Bienvenue – Welcome - Velkommen

Mercredi 17 mai 2017 - 9h/12h et 14h/17h
Initiation à l'apprentissage Artificiel (Machine Learning)
Partie 1 : Fondements et Méthodes

Vendredi 15 septembre 2017 - 9h/12h et 14h/17h
Initiation à l'apprentissage Artificiel (Machine Learning)
Partie 2 : Applications et Mise en pratique

Vendredi 13 octobre 2017 - 9h/12h et 14h/17h
Initiation aux langages R et Python

Mines Alès – EuroMov - Axe Biomedical Signal Processing
Stefan Janaqi
Vincent Derozier
Pierre Jean
Gérard Dray
prenom.nom@mines-ales.fr

IMT Mines Alès
École Mines-Télécom

EuroMov

Initiation à l'apprentissage Artificiel (Machine Learning)
Partie 1 : Fondements et Méthodes

Mercredi 17 mai 2017 - 9h/12h

- Introduction
- Fondements
- Classification supervisée
  - k plus proches voisins
  - Arbre de décision
  - Naive Bayes
  - SVM

Illustration
- Logiciel Weka
- Base de données Diabète

Mercredi 17 mai 2017 - 14h/17h

- Sparse Methods

Vendredi 15 septembre 2017 - 9h/12h

- Classification non supervisée

Mines Alès – EuroMov - Axe Biomedical Signal Processing
Stefan Janaqi
Vincent Derozier
Pierre Jean
Gérard Dray
prenom.nom@mines-ales.fr

IMT Mines Alès
École Mines-Télécom

EuroMov

Initiation à l'apprentissage Artificiel (Machine Learning)
Partie 1 : Fondements et Méthodes

Mercredi 17 mai 2017 - 9h/12h

Introduction

Fondements

Classification supervisée

- k plus proches voisins
- Arbre de décision
- Naive Bayes
- SVM

Illustration
- Logiciel Weka
- Base de données Diabète

Mercredi 17 mai 2017 - 14h/17h

Sparse Methods

Vendredi 15 septembre 2017 - 9h/12h

Classification non supervisée

Mines Alès – EuroMov - Axe Biomedical Signal Processing
Stefan Janaqi
Vincent Derozier
Pierre Jean
Gérard Dray
prenom.nom@mines-ales.fr

IMT Mines Alès
École Mines-Télécom

EuroMov

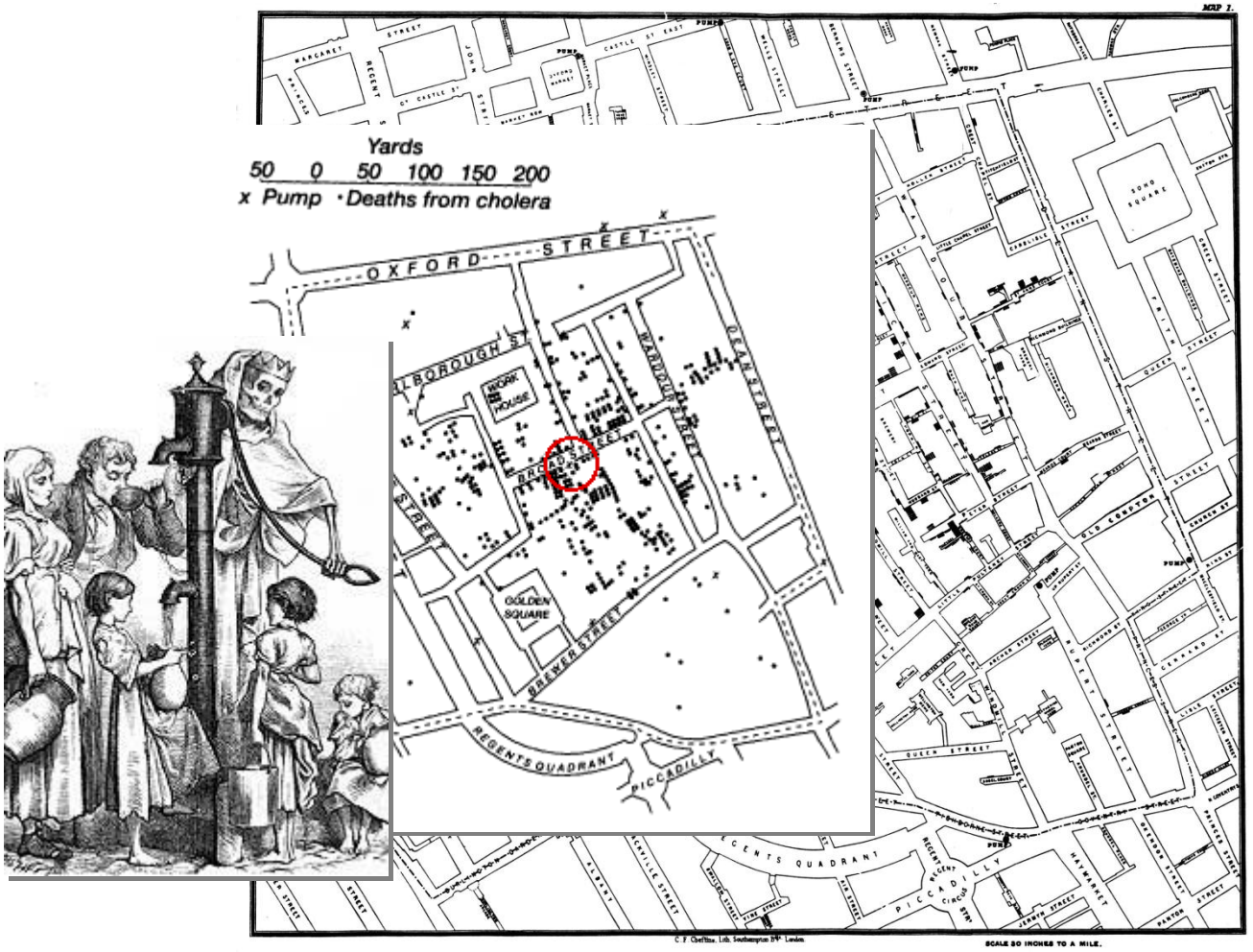
Objectifs

Extraction de Connaissances à partir de Données
Fouille de Données
Apprentissage Artificiel
Classification - Clustering

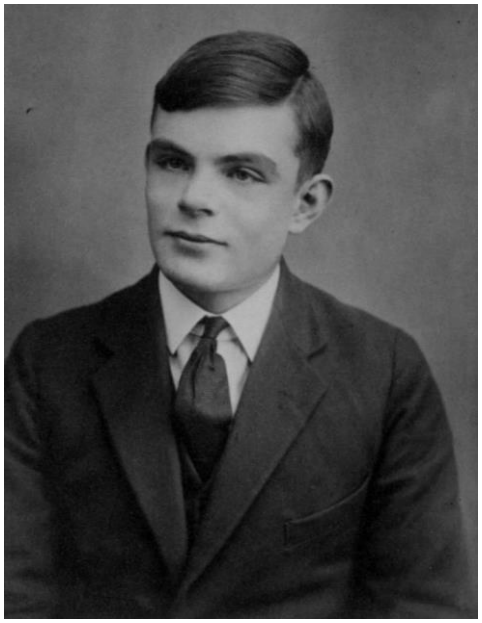
Knowledge Discovery from Data
Data Mining
Machine Learning
Classification - Clustering

Une histoire ancienne

John Snow
Épidémie de choléra de Broad Street (1854)



Une histoire ancienne



Alan Turing 1963

« Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's ? If this were then subjected to an appropriate course of education one would obtain the adult brain. »

1998 : Une légende urbaine ?

Initiation à l'apprentissage Artificiel (Machine Learning)
Partie 1 : Fondements et Méthodes

Mercredi 17 mai 2017 - 9h/12h

Introduction

Fondements

Classification supervisée

- k plus proches voisins
- Arbre de décision
- Naive Bayes
- SVM

Illustration
- Logiciel Weka
- Base de données Diabète

Mercredi 17 mai 2017 - 14h/17h

Sparse Methods

Vendredi 15 septembre 2017 - 9h/12h

Classification non supervisée

Mines Alès – EuroMov - Axe Biomedical Signal Processing
Stefan Janaqi
Vincent Derozier
Pierre Jean
Gérard Dray
prenom.nom@mines-ales.fr

IMT Mines Alès
École Mines-Télécom

EuroMov

Machine Learning
Knowledge Discovery from Data - Data Mining

- **Fouille de données (Data Mining) ou Extraction de connaissances à partir des données (Knowledge Discovery in Data)**

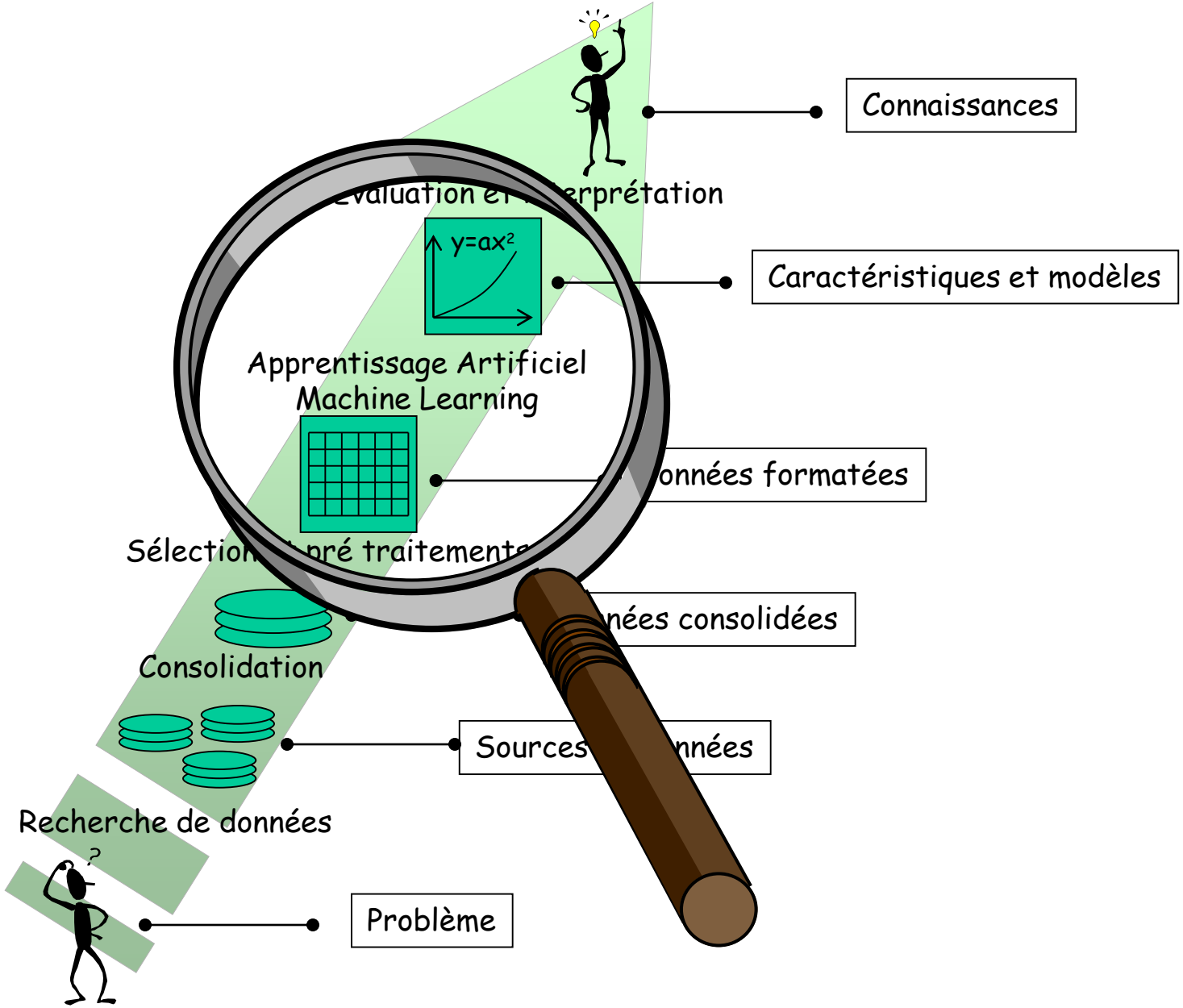
  La fouille de données prend en charge le processus complet d'extraction de connaissances : stockage dans une base de données, sélection des données à étudier, si nécessaire : nettoyage des données, puis utilisation de méthodes d'apprentissage artificiel afin de proposer des modèles à l'utilisateur, et enfin validation des modèles proposés.

- **Apprentissage artificiel (Machine Learning)**

  Cette notion englobe toute méthode permettant de construire un modèle de la réalité à partir de données.

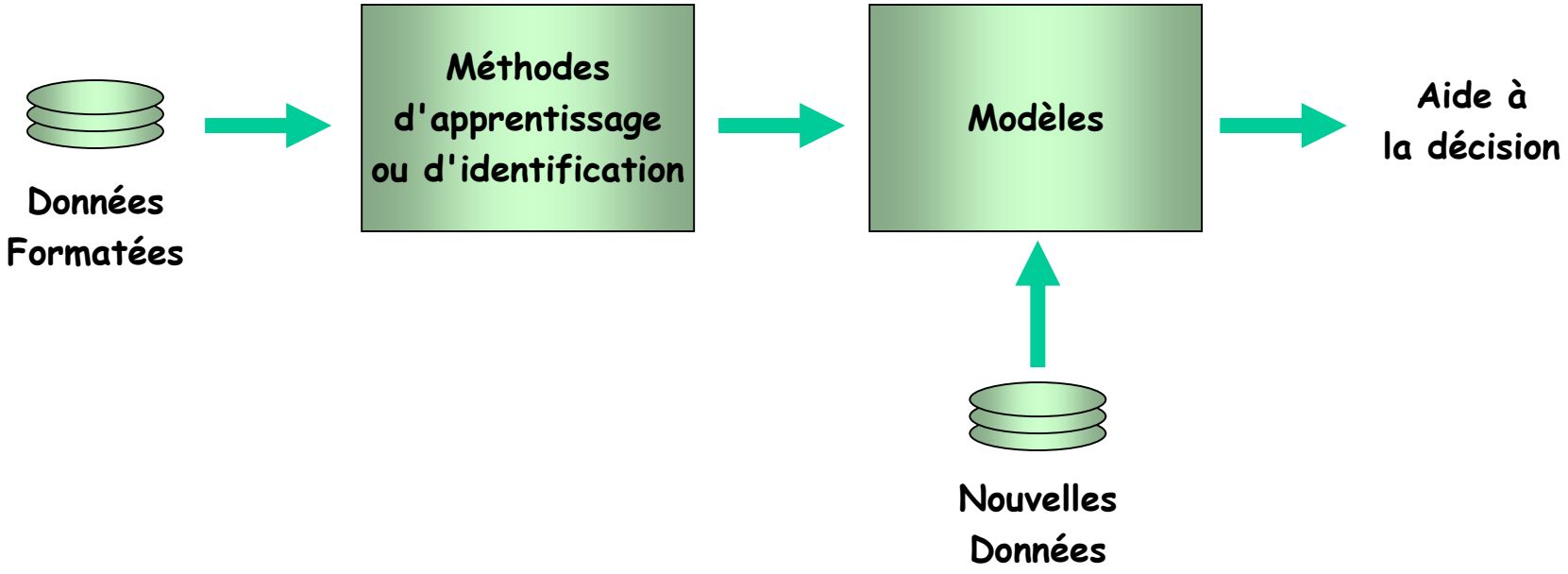
2010 Antoine Cornuéjols et Laurent Miclet : « Apprentissage Artificiel : Concepts et Algorithmes »

