

## Cours 4

# Introduction à l'Apprentissage Automatique (suite)

# Plan du cours

## 1. Généralités sur l'apprentissage automatique

- Problèmes/tâches
- Apprendre sur quelles données ? Représentations ?
- Modèles de représentation des connaissances

## 2. Classer sans apprendre

- Apprentissage Bayésien naïf
- Notions d'erreur

## 3. Classer sans apprendre mais en réglant des paramètres

- Classifieur « plus proches voisins »
- Validation croisée

# Rappels : classification supervisée

On dispose d'un ensemble de données d'apprentissage :  $S = \{(x_i, u_i)\}_{1 \dots m}$

On cherche à induire un modèle/une méthodologie pour **prédir la classe  $u^*$**  d'une nouvelle donnée  $y$ .

Le **classifieur de Bayes** :

- est adapté aux données qualitatives/symboliques (seulement)
- n'est pas paramétrable
- peut s'avérer performant dans certains cas

→ quid des données quantitatives/numériques ?

RI	Na	Mg	Al	Si	K	Ca	Ba	Fe	classe
1.52177	13.2	3.68	1.15	72.75	0.54	8.52	0	0	2
1.51872	12.93	3.66	1.56	72.51	0.58	8.55	0	0.12	2
1.51667	12.94	3.61	1.26	72.75	0.56	8.6	0	0	2
1.52081	13.78	2.28	1.43	71.99	0.49	9.85	0	0.17	2
1.52068	13.55	2.09	1.67	72.18	0.53	9.57	0.27	0.17	2
1.51769	13.65	3.66	1.11	72.77	0.11	8.6	0	0	2

# Classifieur « plus proches voisins »

**Principe** : « Dis-moi qui sont tes amis, je te dirais qui tu es »



**tes voisins**

**Traduction** : Cherchons (dans  $S$ ) ceux qui te ressemblent, ils nous diront quelle est ta classe

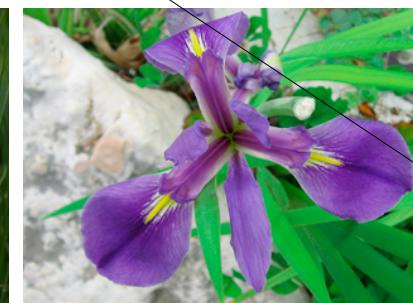
# Exemple

## Classe des Iris Setosa



...

## Classe des Iris Virginica



ressemble à

...

Classe Setosa ←



# Classifieur « plus proches voisins »

**Principe** : « Dis-moi qui sont tes amis, je te dirais qui tu es »



**Traduction** : Cherchons (dans  $S$ ) tes voisins, ils nous diront quelle est ta classe

**Questions/problèmes** à lever :

- Comment « mesurer » la ressemblance entre deux données ? (trouver les voisins)
- Combien de voisins considérer ?
- Comment décider lorsque les voisins ont des étiquettes différentes ?

# Classifieur « plus proches voisins »

**Question 1** : comment mesurer la ressemblance entre deux données ?



↔  
ressemblance ?



# Classifieur « plus proches voisins »

**Question 1 :** comment mesurer la ressemblance entre deux données ?

	Largeur des pétales	Longueur des pétales	Largeur des sépales	Longueur des sépales		Largeur des pétales	Longueur des pétales	Largeur des sépales	Longueur des sépales
x	3.7	7.5	2.1	5.6	y	4.2	6.9	1.1	5.5



ressemblance ?



# Classifieur « plus proches voisins »

**Question 1** : comment mesurer la ressemblance entre deux données ?

→ Calculer une distance/similarité/proximité entre deux données  $x$  et  $y$

Exemples de mesures de distances :

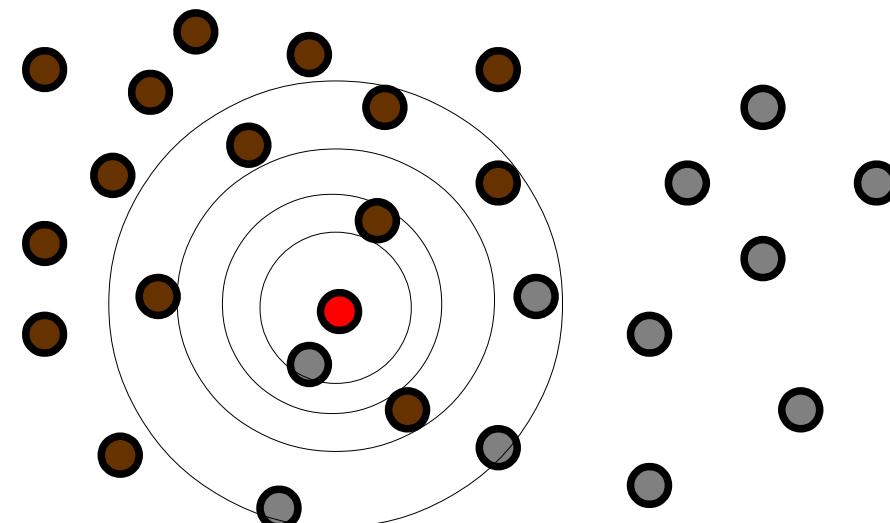
- Distance Euclidienne : 
$$d(x, y) = \left( \sum_{i=1}^n (x_i - y_i)^2 \right)^{1/2}$$
- Distance de Manhattan : 
$$d(x, y) = \sum_{i=1}^n |x_i - y_i|$$
- Distance de Minkowski : 
$$d(x, y) = \left( \sum_{i=1}^n (x_i - y_i)^p \right)^{1/p}$$
- Distance de Tanimoto : 
$$d(s_1, s_2) = \frac{|s_1| + |s_2| - 2|s_1 \cap s_2|}{|s_1| + |s_2| - |s_1 \cap s_2|}$$

