

Cours 3

Introduction à l'Apprentissage Automatique

(branche de l'Intelligence Artificielle)

Plan du cours

1. Généralités sur l'apprentissage automatique

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

2. Classifier sans apprendre

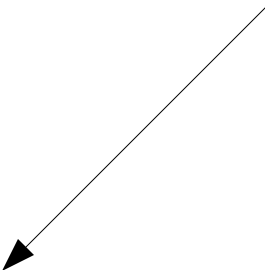
- Apprentissage Bayésien naïf
- Notions d'erreur

L'apprentissage automatique : c'est quoi ?

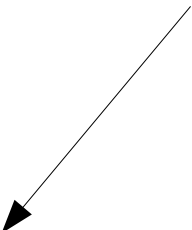
En anglais *Machine Learning*.

Branche de l'**Intelligence Artificielle** (*Artificial Intelligence*)
et de la **Fouille de Données** (*Data Mining*).

Principe/objectif : Extraire de l'**information** (ou connaissance) à partir de **données**



Des régularités
Une organisation
Des motifs (règles)
... ?



Données attribut/valeur
Bases de données relationnelle
Textes, Images, Vidéos
Données séquentielles
Données complexes (molécules)
Requêtes HTTP
...

Autant de situations/problèmes d'apprentissage !

L'apprentissage automatique : c'est quoi ?

En anglais *Machine Learning*.

Branche de l'**Intelligence Artificielle** (*Artificial Intelligence*)
et de la **Fouille de Données** (*Data Mining*).

Principe/objectif : Extraire de l'**information** (ou connaissance) à partir de **données**

APPRENDRE = GENERALISER

Produire un **modèle** qui « explique » les données.

Ce modèle doit permettre de faire des **prédictions** sur de nouvelles données.

L'apprentissage automatique

Animaux



Végétaux



???

L'apprentis

Animaux



Végétaux



???

Questions essentielles

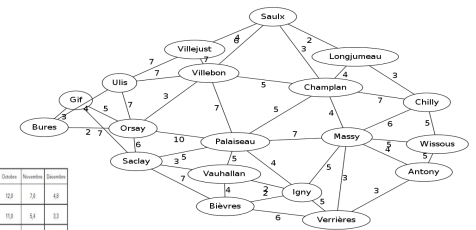
Données et connaissances a priori

- Quelles données sont disponibles ?
- Que sait-on du problème ? Du besoin ? De la demande ?



Représentation

- Comment représenter les exemples ?
- Comment représenter les hypothèses/modèles ?



	Janvier	Février	Mars	Avril	Mai	Juin	Juillet	Août	Septembre	Octobre	Novembre	Décembre
97a	38	44	18	87	12	93	93	90	91	94	92	79
97b	21	18	84	81	108	96	102	91	109	73	54	33
97c	22	42	19	104	104	91	209	209	109	123	61	31
97d	19	82	69	102	101	98	101	102	104	101	85	69
97e	64	12	67	113	164	91	203	204	119	168	88	14
97f	55	11	87	112	103	97	213	214	102	168	83	64
97g	61	12	81	118	111	207	204	228	102	165	61	61
97h	53	88	117	122	112	205	208	209	119	118	101	101

Méthode et estimation

- Quel est l'espace des hypothèses ?
- Comment évaluer une hypothèse en fonction des exemples connus ? Inconnus ?

Apprentissage supervisé

A partir d'un échantillon d'apprentissage $S = \{(x_i, u_i)\}_{1 \dots m}$
on cherche une loi de dépendance sous-jacente

- Par exemple une **fonction** h aussi proche possible de f (*fonction cible*) telle que :

$$\forall i, \quad u_i = f(x_i)$$

- Ou encore une **distribution de probabilités** $P(x_i, u_i)$

Afin de prédire la classe de nouvelles données

Induction supervisée

- Si f est une **fonction continue**
 - Régression
 - Estimation de densité
- Si f est une **fonction discrète**
 - Classification
- Si f est une **fonction binaire** (booléenne)
 - Apprentissage de concept (règles)

Induction supervisée

- Si f est une **fonction continue**

- Régression
- Estimation de densité

- Si f est une **fonction discrète**

- Classification

← TD/TP/CHALLENGE

- Si f est une **fonction binaire** (booléenne)

- Apprentissage de concept (règles)