Data Mining

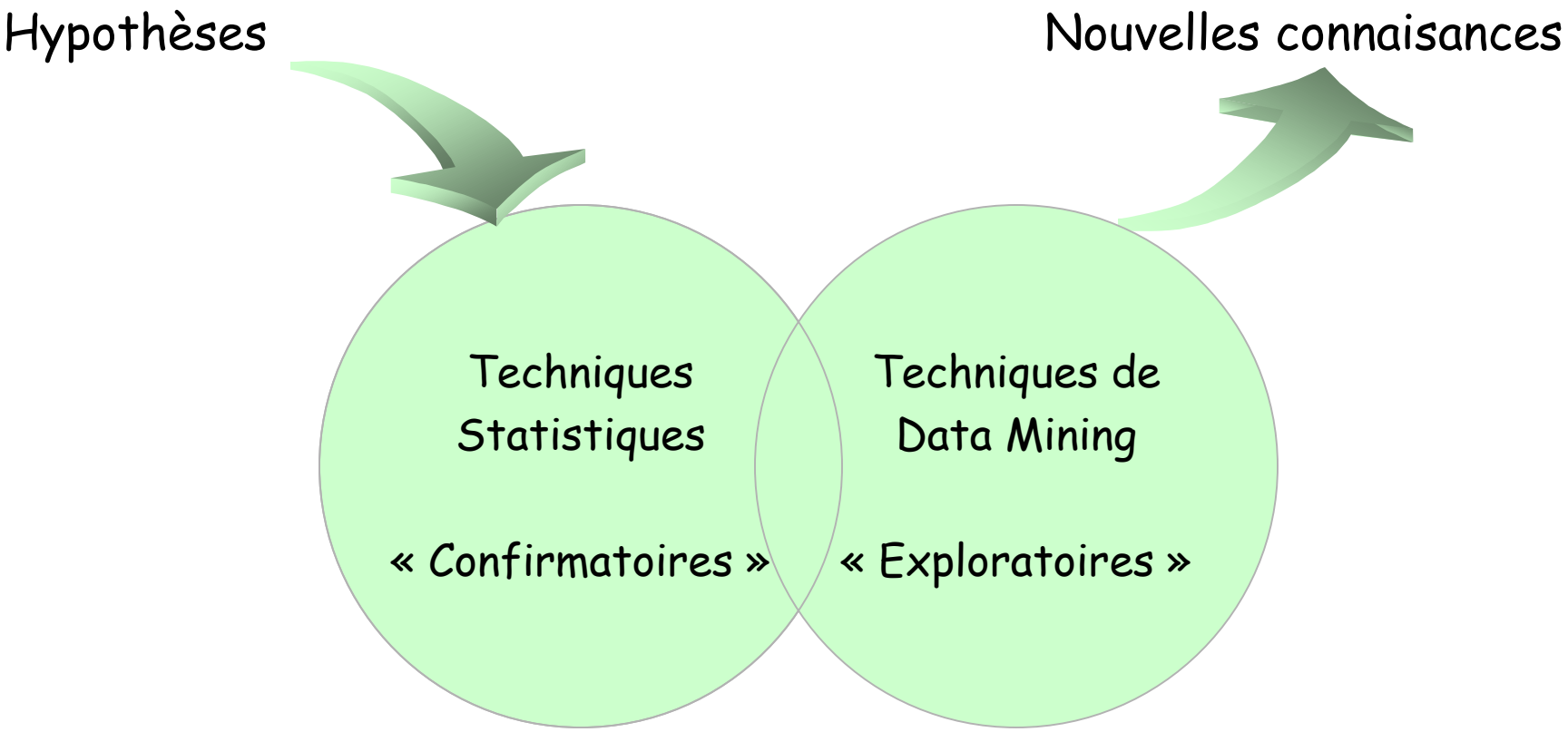

Connaissances
Caractéristiques et modèles
Apprentissage Artificiel
Machine Learning
Consolidation
Recherche de données
Problème

Data Mining

- Processus d'induction à partir des données



Data Mining vs Statistiques



Nature des données

- Tableau (Individus x Caractères) de dimension (n x p)

		Caractères										
		Age	Revenu	Sexe		Situation Matrimoniale			...	Caractère j	...	Caractère p
		X^1	X^2	M	F	Marié	Célibataire	Veuf Divorcé	...	X^j	...	X^p
				X^3	X^4	X^5	X^6	X^7				
Individus	X_1	x_1^1	x_1^2	x_1^3	x_1^4	x_1^5	x_1^6	x_1^7	...	x_1^j	...	x_1^p
	X_2	x_2^1	x_2^2	x_2^3	x_2^4	x_2^5	x_2^6	x_2^7	...	x_2^j	...	x_2^p
	...	...	...	...	...	...	...	...	...	...	...	...
	X_i	x_i^1	x_i^2	x_i^3	x_i^4	x_i^5	x_i^6	x_i^7	...	x_i^j	...	x_i^p
	...	...	...	...	...	...	...	...	...	...	...	...
	X_n	x_n^1	x_n^2	x_n^3	x_n^4	x_n^5	x_n^6	x_n^7	...	x_n^j	...	x_n^p

$$X : \begin{array}{c|cccccc} & X^1 & \dots & X^j & \dots & X^p \\ X_1 & x_1^1 & \dots & x_1^j & \dots & x_1^p \\ \dots & \dots & \dots & \dots & \dots & \dots \\ X_i & x_i^1 & \dots & x_i^j & \dots & x_i^p \\ \dots & \dots & \dots & \dots & \dots & \dots \\ X_n & x_n^1 & \dots & x_n^j & \dots & x_n^p \end{array}$$

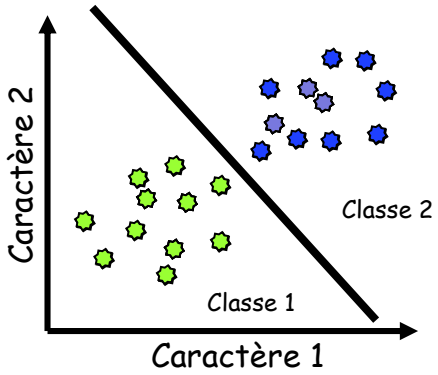
Vecteur Individus $X_i = [x_i^1, x_i^2, \dots, x_i^j, \dots, x_i^p]$

Vecteur Caractère $X^j = [x_1^j, x_2^j, \dots, x_i^j, \dots, x_n^j]^T$

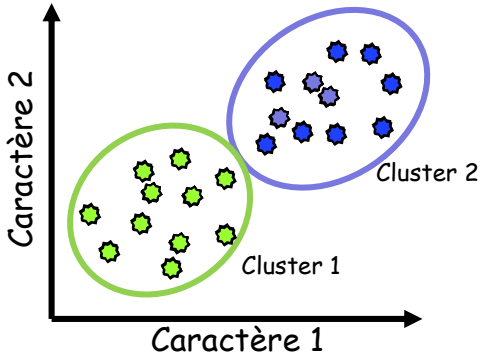
- Les caractères Age et Revenu sont quantitatifs
- Les caractères Sexe et Situation matrimoniale sont qualitatifs
- Modalité du caractère Situation matrimoniale : (Célibataire, Marié, Veuf ou divorcé)
- Les variables : X^3, X^4, X^5, X^6 et X^7 sont booléennes (1 = vrai, 0 = faux)

Machine Learning
Classification - Clustering

Classification Supervisée

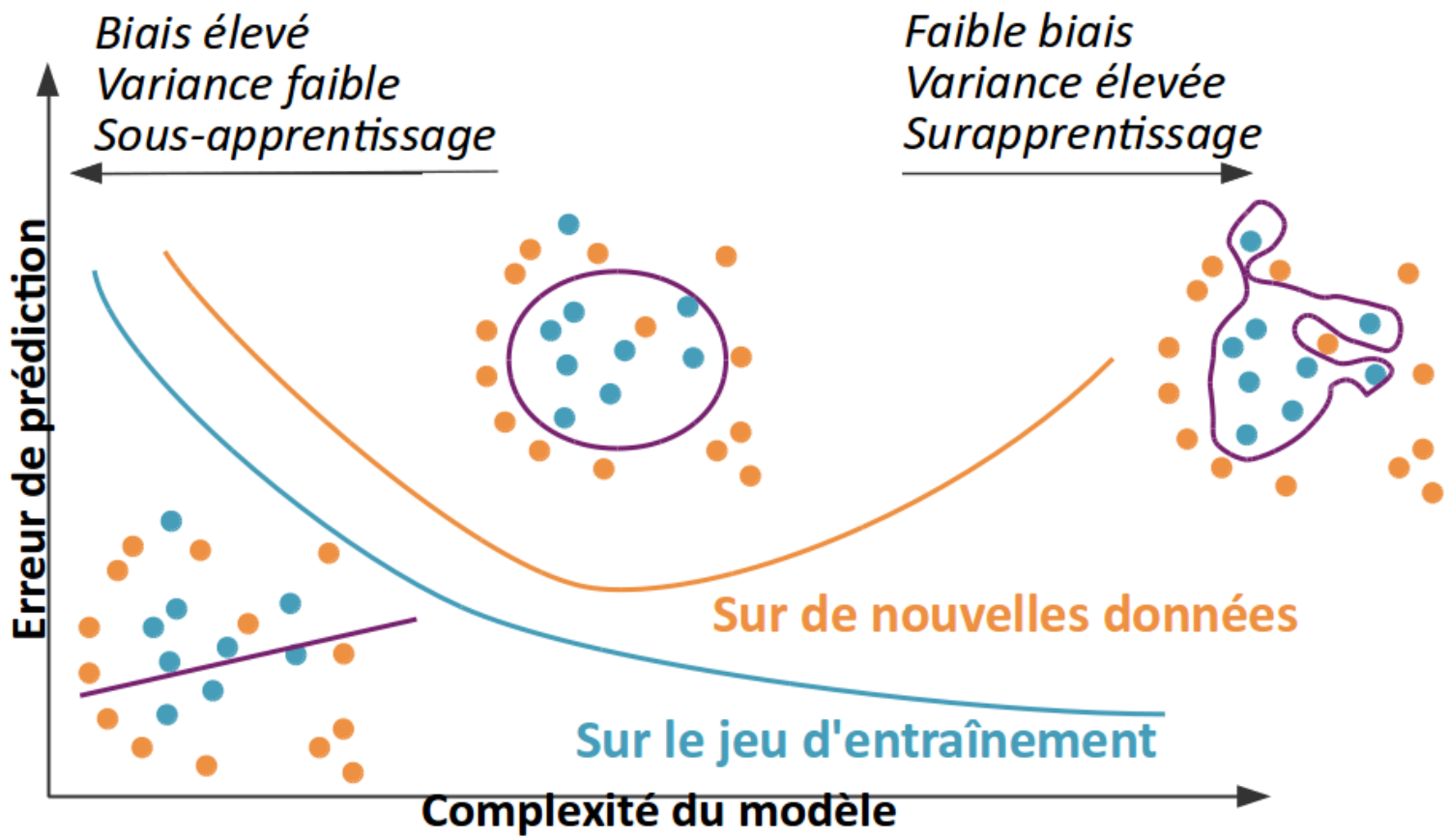


Classification Non Supervisée « Clustering »



Principe de parcimonie - « Rasoir d'Ockham »

- Sous apprentissage et Sur apprentissage



https://openclassrooms.com/courses/evaluez-et-ameliorez-les-performances-d-un-modele-de-machine-learning/comprenez-ce-qui-fait-un-bon-modele-d-apprentissage

Les étapes du processus de Data Mining

- Étape 1 : Poser le problème
- Étape 2 : Rechercher les données
- Étape 3 : Sélectionner les données pertinentes
- Étape 4 : Nettoyer les données
- Étape 5 : Transformer les données
- Étape 6 : Rechercher les caractéristiques et les modèles
- Étape 7 : Évaluer et valider les résultats

Les étapes du processus de Data Mining
Étape 1: Poser le problème

Un exemple : Classification de fleurs d'iris

- Problème :

  Connaissant :

  - la longueur des sépales d'une fleur d'iris
  - la largeur des sépales d'une fleur d'iris
  - la longueur des pétales d'une fleur d'iris
  - la largeur des pétales d'une fleur d'iris

  Est-il possible de classer les fleurs dans les catégories suivantes ?

  - Iris Setosa - classe 1
  - Iris Versicolour - classe 2
  - Iris Virginica - classe 3

- L'objectif est de fournir un modèle de classification qui accepte en entrée les 4 caractéristiques des sépales et des pétales des fleurs d'iris et qui fournit en sortie le numéro de la classe de la fleur (1, 2 ou 3).

- Les résultats de la classification seront évalués sur une base de données de test représentant 25% des données collectées sous la forme d'une matrice de confusion.

Les étapes du processus de Data Mining
Étape 2 : Rechercher les données

Un exemple : Classification de fleurs d'iris

- Title: Iris Plants Database Updated Sept 21 2000 by C.Blake
- File : iris.dat
- Sources:
  - (a) Creator: R.A. Fisher
  - (b) Donor: Michael Marshall
  - (c) Date: July, 1988
- Number of Instances: 150 (50 in each of three classes)
- Number of Attributes: 4 numeric attributes and the class
- Attribute Information:
  - 1. sepal length in cm
  - 2. sepal width in cm
  - 3. petal length in cm
  - 4. petal width in cm
  - 5. class:
  - -- Iris Setosa
  - -- Iris Versicolour
  - -- Iris Virginica
- Missing Attribute Values: None

Les étapes du processus de Data Mining
Étape 2 : Rechercher les données

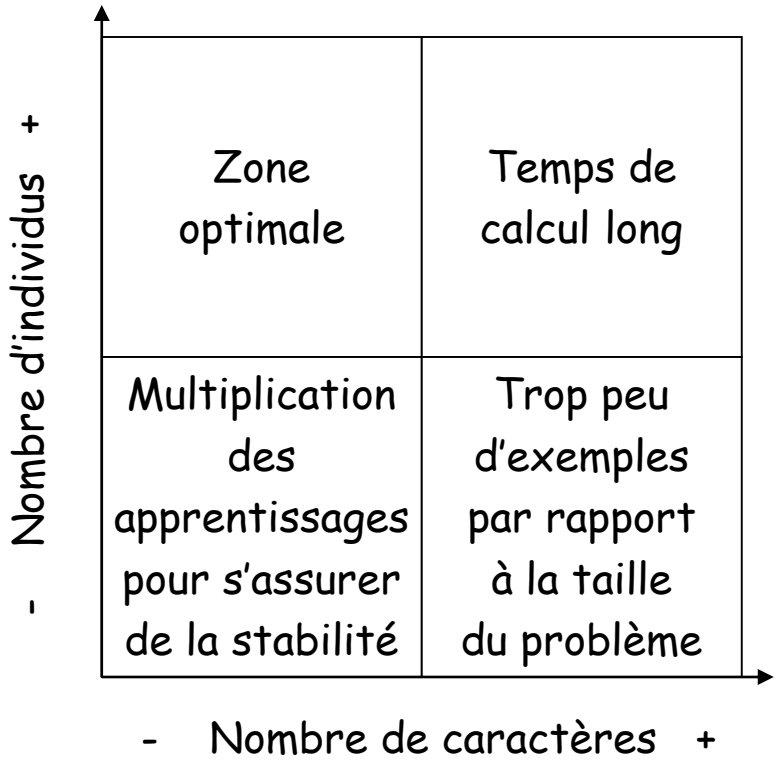
Un exemple : Classification de fleurs d'iris

Extrait de la base de données

5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
...
7.0,3.2,4.7,1.4,Iris-versicolor
6.4,3.2,4.5,1.5,Iris-versicolor
6.9,3.1,4.9,1.5,Iris-versicolor
...
6.3,3.3,6.0,2.5,Iris-virginica
5.8,2.7,5.1,1.9,Iris-virginica
7.1,3.0,5.9,2.1,Iris-virginica
...