# Classifieur « plus proches voisins »

**Question 2** : combien de voisins considérer ?

→ Paramètre à ajuster :  $k$ -PPV (Plus Proches Voisins)

$k$	Voisinage
1	○
2	○○
3	○○○
...	...
10	○○○○○○○○○○



- $k$  petit : fiabilité de la décision mais sensibilité aux données « bruitées »
  - $k$  grand : moins de fiabilité mais plus robuste aux données « bruitées »
  -
- calculer l'**erreur d'apprentissage** et/ou **de généralisation** pour différentes valeurs de  $k$

# Classifieur « plus proches voisins »

## Question 3 : décision finale ?

→ choix de la **classe majoritaire** parmi les  $k$  plus proches voisins

$k$	Voisinage	
1	○	→ ○
2	○ ●	→ ?
3	○ ● ●	→ ●
...	...	
10	○ ● ● ● ● ● ● ● ● ●	→ ●

# Classifieur « plus proches voisins »

## Discussion

- **Interprétation** des prédictions : la classe prédite peut être expliquée en montrant le(s) plus proche(s) voisin(s) ayant induit cette décision (mais pas vraiment de modèle appris).
- La méthode peut s'appliquer dès qu'il est possible de définir une **distance** sur des paires d'individus.
- Tous les calculs doivent être effectués lors de l'étape de prédiction (pas de modèle) → d'autant plus **coûteux** que :
  - $S$  est grand
  - Que la dimension ( $n$ ) de l'espace de représentation est grand

# Plan du cours

## 1. Généralités sur l'apprentissage automatique

- Problèmes/tâches
- Apprendre sur quelles données ? Représentations ?
- Modèles de représentation des connaissances

## 2. Classer sans apprendre

- Apprentissage Bayésien naïf
- Notions d'erreur

## 3. Classer sans apprendre mais en réglant des paramètres

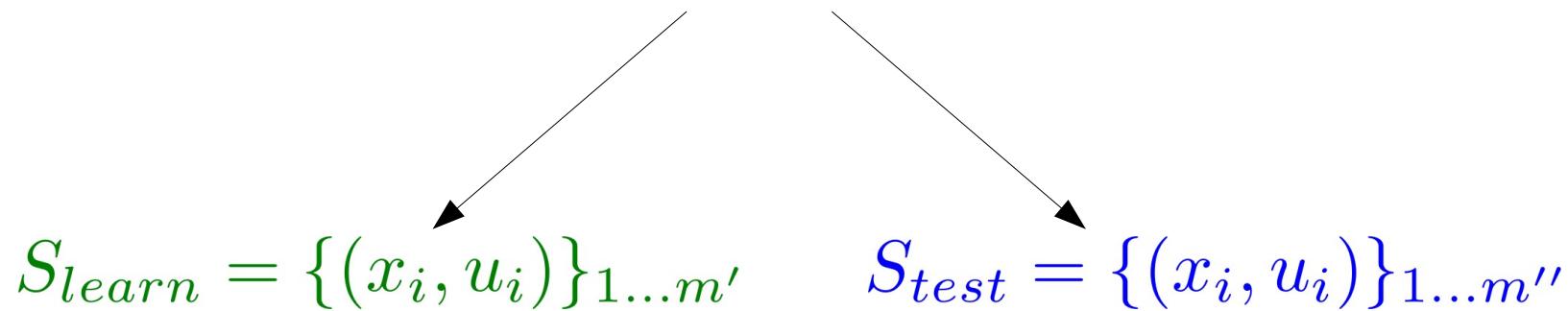
- Classifieur « plus proches voisins »
- Validation croisée

# Evaluation de l'apprentissage

**RAPPEL**

Supposons que l'on ait à notre disposition :

- des données étiquetées  $S = \{(x_i, u_i)\}_{1\dots m}$  (données d'apprentissage)



Comment évaluer un classifieur ?

- On oublie l'ensemble de test
- On coupe **COMMENT ?**  $S$  en 2 sous-ensembles : **apprentissage** vs. **Test**
- On classe les données **d'apprentissage** en utilisant les informations de l'ensemble **d'apprentissage** → **erreur d'apprentissage**
- On classe les données de **test** en utilisant les informations de l'ensemble **d'apprentissage** → **erreur de généralisation**

# Evaluation de l'apprentissage : par validation croisée

$$S = \{(x_i, u_i)\}_{1 \dots m}$$

50%

50%

Evaluation 1

$$S_{learn} = \{(x_i, u_i)\}_{1 \dots m'} \quad S_{test} = \{(x_i, u_i)\}_{1 \dots m''}$$

Evaluation 2

$$S_{learn} = \{(x_i, u_i)\}_{1 \dots m'} \quad S_{test} = \{(x_i, u_i)\}_{1 \dots m''}$$

→ **Erreur de généralisation** = moyenne des deux évaluations (croisées)

**2-fold cross validation**

# Evaluation de l'apprentissage : par validation croisée

$$S = \{(x_i, u_i)\}_{1 \dots m}$$

50%

Evaluation 1

$$S_{learn} = \{(x_i, u_i)\}$$

50%

$$= \{(x_i, u_i)\}_{1 \dots m''}$$

Evaluation 2

$$S_{learn} = \{(x_i, u_i)\}_{1 \dots m}$$

$$= \{(x_i, u_i)\}_{1 \dots m''}$$



→ **Erreur de généralisation** = moyenne des ?? évaluations (croisées)

**10-fold cross validation**

Questions ?