(x,y)=\text{sqrt}((10/15)^2 + 1^2 + (1200/3000)^2) = 1.27$
- $d(x,z)=\text{sqrt}((15/15)^2 + 0^2 + (3000/3000)^2) = 1.41$
- $d(y,z)=\text{sqrt}((5/15)^2 + 1^2 + (1800/3000)^2) = 1.21$
- plus proche voisin de $x = y$
- Distances normalisées.
- Sommation : $d(x,y)=d_1(x_1,y_1) + \dots + d_n(x_n,y_n)$

Classification par plus proche voisin

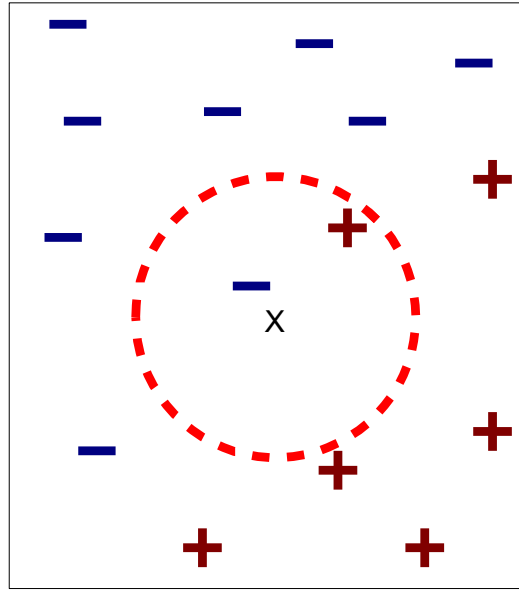
- Choisir k :
 - Si k est trop petit, knn sera sensible au bruit
 - Si k est trop grand, le voisinage pourrait inclure des points d'autres classes



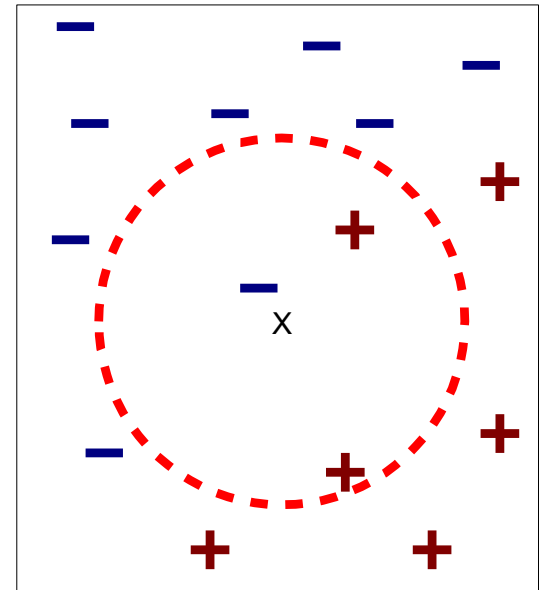
Definition de Plus Proche Voisin



(a) 1-nearest neighbor



(b) 2-nearest neighbor



(c) 3-nearest neighbor

Algorithme kNN : sélection de la classe

- Basé sur l'apprentissage par analogie
- Basée sur une notion de distance et Similarité
- **Solution simple** : rechercher le cas le plus proche et prendre la même décision (Méthode 1-NN).
- **Combinaison des k classes** :
 - Heuristique : $k = \text{nombre d'attributs} + 1$
 - Vote majoritaire : prendre la classe majoritaire.
 - Vote majoritaire pondéré : chaque classe est pondérée. Le poids de $c(x_i)$ est inversement proportionnel à la distance $d(y, x_i)$.
- **Confiance** : Définir une confiance dans la classe attribuée = rapport entre les votes gagnants et le total des votes.