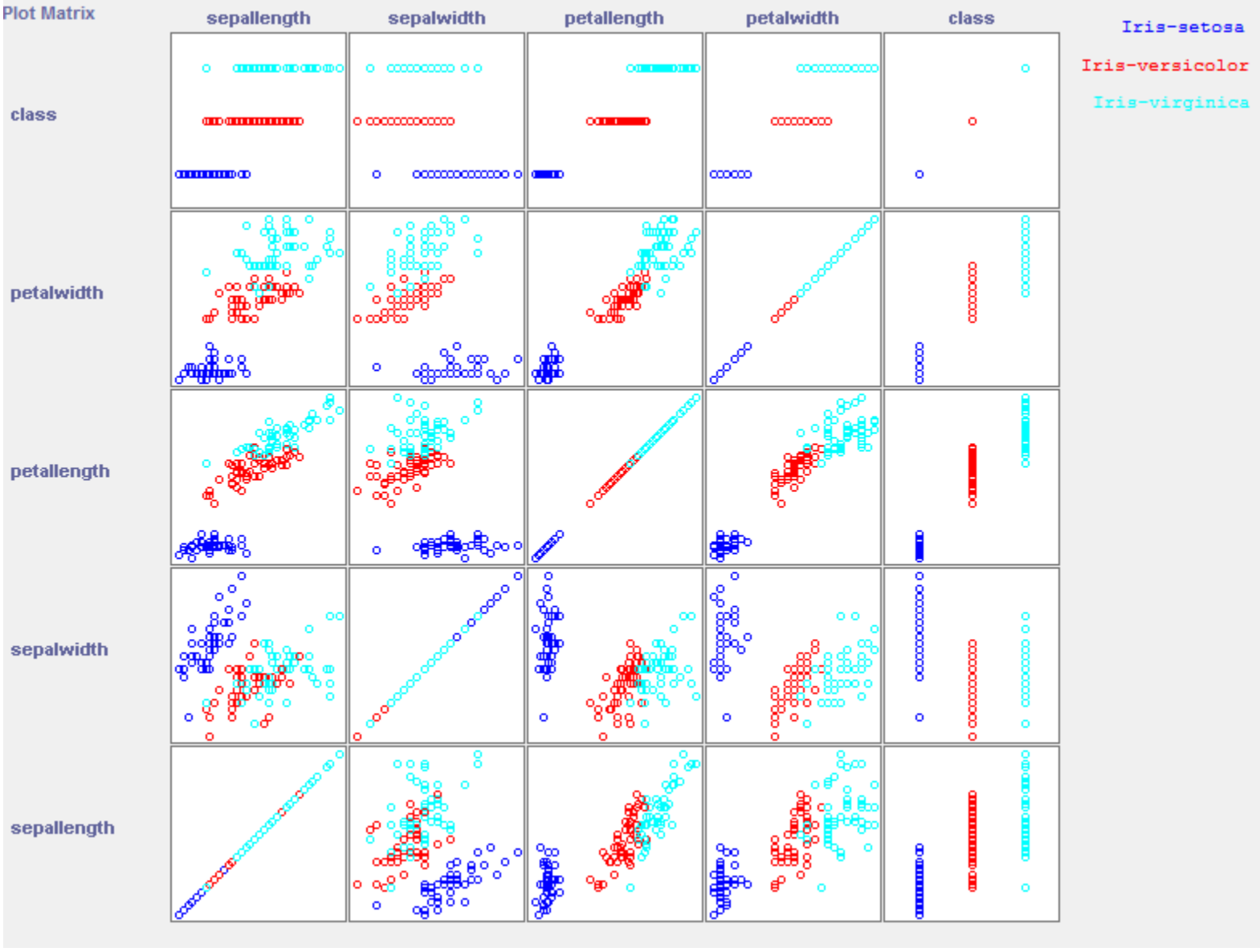
Les étapes du processus de Data Mining
Étape 3 : Sélectionner les données pertinentes

- Réduire les dimensions
  - Expertise humaine
  - Analyses graphiques
  - Analyses de corrélation
  - Analyse en composantes principales
  - ...



Les étapes du processus de Data Mining
Étape 3 : Sélectionner les données pertinentes

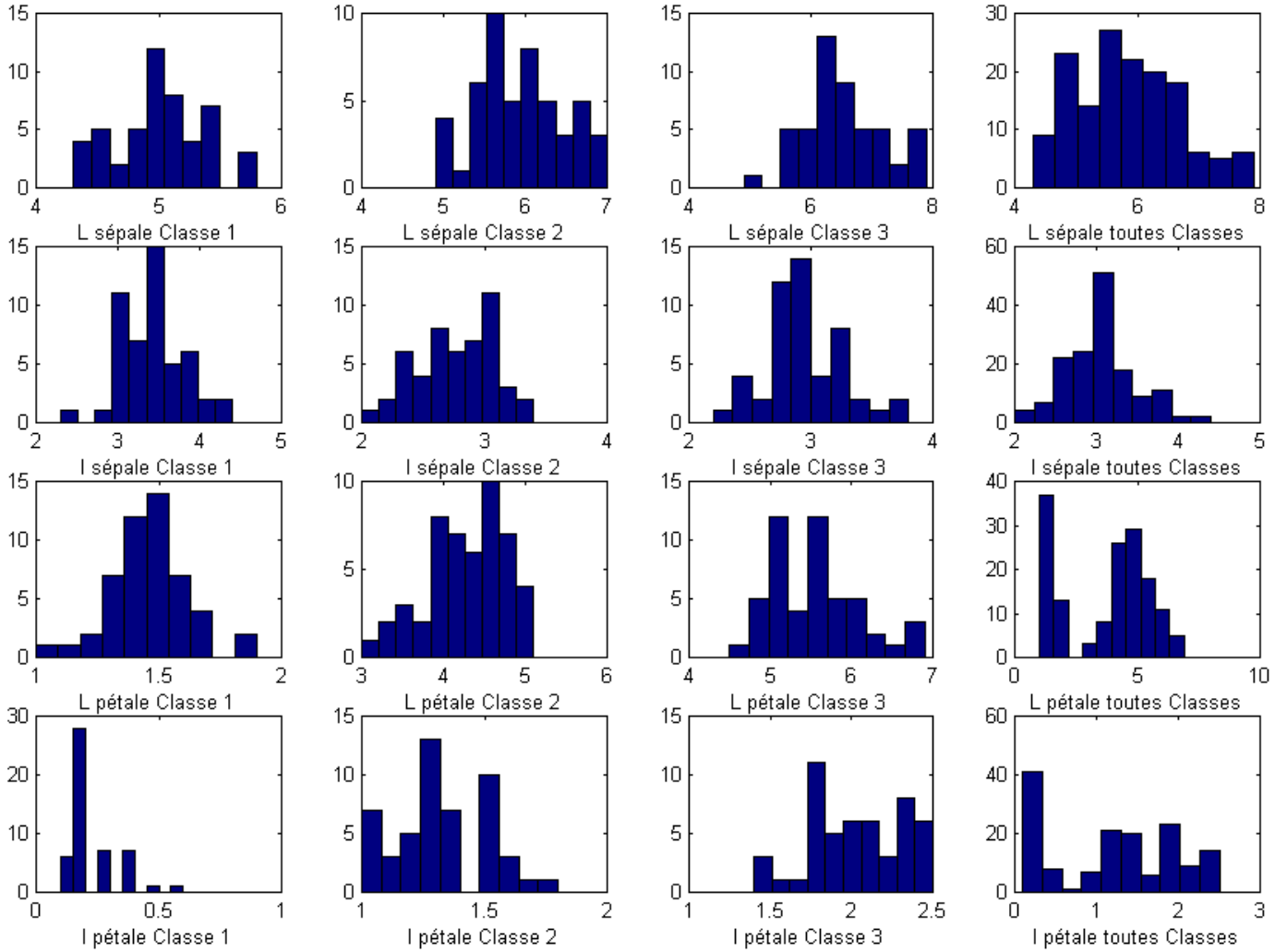
Un exemple : Classification de fleurs d'iris



Les étapes du processus de Data Mining
Étape 3 : Sélectionner les données pertinentes

Un exemple : Classification de fleurs d'iris

Histogrammes des caractéristiques



Les étapes du processus de Data Mining
Étape 3 : Sélectionner les données pertinentes

Un exemple : Classification de fleurs d'iris

Quelques mesures statistiques

Coefficient de corrélation	L sépale	l sépale	L pétale	l pétale	Classe
L sépale	1,00	-0,11	0,87	0,82	0,78
l sépale	-0,11	1,00	-0,42	-0,36	-0,42
L pétale	0,87	-0,42	1,00	0,96	0,95
l pétale	0,82	-0,36	0,96	1,00	0,96
Classe	0,78	-0,42	0,95	0,96	1,00
Minimum	4,30	2,00	1,00	0,10	1,00
Maximum	7,90	4,40	6,90	2,50	3,00
Moyenne	5,84	3,05	3,76	1,20	2,00
Ecart-type	0,83	0,43	1,76	0,76	0,82
Médiane	5,80	3,00	4,35	1,30	2,00
Interquartile	1,30	0,50	3,50	1,50	2,00

Les étapes du processus de Data Mining
Étape 4 : Nettoyer les données

- Vérifier l'origine des données
- Traiter les valeurs aberrantes
- Traiter les valeurs manquantes
- Traiter les valeurs nulles
- …

Les étapes du processus de Data Mining
Étape 5 : Transformer les données

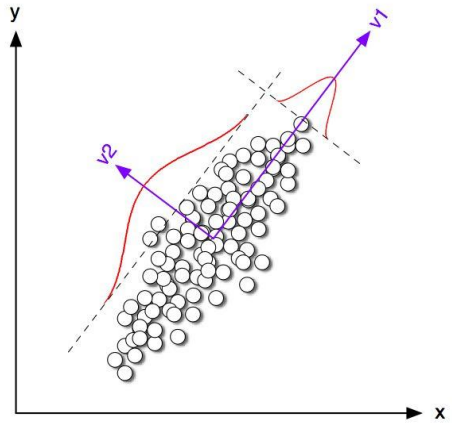
- Cette étape peut consister à :
  - Coder les informations qualitatives
  - Coder en ratios (pourcentages)
  - Normaliser les données
  - Transformer les dates en durées
  - Transcoder les données,

    exemple : code postal en coordonnées géographiques
  - Exprimer des fréquences
  - Réaliser des combinaisons de variables
  - ...

Les étapes du processus de Data Mining
Étape 5 : Transformer les données

Un exemple : Classification de fleurs d'iris

Analyse en composantes principales

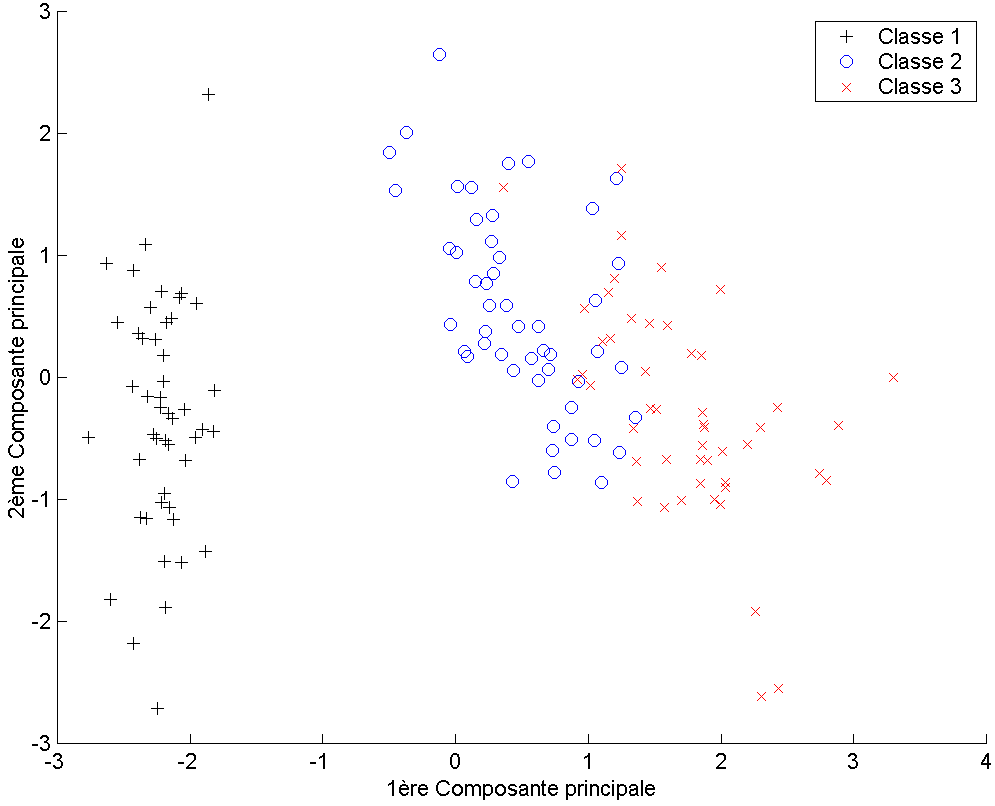


Coefficients

	CP1	CP2	CP3	CP4
L sépale	0.5224	-0.3723	0.7210	-0.2620
l sépale	-0.2634	-0.9256	-0.2420	0.1241
L pétale	0.5813	-0.0211	-0.1409	0.8012
l pétale	0.5656	-0.0654	-0.6338	-0.5235

Valeurs propres et % d'inertie

CP1	2.9108	72.77%
CP2	0.9212	23.03%
CP3	0.1474	03.69%
CP4	0.0206	00.51%



Les étapes du processus de Data Mining
Étape 5 : Transformer les données

- Données Centrées - Réduites

$$X^j = [x_1^j, x_2^j, \ldots, x_i^j, \ldots, x_n^j]^T$$

$$x_i^{j*} = \frac{x_i^j - \overline{X^j}}{s_{X^j}}$$

- Données Bornées [0,1]

$$x_i^{j*} = \frac{x_i^j - X_{min}^j}{X_{max}^j - X_{min}^j}$$

Données Bornées [a,b]

$$x_i^{j*} = \frac{x_i^j - offset}{quotient}$$

$$quotient = \left(X_{max}^j - X_{min}^j\right)/(b-a)$$

$$offset = X_{max}^j - quotient \times b$$

Les étapes du processus de Data Mining
Étape 5 : Transformer les données

$$X_i = [x_i^1, x_i^2, \ldots, x_i^j, \ldots, x_i^p] \qquad X_k = [x_k^1, x_k^2, \ldots, x_k^j, \ldots, x_k^p]$$

Mesures de distance

Formule Générale : $d_{ik}^2 = d^2(X_i; X_k) = (X_i - X_k)^T M (X_i - X_k)$

Distance Euclidienne : $M = I = \begin{bmatrix} 1 & & & & 0 \\ & 1 & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & \cdot \\ 0 & & & & 1 \end{bmatrix}$

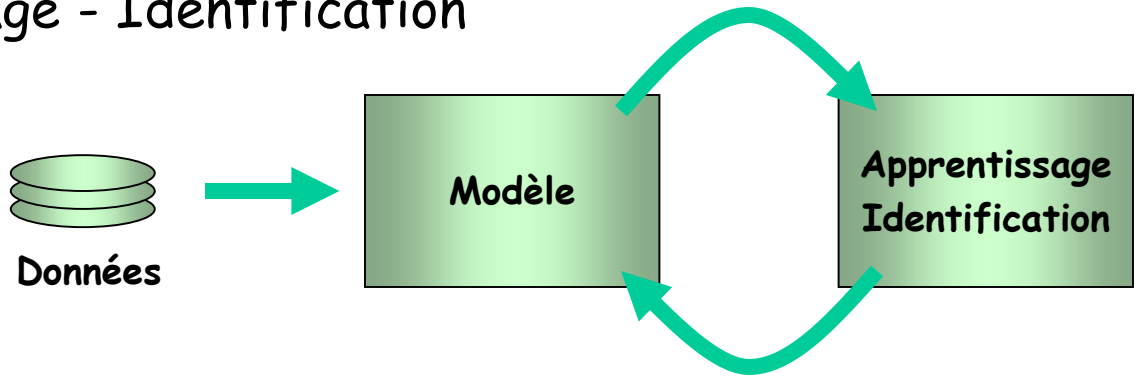
Distance standardisée : $M = D_{1/s^2} = \begin{bmatrix} 1/s_1^2 & & & & 0 \\ & 1/s_2^2 & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & \cdot \\ 0 & & & & 1/s_p^2 \end{bmatrix}$