Plan du cours

1. Généralités sur l'apprentissage automatique

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

2. Classifier sans apprendre

- Apprentissage Bayésien naïf
- Notions d'erreur

Représentation des données : attribut/valeur

Souvent les données sont stockées dans un tableau (attribut/valeur)

Champignons	Hauteur	Couleur	Dessous	Anneau	Volve	Classe
C1	grand	blanc	lamelles	non	non	+
C2	moyen	blanc	mousse	non	non	+
C3	petit	marron	lamelles	oui	oui	+
C4	petit	noir	lamelles	non	oui	+
C5	grand	blanc	mousse	non	non	+
C6	petit	blanc	lamelles	non	non	+
C7	grand	blanc	mousse	oui	oui	-
C8	petit	marron	mousse	oui	oui	-
C9	moyen	marron	lamelles	non	oui	-
C10	moyen	blanc	lamelles	oui	non	?

Ici les valeurs sont « symboliques »

Représentation des données : attribut/valeur

Souvent les données sont stockées dans un tableau (attribut/valeur)

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	3
1.5161	13.33	3.53	1.34	72.67	0.56	8.33	0	0	3
1.5167	13.24	3.57	1.38	72.7	0.56	8.44	0	0.1	3
1.51643	12.16	3.52	1.35	72.89	0.57	8.53	0	0	3
1.51665	13.14	3.45	1.76	72.48	0.6	8.38	0	0.17	3
1.52127	14.32	3.9	0.83	71.5	0	9.49	0	0	?

Ici les valeurs sont «numériques»

Représentation des données : attribut/valeur

Souvent les données sont stockées dans un tableau (attribut/valeur)

Champignons	Hauteur	Couleur	Dessous	Anneau	Volve	Classe
C1	22	blanc	lamelles	non	non	+
C2	17		mousse	non	non	+
C3	8	marron	lamelles	oui	oui	+
C4	9	noir	lamelles	non	oui	+
C5		blanc	mousse		non	+
C6	5	blanc	lamelles	non	non	+
C7	24	blanc		oui	oui	-
C8	6		mousse	oui		-
C9	17	marron	lamelles	non	oui	-
C10	12	blanc	lamelles		non	?

Ici les données sont « mixtes » (mélange de données numériques et symboliques) et certaines sont manquantes

Représentation des données

On peut envisager n'importe quelles données :

- Des fichiers images, vidéo, audio
- Des pages web, des mails, des requêtes web (détection d'attaques)
- Des flux de données, des séquences (génomique)
- ...

Qui souvent seront finalement représentées à l'aide de caractéristiques extraites, et donc dans un format attribut/valeur

Plan du cours

1. Généralités sur l'apprentissage automatique

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

2. Classifier sans apprendre

- Apprentissage Bayésien naïf
- Notions d'erreur

Représentation des connaissances par des règles

Données : Attribut/valeur symboliques

Champignons	Hauteur	Couleur	Dessous	Anneau	Volve	Classe
C1	grand	blanc	lamelles	non	non	+
C2	moyen	blanc	mousse	non	non	+
C3	petit	marron	lamelles	oui	oui	+
C4	petit	noir	lamelles	non	oui	+
C5	grand	blanc	mousse	non	non	+
C6	petit	blanc	lamelles	non	non	+
C7	grand	blanc	mousse	oui	oui	-
C8	petit	marron	mousse	oui	oui	-
C9	moyen	marron	lamelles	non	oui	-
C10	moyen	blanc	lamelles	oui	non	?

Connaissance : Règles symboliques

Si (*couleur = blanc*) et (*volve=non*) alors +

Représentation des connaissances par des règles

Données : Attribut/valeur 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	3
1.5161	13.33	3.53	1.34	72.67	0.56	8.33	0	0	3
1.5167	13.24	3.57	1.38	72.7	0.56	8.44	0	0.1	3
1.51643	12.16	3.52	1.35	72.89	0.57	8.53	0	0	3
1.51665	13.14	3.45	1.76	72.48	0.6	8.38	0	0.17	3
1.52127	14.32	3.9	0.83	71.5	0	9.49	0	0	?

Connaissance :

- Régression : $3 * Mg^2 - 1.5 \frac{Na}{K} = 7.8$
- Séparateur (SVM, réseau de neurones) : $3 * Mg^2 - 1.5 \frac{Na}{K} \geq 7.8$