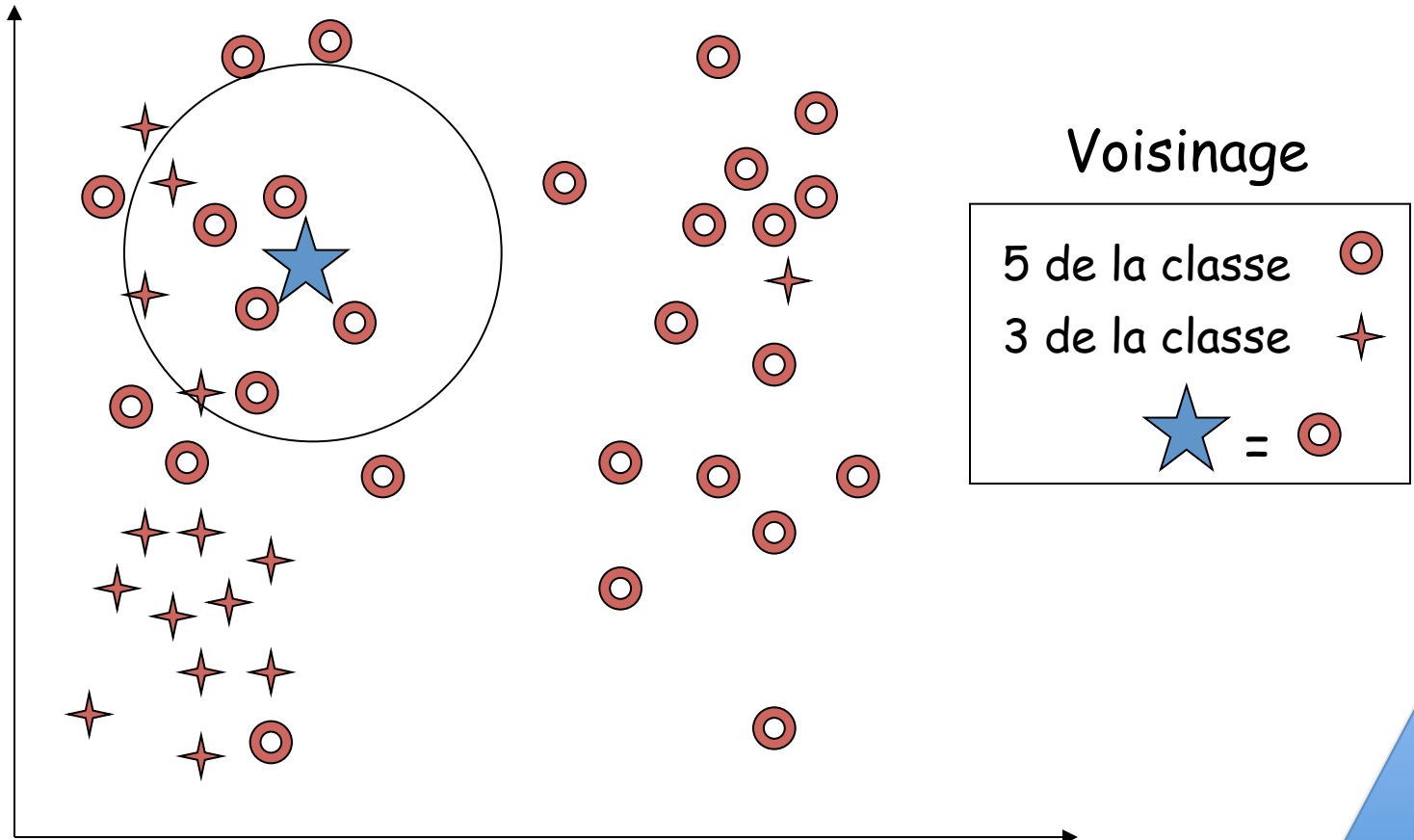
Vote pondéré

- Rectangulaire (loi uniforme): $\frac{1}{2}I(|d| \leq 1)$
- Triangulaire : $(1 - |d|)I(|d| \leq 1)$
- Epanechnikov : $\frac{3}{4}(1 - d^2)I(|d| \leq 1)$
- Bi-poids : $\frac{15}{16}(1 - d^2)^2 I(|d| \leq 1)$
- Tri-poids : $\frac{35}{32}(1 - d^2)^3 I(|d| \leq 1)$
- Cosine $\frac{\pi}{4} \cos\left(\frac{\pi}{2} d\right) I(|d| \leq 1)$
- *gaussien* : $\frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2} d^2)$
- Inverse $\frac{1}{|d|}$

Classiquement weight factor, $w = 1/d^2$

Exemple

8 plus proches voisins



Forces et faiblesses

- Les attributs ont le même poids
 - centrer et réduire pour éviter les biais
 - certains peuvent être moins classant que d'autres
- Apprentissage paresseux
 - rien n'est préparé avant le classement
 - tous les calculs sont fait lors du classement
 - nécessité de technique d'indexation pour large BD
- Calcul du score d'une classe
 - peut changer les résultats; variantes possibles