Distance de Mahalanobis : $M = V^{-1}$ $\quad V =$ matrice de covariance

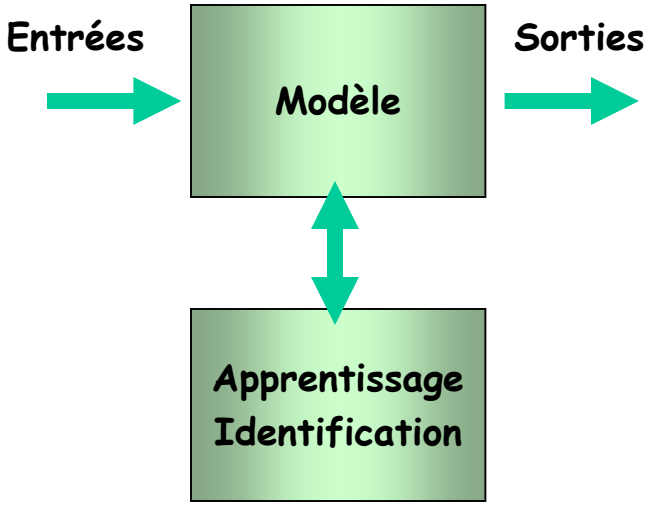
Distance "City Block" : $d_{jk}^2 = \sum_{i=1}^{n} \left| x_i^j - x_i^k \right|$

Les étapes du processus de Data Mining
Étape 6 : Rechercher les caractéristiques et les modèles

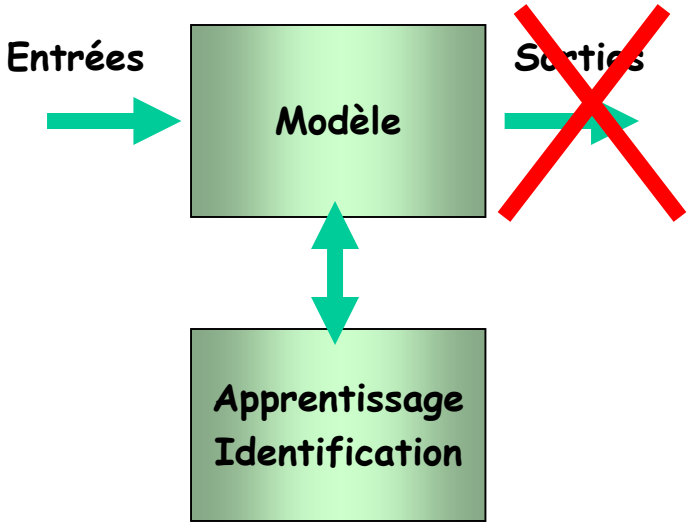
- Apprentissage - Identification



- Apprentissage Supervisé



Apprentissage Non supervisé



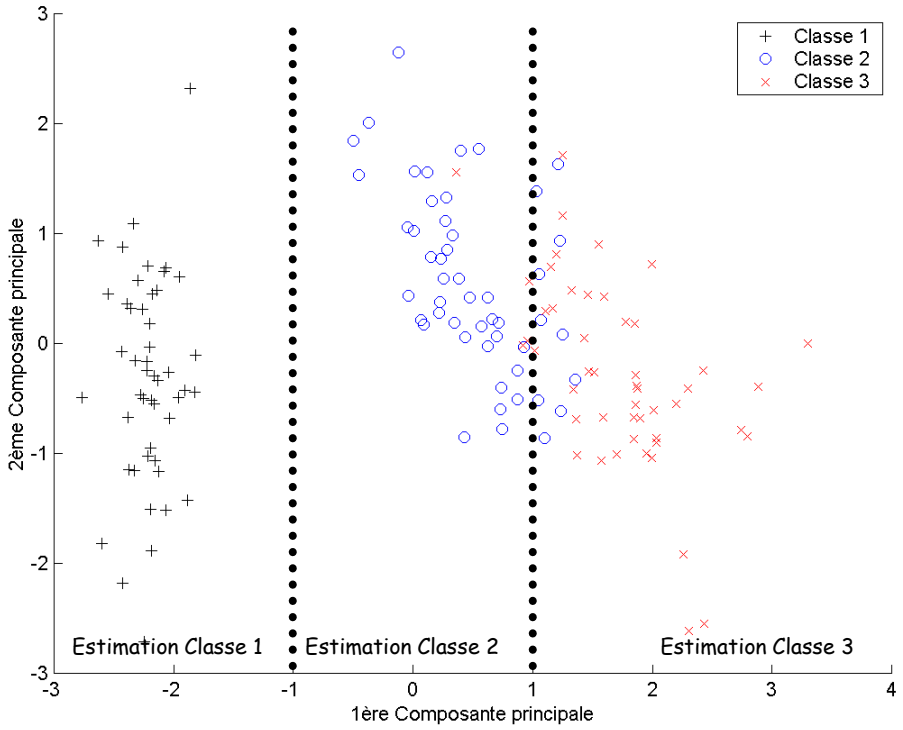
Les étapes du processus de Data Mining
Étape 6 : Rechercher les caractéristiques et les modèles

- Choix d'une méthode ou d'une technique :
  - K plus proches voisins
  - Méthodes bayésiennes
  - Arbres de décision
  - Support Vector Machine
  - Réseaux de neurones
  - Régression
  - Règles
  - Logique Floue
  - Algorithmes génétiques
  - Le raisonnement à base de cas
  - Les réseaux Bayésiens
  - ...

Les étapes du processus de Data Mining
Étape 6 : Rechercher les caractéristiques et les modèles

Un exemple : Classification de fleurs d'iris

Analyse en composantes principales



Arbre de décision

Decision tree
---

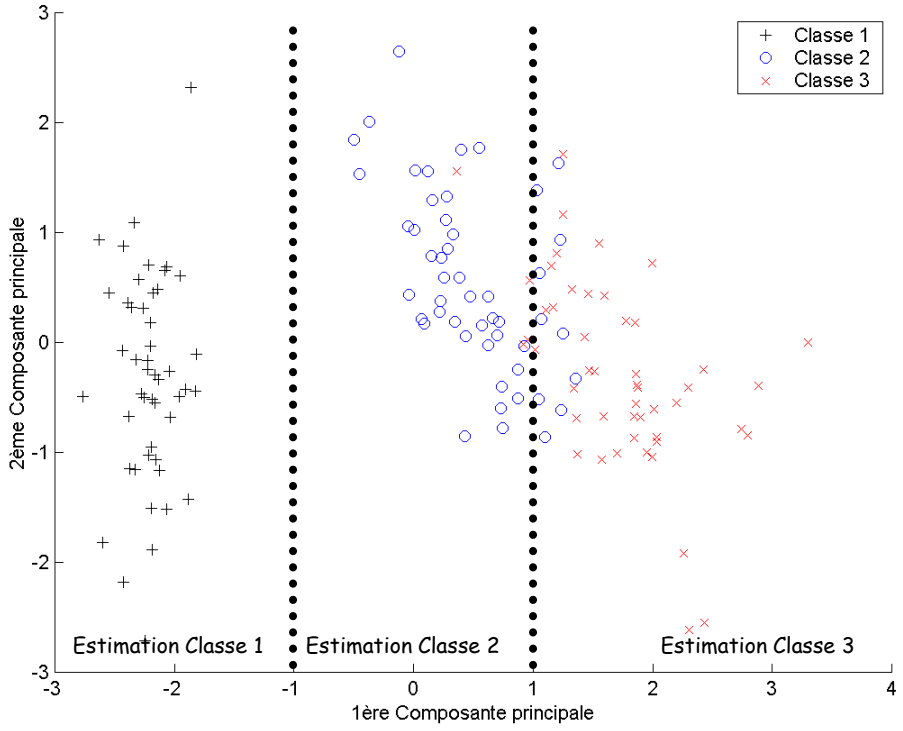
petalwidth <= 0.6: Iris-setosa
petalwidth > 0.6
| petalwidth <= 1.7
| | petallength <= 4.9: Iris-versicolor
| | petallength > 4.9
| | | petalwidth <= 1.5: Iris-virginica
| | | petalwidth > 1.5: Iris-versicolor
| petalwidth > 1.7: Iris-virginica

Les étapes du processus de Data Mining
Étape 7 : Évaluer et valider les résultats

- Évaluation quantitative :
  - matrice de confusion

    exemple : Classification de fleurs d'iris

		Classes estimées		
		C1	C2	C3
Classes réelles	C1	50	0	0
	C2	0	40	10
	C3	0	4	46



Les étapes du processus de Data Mining
Étape 7 : Évaluer et valider les résultats

- Évaluation quantitative :
  - matrice de confusion

    exemple : Classification de fleurs d'iris

		Classes estimées		
		C1	C2	C3
Classes réelles	C1	49	1	0
	C2	0	47	3
	C3	0	2	48

Arbre de décision

```
petalwidth <= 0.6: Iris-setosa
petalwidth > 0.6
|   petalwidth <= 1.7
|   |   petallength <= 4.9: Iris-versicolor
|   |   petallength > 4.9
|   |   |   petalwidth <= 1.5: Iris-virginica
|   |   |   petalwidth > 1.5: Iris-versicolor
|   petalwidth > 1.7: Iris-virginica
```

Les étapes du processus de Data Mining
Étape 7 : Évaluer et valider les résultats

- Évaluation quantitative :
  - Correctly Classified Instances et Incorrectly Classified Instances

    exemple : Classification de fleurs d'iris

		Classes estimées		
		C1	C2	C3
Classes réelles	C1	49	1	0
	C2	0	47	3
	C3	0	2	48

Arbre de décision

Correctly Classified Instances	**144**	**96%**
Incorrectly Classified Instances	**6**	**4%**

Les étapes du processus de Data Mining
Étape 7 : Évaluer et valider les résultats

- Évaluation quantitative :
  - Notions de Vrais/Faux Positifs et Vrais/Faux Négatifs

		Classes estimées	
		C	$\overline{C}$
Classe réelles	C	VP	FN
	$\overline{C}$	FP	VN

VP :Vrais Positifs, VN : Vrais Négatifs,

FN : Faux Négatifs, FP : Faux Positifs

Les étapes du processus de Data Mining
Étape 7 : Évaluer et valider les résultats

- Évaluation quantitative :
  - Taux de VP et Taux de FP
  - exemple : Classification de fleurs d'iris

		Classes estimées	
		C	$\overline{C}$
Classe réelles	C	VP	FN
	$\overline{C}$	FP	VN

	$C1$	$\overline{C1}$
$C1$	49	1
$\overline{C1}$	0	100

	$C2$	$\overline{C2}$
$C2$	47	3
$\overline{C2}$	3	97

	$C3$	$\overline{C3}$
$C3$	48	2
$\overline{C3}$	3	97

Arbre de décision

TP Rate	FP Rate	Class
0.98	0	Iris-setosa
0.94	0.03	Iris-versicolor
0.96	0.03	Iris-virginica

Taux de VP = VP/(VP+FN) Taux de FP = FP/(FP+VN)

Les étapes du processus de Data Mining
Étape 7 : Évaluer et valider les résultats

- Évaluation quantitative :
  - Précision, Rappel, F-Mesure
  - exemple : Classification de fleurs d'iris

		Classes estimées	
		C	$\overline{C}$
Classe réelles	C	VP	FN
	$\overline{C}$	FP	VN

	$C1$	$\overline{C1}$
$C1$	49	1
$\overline{C1}$	0	100

	$C2$	$\overline{C2}$
$C2$	47	3
$\overline{C2}$	3	97

	$C3$	$\overline{C3}$
$C3$	48	2
$\overline{C3}$	3	97

Arbre de décision

Precision	Recall	F-Measure	Class
1	0.98	0.99	Iris-setosa
0.94	0.94	0.94	Iris-versicolor
0.94	0.96	0.95	Iris-virginica

Précision = VP/(VP+FP) Rappel = VP/(VP+FN) = Taux VP

F-Mesure = 2xPrécisionxRappel/(Précision+Rappel)

Les étapes du processus de Data Mining
Étape 7 : Évaluer et valider les résultats

		Condition (as determined by "Gold standard")		
		Condition positive	Condition negative	
Test outcome	Test outcome positive	**True positive**	**False positive** (Type I error)	Precision = Σ True positive / Σ Test outcome positive
	Test outcome negative	**False negative** (Type II error)	**True negative**	Negative predictive value = Σ True negative / Σ Test outcome negative
		Sensitivity = Σ True positive / Σ Condition positive	**Specificity** = Σ True negative / Σ Condition negative	Accuracy = Σ True positive + Σ True negative / Σ Total population

Les étapes du processus de Data Mining
Étape 7 : Évaluer et valider les résultats

- Évaluation quantitative

A

63	37
28	72

TPR = 0.63
FPR = 0.28

B

77	23
77	23

TPR = 0.77
FPR = 0.77

C

24	76
88	12

TPR = 0.24
FPR = 0.88

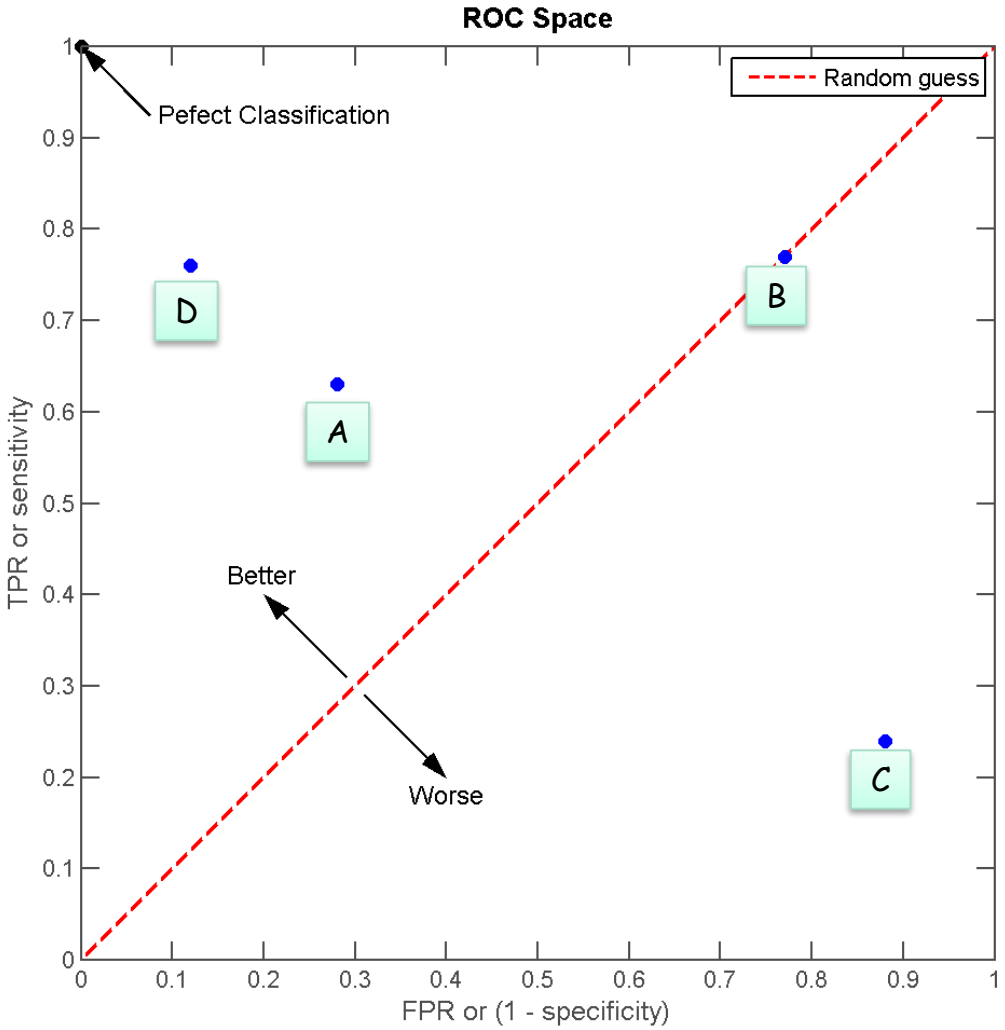
D

76	24
12	88

TPR = 0.76
FPR = 0.12

TP	FN
FP	TN

TPR= TP/(TP+FN)
FPR= FP/(FP+TN)



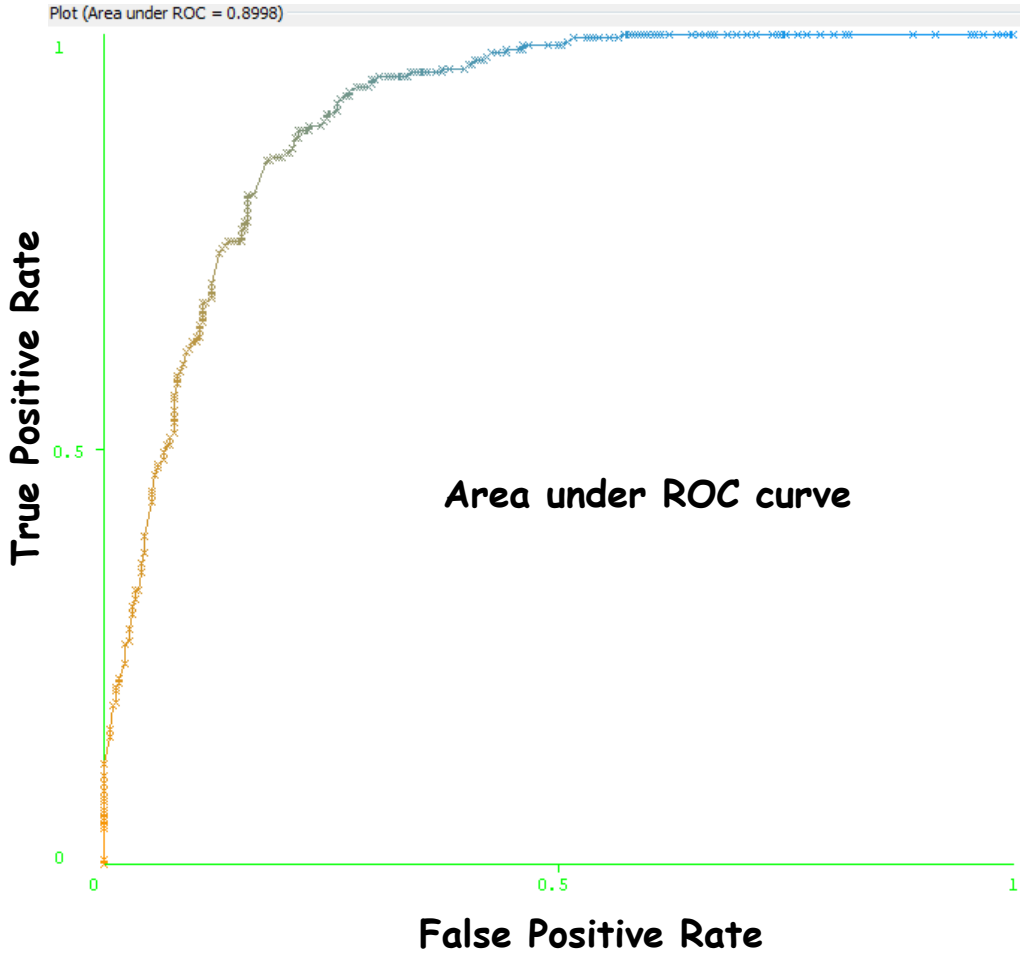
ROC : Receiver Operating Characteristic

Les étapes du processus de Data Mining
Étape 7 : Évaluer et valider les résultats

Model			
Positive	Negative		
TP	FN	Positive	Real
FP	TN	Negative	

True Positive Rate = TP / (TP + FN)

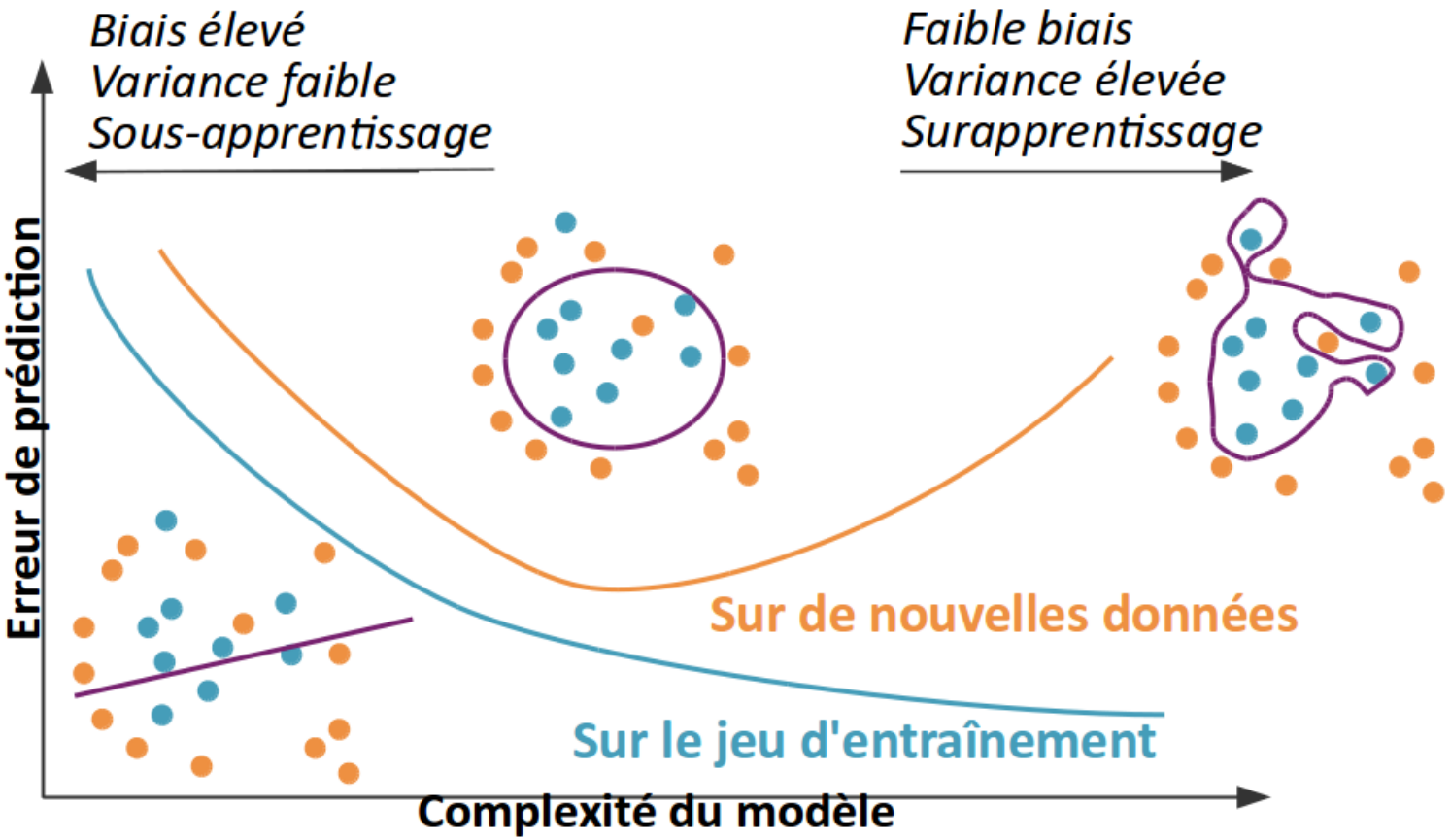
False Positive Rate = FP / (FP + TN)



ROC : Receiver Operating Characteristic

Les étapes du processus de Data Mining
Étape 7 : Évaluer et valider les résultats

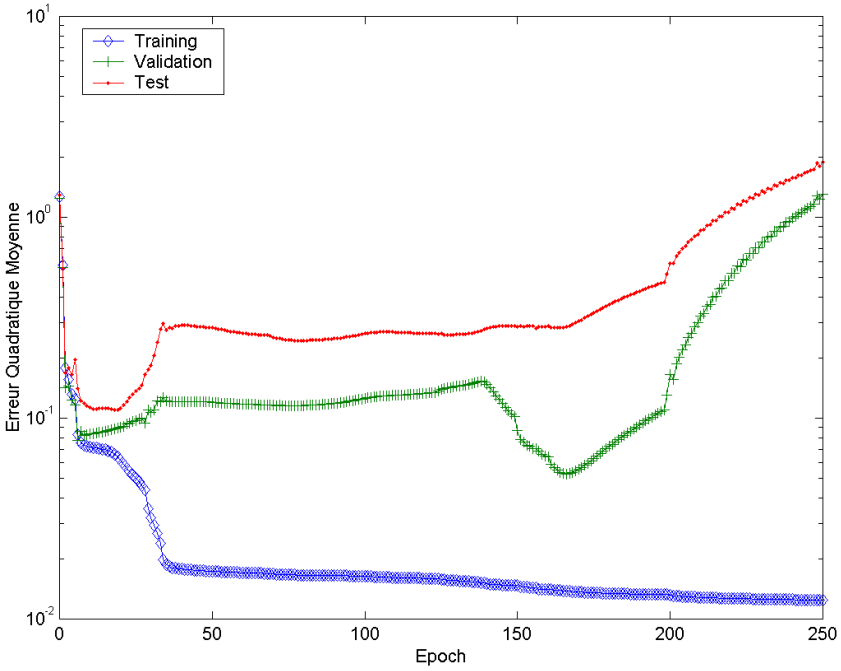
- Sous apprentissage et Sur apprentissage
- Principe de parcimonie : « Rasoir d'Ockham »

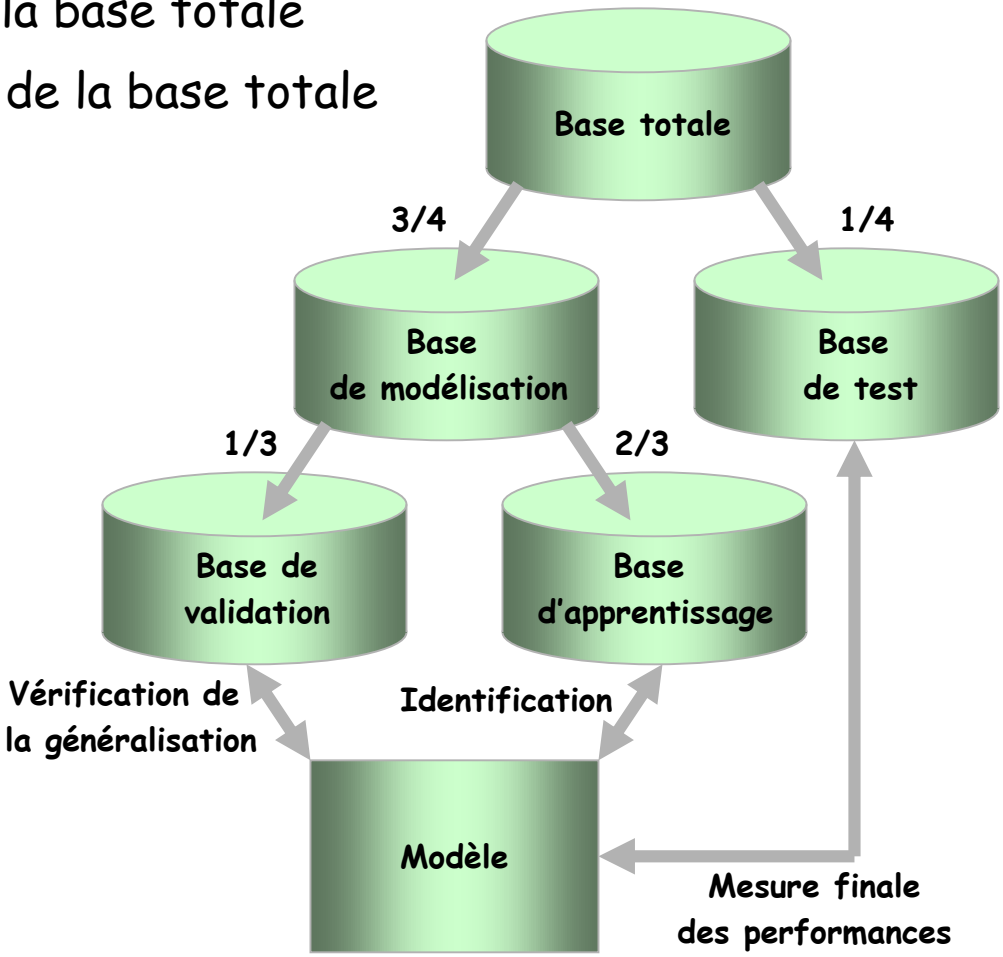


https://openclassrooms.com/courses/evaluez-et-ameliorez-les-performances-d-un-modele-de-machine-learning/comprenez-ce-qui-fait-un-bon-modele-d-apprentissage

Les étapes du processus de Data Mining
Étape 7 : Évaluer et valider les résultats

- Évaluation par le test
  - Base de test 25% de la base totale
  - Base de validation 25% de la base totale
  - Base d'apprentissage 50% de la base totale

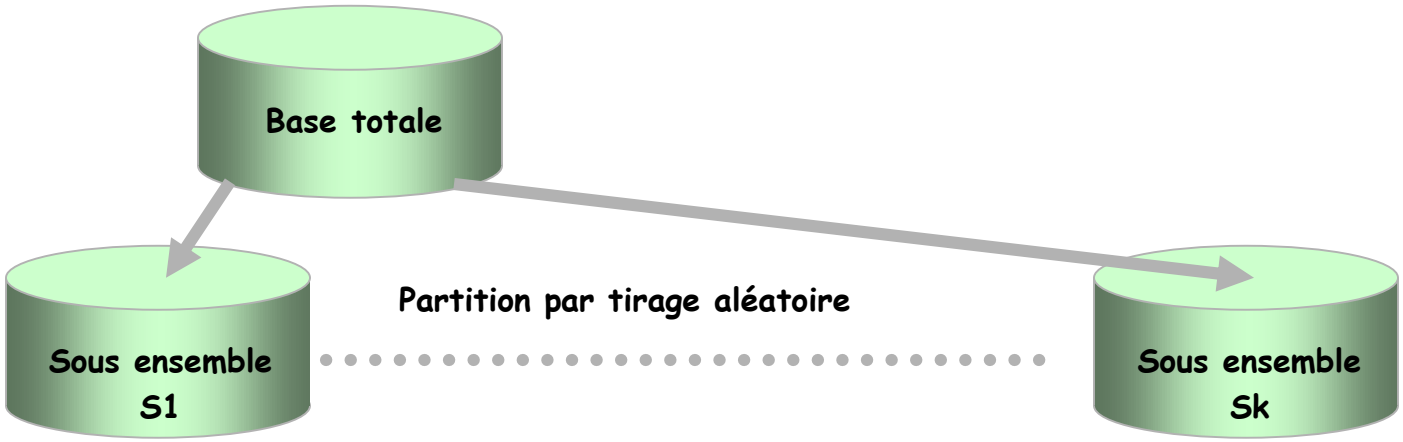




Les étapes du processus de Data Mining
Étape 7 : Évaluer et valider les résultats

- Validation croisée

Attention cette technique ne fournit pas de modèles mais permet d'évaluer les performances de la modélisation dans les cas où la base totale est petite



Pour tout i de 1 à k

- Appliquer une méthode d'identification sur (Base totale - Si)
- Calculer la performance du modèle sur Si

Fin

Agréger les performances individuelles pour établir une mesure globale

Si la taille des Si est de 1 individu on parle alors de « Leave one out »

Initiation à l'apprentissage Artificiel (Machine Learning)
Partie 1 : Fondements et Méthodes

Mercredi 17 mai 2017 - 9h/12h

Introduction

Fondements

Classification supervisée

k plus proches voisins

Arbre de décision

Naive Bayes

SVM

Illustration
- Logiciel Weka
- Base de données Diabète

Mercredi 17 mai 2017 - 14h/17h

Sparse Methods

Vendredi 15 septembre 2017 - 9h/12h

Classification non supervisée

Mines Alès – EuroMov - Axe Biomedical Signal Processing
Stefan Janaqi
Vincent Derozier
Pierre Jean
Gérard Dray
prenom.nom@mines-ales.fr

IMT Mines Alès
École Mines-Télécom

EuroMov

Classification

- Objectif : identifier les classes auxquelles appartiennent des objets à partir de traits descriptifs

Π est la population,

D est l'ensemble des descriptions,

C est l'ensemble des classes

$X : \Pi \rightarrow D$

fonction qui associe une description à tout élément de la population

$Y : \Pi \rightarrow C$

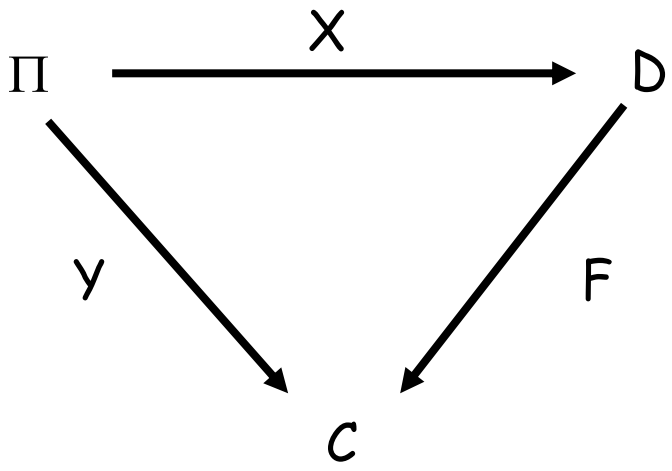
fonction qui associe une classe à tout élément de la population

$F : D \rightarrow C$

fonction de classement recherchée

Classification

- Le but est de rechercher une fonction de classement F telle que $F \circ X$ soit une bonne approximation de Y.



Méthodes de classification

- 4 grandes catégories de méthodes de classification
  - Distances - exemple : k plus proches voisins
  - Probabiliste - exemple : Naives Bayes
  - Arbres de décision - exemple : ID3
  - Distances et optimisation - exemple : machines à vecteurs de support

- Il existe un nombre important de méthodes de classification basées sur ces 4 catégories

Algorithme des k plus proches voisins

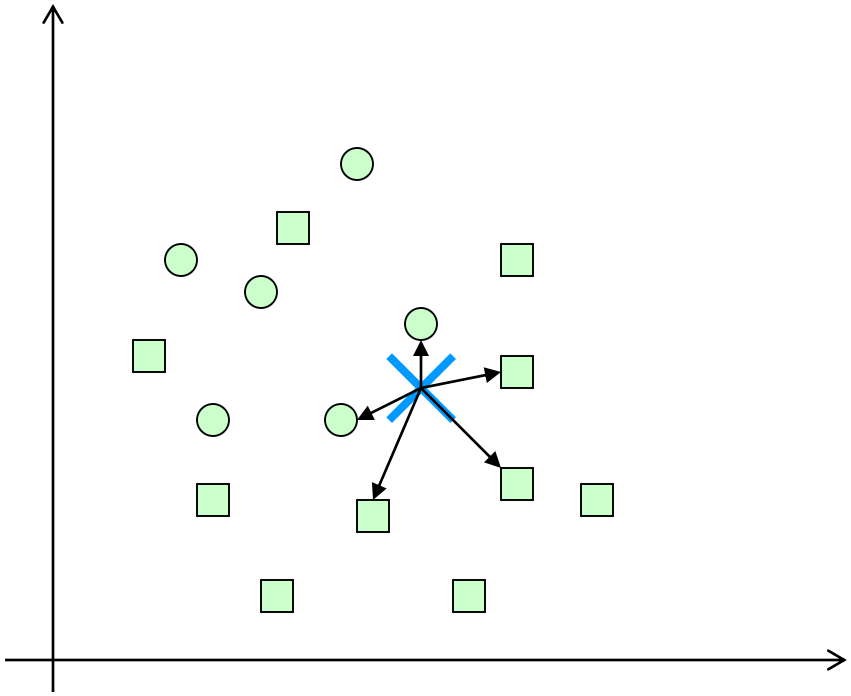
- On considère l'espace de n points de dimension p suivant :

$$X : \begin{array}{c|cccccc} & X^1 & \dots & X^j & \dots & X^p \\ X_1 & x_1^1 & \dots & x_1^j & \dots & x_1^p \\ \dots & \dots & \dots & \dots & \dots & \dots \\ X_i & x_i^1 & \dots & x_i^j & \dots & x_i^p \\ \dots & \dots & \dots & \dots & \dots & \dots \\ X_n & x_n^1 & \dots & x_n^j & \dots & x_n^p \end{array}$$

- A chaque point est associée une classe connue à l'avance
- Soit $X_T = [x_T^1, x_T^2, \dots, x_T^j, \dots, x_T^p]$ un point que l'on souhaite classifier
- On calcule toutes les distances entre le point X_T et les n points de l'espace
- On conserve les k points les plus proches de X_T
- La classe majoritaire dans l'ensemble de ces k points est attribuée à X_T

Algorithme des k plus proches voisins

- Exemple :



Si k=3 le nouveau point sera associé à la classe des cercles

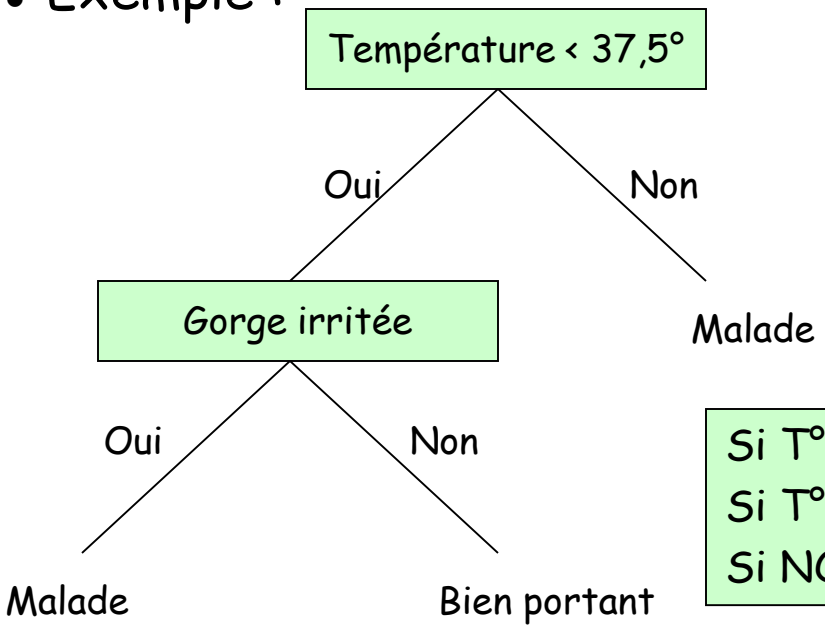
Si k=5 le nouveau point sera associé à la classe des carrés

Arbres de Décision

- Un arbre de décision est une représentation graphique d'une procédure de classification

- Un arbre de décision peut être traduit sous forme de règles de décision

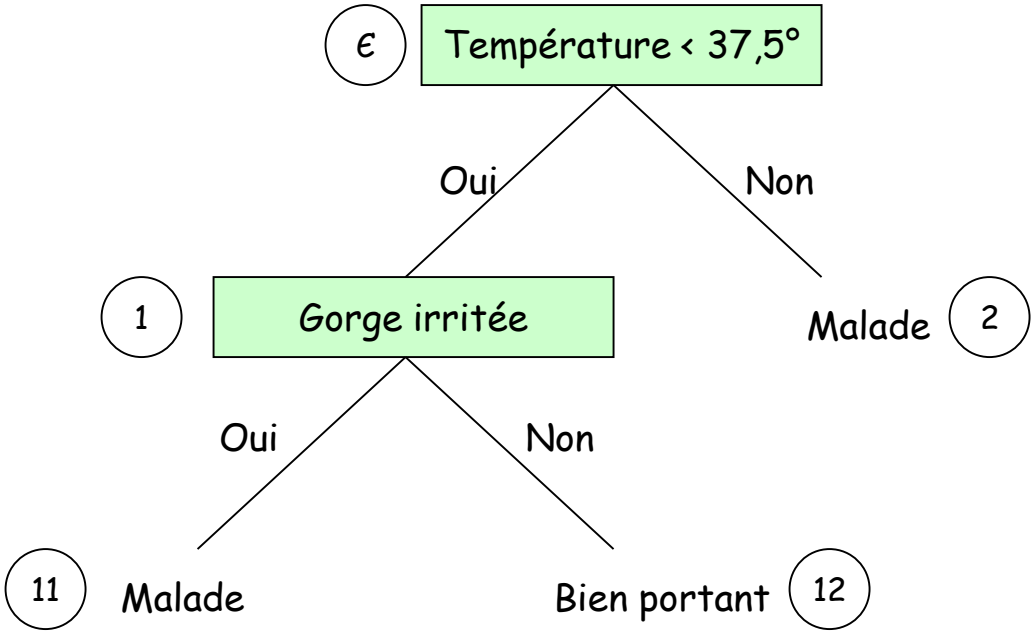
- Exemple :



Si T° < 37,5 ET Gorge irritée ALORS Malade
Si T° < 37,5 ET NON(Gorge irritée) ALORS Bien portant
Si NON(T° < 37,5) ALORS Malade

Arbres de Décision

- Un arbre de décision est un arbre au sens informatique du terme.
  - Chaque nœud interne test un caractère
  - Chaque branche correspond à une valeur d'un caractère
  - Chaque nœud feuille est une classe

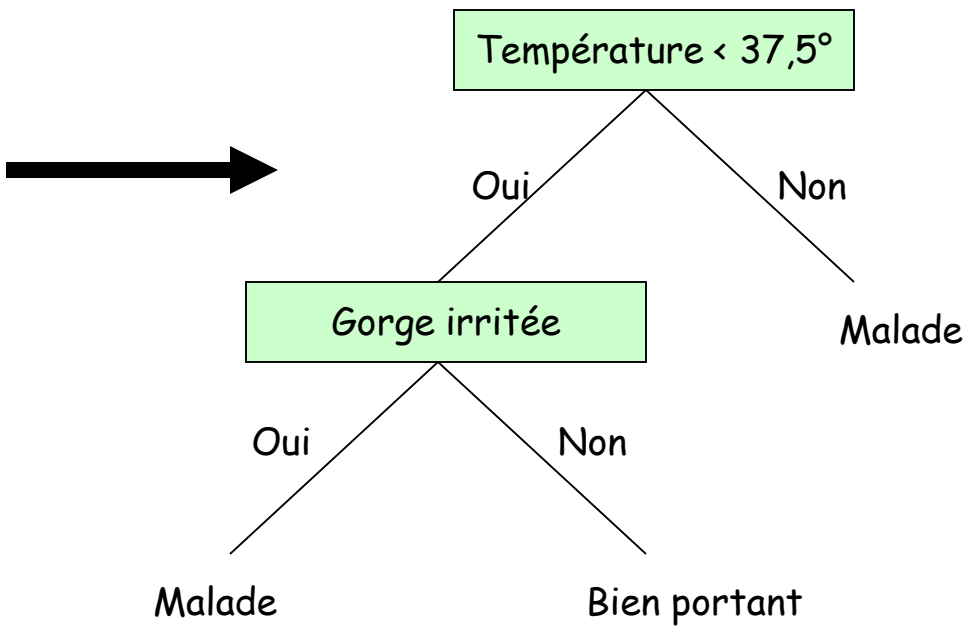


Arbres de Décision

- Comment généré automatiquement un arbre à partir de données ?

- Exemple :

Patients	T°	Gorge	Malade
Dupond	37,2	Normale	Non
Durand	38,5	Normale	Oui
⋮	⋮	⋮	⋮
Martin	37,2	Irritée	Oui



Arbres de Décision

- Notation

S : échantillon

$\{1, \cdots, c\}$: ensemble de classes

t : arbre de décision

p : position dans l'arbre

$N(p)$: cardinal de l'ensemble des exemples associé à p

$N(k/p)$: cardinal de l'ensemble des exemples associé à p et de classe k

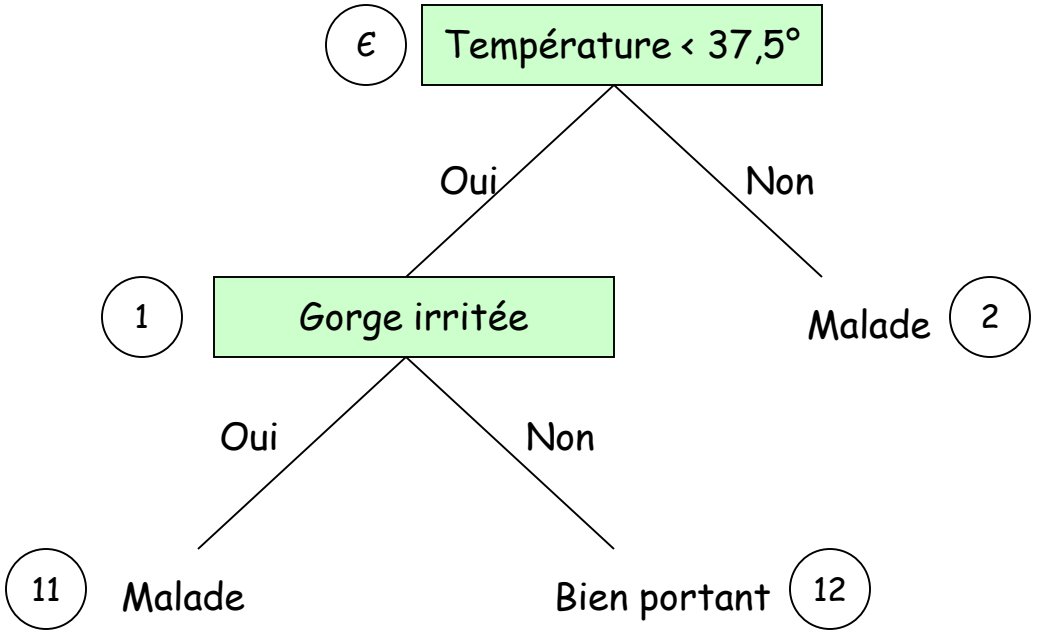
$P(k/p) = N(k/p)/N(p)$: proportion d'éléments de classe k à la position p

Arbres de Décision

- Exemple

	Gorge irritée	Gorge non irritée
T° < 37,5	(6 S, 37 M)	(91 S, 1 M)
T° ≥ 37,5	(2 S, 21 M)	(1 S, 41 M)

M : Malade
S : Bien portant



$$N(11) = 43$$
$$N(S/11) = 6$$
$$N(M/11) = 37$$
$$P(S/11) = N(S/11)/N(11) = 6/43$$
$$P(M/11) = N(M/11)/N(11) = 37/43$$

Arbres de Décision

- Exemple introductif à l'apprentissage automatique d'arbre de décision
- Données banque

	M	A	R	E	I
1	moyen	moyen	village	oui	oui
2	élevé	moyen	bourg	non	non
3	faible	âgé	bourg	non	non
4	faible	moyen	bourg	oui	oui
5	moyen	jeune	ville	oui	oui
6	élevé	âgé	ville	oui	non
7	moyen	âgé	ville	oui	non
8	faible	moyen	village	non	non

Moyenne du solde du compte :

$M = \{\text{élevé}, \text{moyen}, \text{faible}\}$

Tranche d'âge du client :

$A = \{\text{jeune}, \text{moyen}, \text{âgé}\}$

Type de localité de résidence du client :

$R = \{\text{village}, \text{bourg}, \text{ville}\}$

Niveau d'études supérieures du client :

$E = \{\text{oui}, \text{non}\}$

Consultation du compte par Internet :

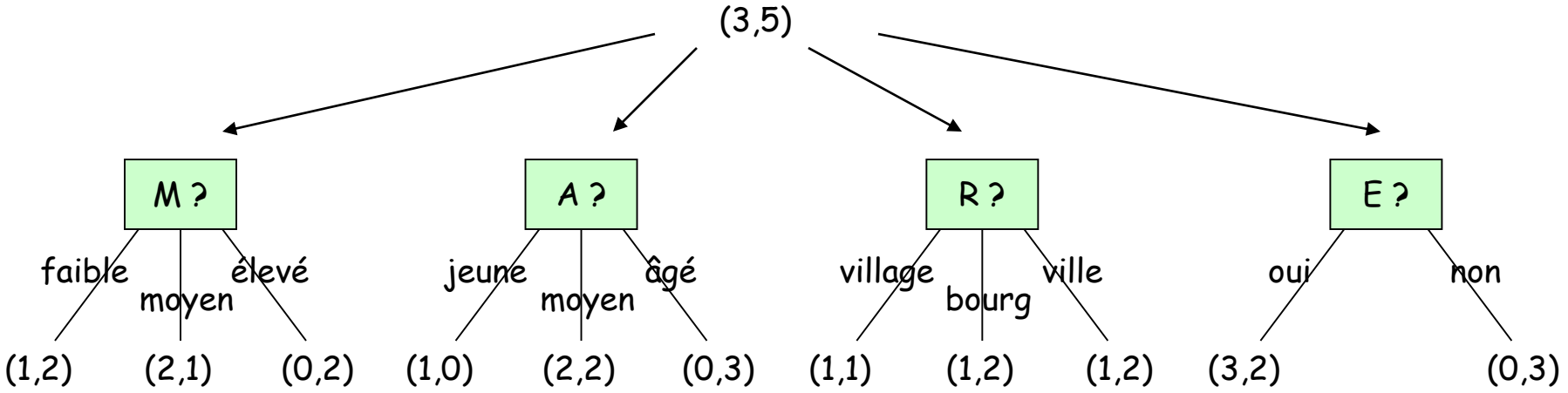
$I = \{\text{oui}, \text{non}\}$

- On souhaite construire un arbre de décision qui soit capable de déterminer si un client consultera son compte par Internet en fonction des attributs : M, A, R et E

Arbres de Décision

- Racine de l'arbre (pas de test)
  Etiquette (3,5) correspondant à : (N(Є,oui), N(Є,non))

- Quel est premier test à réaliser ?



Intuitivement :

Le test sur R n'est pas discriminatoire

Le test sur A est interessant sur les branches jeune et âgé

Arbres de Décision

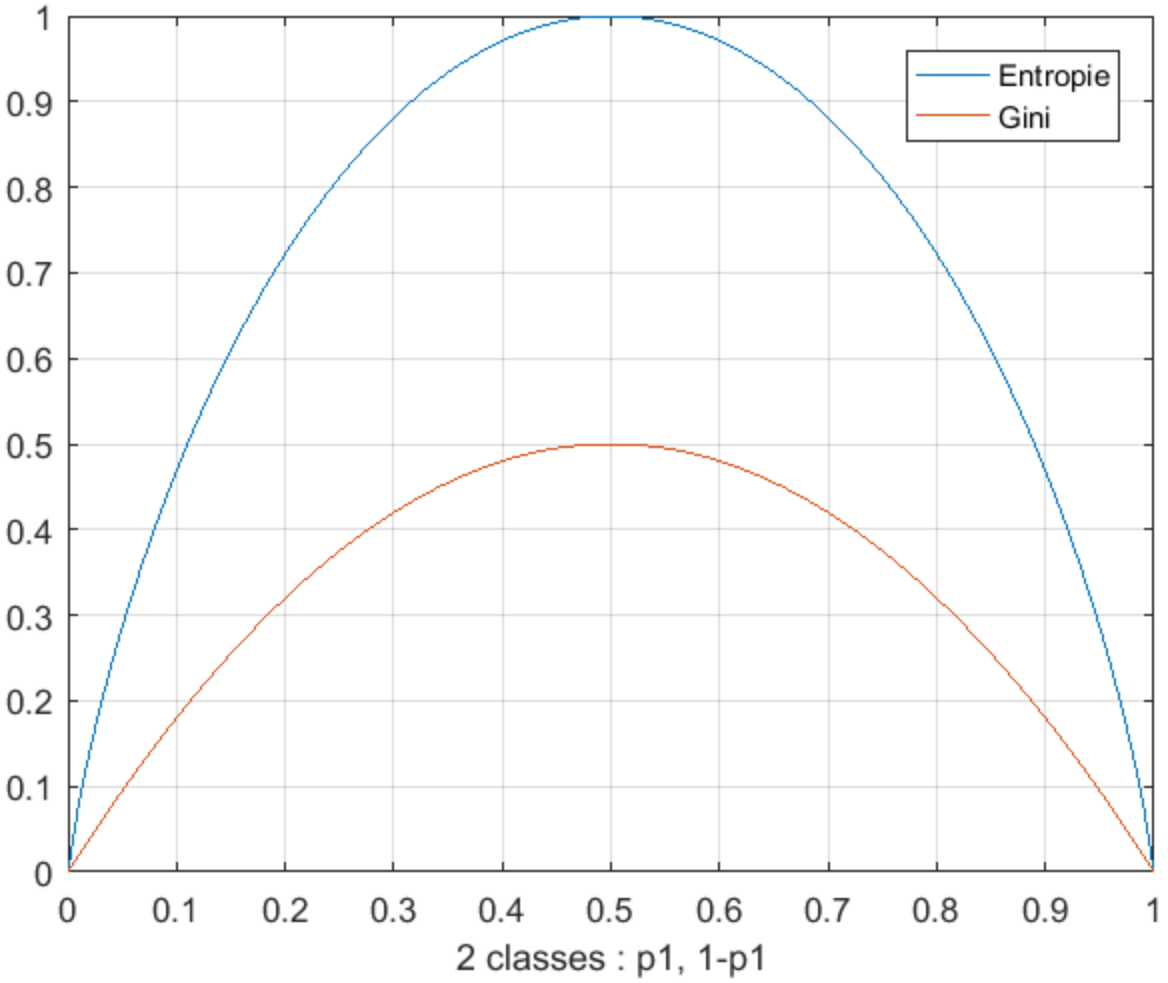
- Quelles fonctions permettraient de réprésenter ces intuitions ?

- Fonctions qui seraient :
  - Minimum lorsque le nœud est pur (tous les exemples sont dans une même classe
  - et Maximum lorsque les exemples sont équirépartis.

- Exemples de fonctions possédant ces propriétés :
  - Entropie : $Entropie(p) = -\sum_{k=1}^{c} P(k/p) \times \log(P(k/p))$
  - Fonction de Gini : $Gini(p) = 1 - \sum_{k=1}^{c} P(k/p)^2$

Arbres de Décision

- Quelles fonctions permettraient de réprésenter ces intuitions ?



Arbres de Décision

- Quelles fonction permettrait choisir un test ?
  - Fonction gain : $Gain(p, test) = i(p) - \sum_{j=1}^{n} P_j \times i(p_j)$

    p : position

    test : test d'arité n

    P_j : proportion d'éléments de S à la position p qui vont en position p_j

    $i(p)$: $Entropie(p)$ ou $Gini(p)$

  - Le test choisi est celui qui possède le gain le plus grand

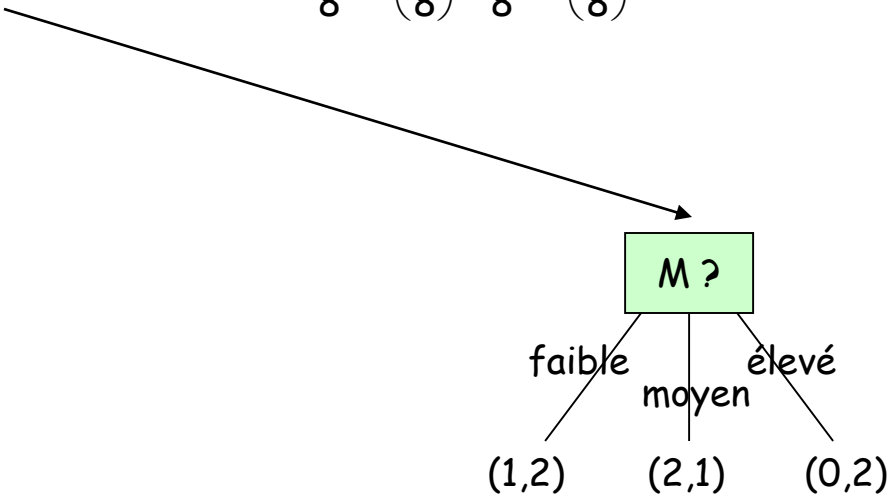
Arbres de Décision

- Exemple traité avec l'entropie :

$$Entropie(p) = -\sum_{k=1}^{c} P(k/p) \times \log\left(P(k/p)\right)$$

$$Gain(p, test) = i(p) - \sum_{j=1}^{n} P_j \times i(p_j)$$

(3,5) $Entropie(\in) = -\frac{3}{8}\log\left(\frac{3}{8}\right) - \frac{5}{8}\log\left(\frac{5}{8}\right) \approx 0{,}954$



$$Entropie(1) = -\frac{1}{3}\log\left(\frac{1}{3}\right) - \frac{2}{3}\log\left(\frac{2}{3}\right) \approx 0{,}918$$

$$Entropie(2) = -\frac{2}{3}\log\left(\frac{2}{3}\right) - \frac{1}{3}\log\left(\frac{1}{3}\right) \approx 0{,}918$$

$$Entropie(3) = -\frac{0}{2}\log\left(\frac{0}{2}\right) - \frac{2}{2}\log\left(\frac{2}{2}\right) = 0$$

$$Gain(\in, M) = Entropie(\in) - \left(\frac{3}{8}Entropie(1) + \frac{3}{8}Entropie(2) + \frac{2}{8}Entropie(3)\right)$$

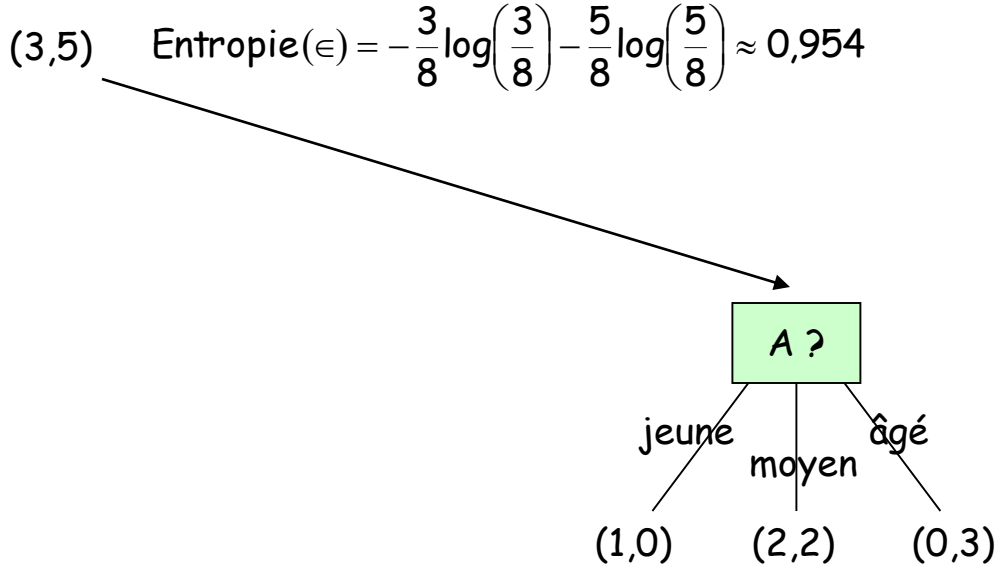
$$= Entropie(\in) - 0{,}688$$

Arbres de Décision

- Exemple traité avec l'entropie :

$$Entropie(p) = -\sum_{k=1}^{c} P(k/p) \times \log\left(P(k/p)\right)$$

$$Gain(p, test) = i(p) - \sum_{j=1}^{n} P_j \times i(p_j)$$



$$Entropie(1) = -\frac{1}{1}\log\left(\frac{1}{1}\right) - \frac{0}{1}\log\left(\frac{0}{1}\right) = 0$$

$$Entropie(2) = -\frac{2}{4}\log\left(\frac{2}{4}\right) - \frac{2}{4}\log\left(\frac{2}{4}\right) = 1$$

$$Entropie(3) = -\frac{0}{3}\log\left(\frac{0}{3}\right) - \frac{3}{3}\log\left(\frac{3}{3}\right) = 0$$

$$Gain(\in, A) = Entropie(\in) - \left(\frac{1}{8}Entropie(1) + \frac{4}{8}Entropie(2) + \frac{3}{8}Entropie(3)\right)$$

$$= Entropie(\in) - 0{,}5$$

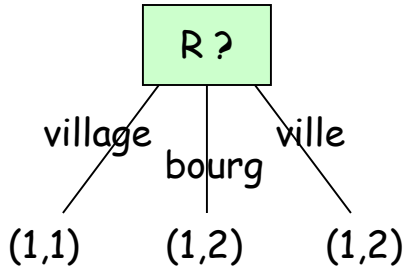
Arbres de Décision

- Exemple traité avec l'entropie :

$$Entropie(p) = -\sum_{k=1}^{c} P(k/p) \times \log\left(P(k/p)\right)$$

$$Gain(p, test) = i(p) - \sum_{j=1}^{n} P_j \times i(p_j)$$

(3,5) $\quad Entropie(\in) = -\frac{3}{8}\log\left(\frac{3}{8}\right) - \frac{5}{8}\log\left(\frac{5}{8}\right) \approx 0{,}954$



$$Entropie(1) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) = 1$$

$$Entropie(2) = -\frac{1}{3}\log\left(\frac{1}{3}\right) - \frac{2}{3}\log\left(\frac{2}{3}\right) \approx 0{,}918$$

$$Entropie(3) = -\frac{1}{3}\log\left(\frac{1}{3}\right) - \frac{2}{3}\log\left(\frac{2}{3}\right) \approx 0{,}918$$

$$Gain(\in, R) = Entropie(\in) - \left(\frac{2}{8}Entropie(1) + \frac{3}{8}Entropie(2) + \frac{3}{8}Entropie(3)\right)$$

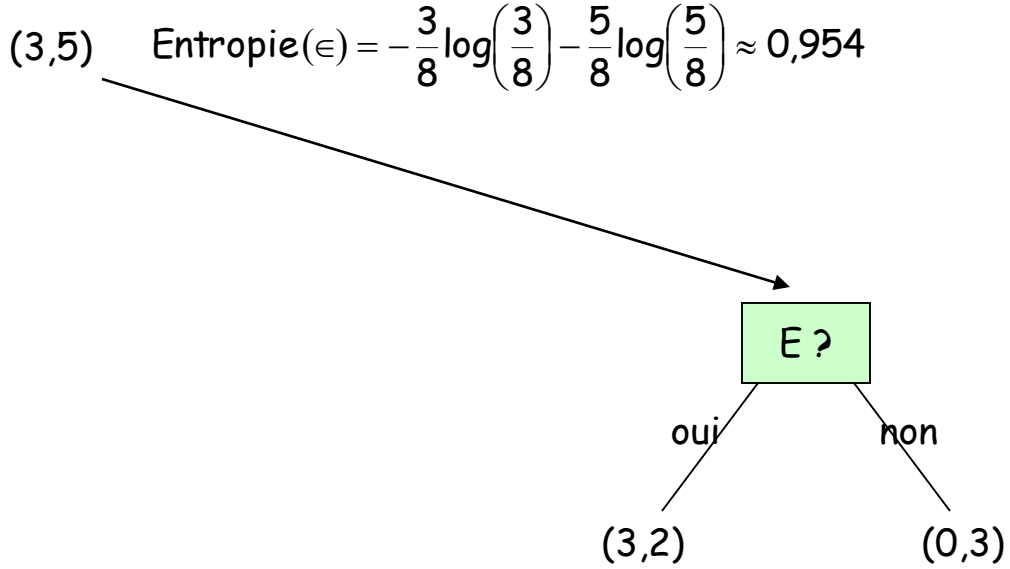
$$= Entropie(\in) - 0{,}938$$

Arbres de Décision

- Exemple traité avec l'entropie :

$$Entropie(p) = -\sum_{k=1}^{c} P(k/p) \times \log\left(P(k/p)\right)$$

$$Gain(p, test) = i(p) - \sum_{j=1}^{n} P_j \times i(p_j)$$



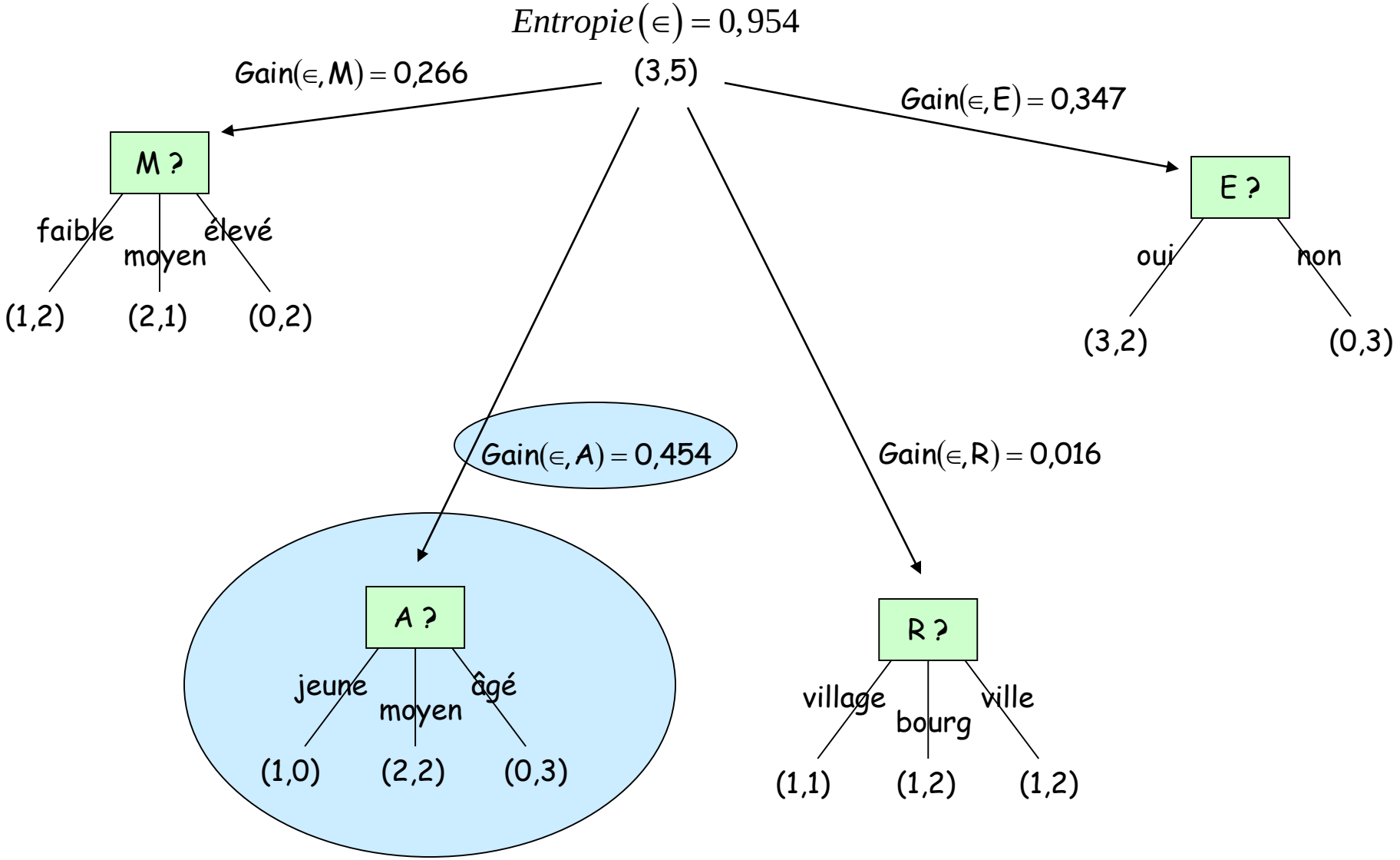
$$Entropie(1) = -\frac{3}{5}\log\left(\frac{3}{5}\right) - \frac{2}{5}\log\left(\frac{2}{5}\right) \approx 0.97$$

$$Entropie(2) = -\frac{0}{3}\log\left(\frac{0}{3}\right) - \frac{3}{3}\log\left(\frac{3}{3}\right) = 0$$

$$Gain(\in, E) = Entropie(\in) - \left(\frac{5}{8}Entropie(1) + \frac{3}{8}Entropie(2)\right) = Entropie(\in) - 0,607$$

Arbres de Décision

- Exemple traité avec l'entropie :



Arbres de Décision

- Algorithme générique de construction d'un arbre

entrée : Caractères; échantillon S
début
    Initialiser à l'arbre vide; la racine est le nœud courant
    répéter
        Décider si le nœud courant est terminal
        si le nœud est terminal **alors**
            Affecter une classe
        sinon
            Sélectionner un test et créer le sous-arbre
        finsi
            Passer au nœud suivant non exploré si il en existe
    jusqu'à obtenir un arbre de décision
    fin

Arbres de Décision

- Algorithme ID3

Un nœud p est terminal si : tous les éléments associés à ce nœud sont dans une même classe ou si aucun test n'a pu être séléctionné

On choisit le test qui maximise :

$$Gain(p, test) = Entropie(p) - \sum_{j=1}^{n} P_j \times Entropie(p_j)$$

On attribut la classe majoritaire à une feuille

Arbres de Décision : Exemple

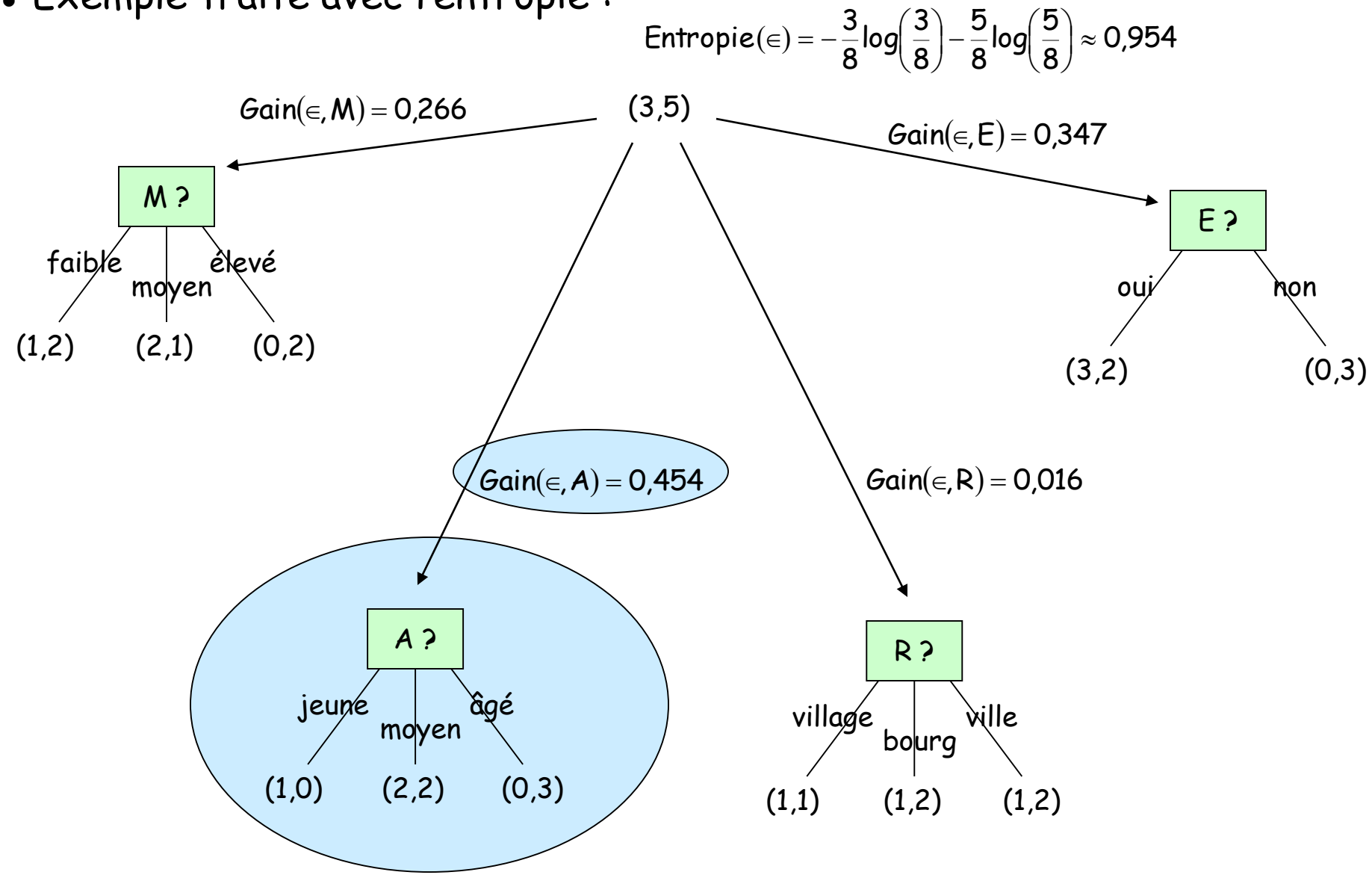
```
Fonction ID3 (I,O,T)
I ensemble des attributs d'entrée
O attribut de sortie
T ensemble des individus d'apprentissage

Si (T est vide)
          Renvoyer Erreur

Si (tous les individus de T appartiennent à la même classe)
          Renvoyer un nœud avec le label de la classe

Si (I est vide)
          Renvoyer un nœud avec le label le plus fréquent sur l'attribut de sortie de T

Calculer le gain d'information pour tous les attributs de I relativement à T

X est l'attribut avec le plus grand gain

{x_j / j=1,2,…,m} sont les valeurs de X

{T_j / j=1,2,…,m} sont les sous ensembles de T décomposé par rapport aux x_j

Renvoyer un arbre avec X comme label du nœud racine
          et x_1, x_2, …, x_m comme labels des arcs allant aux arbres
          ID3(I-{X},O,T_1), ID3(I-{X},O,T_2), …, ID3(I-{X},O,T_m)
```

Arbres de Décision : Exemple

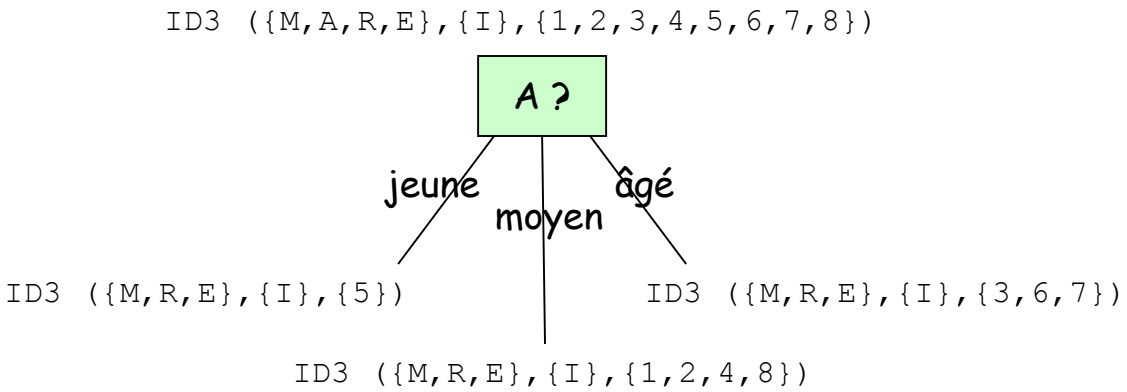
```
ID3 ({M,A,R,E},{I},{1,2,3,4,5,6,7,8})
```

Arbres de Décision : Exemple

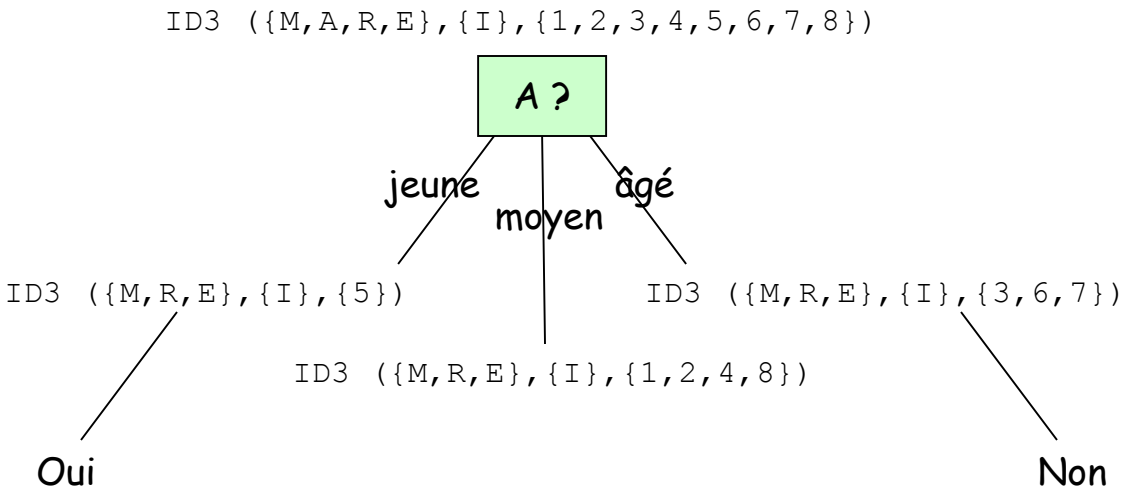
- Exemple traité avec l'entropie :



Arbres de Décision : Exemple

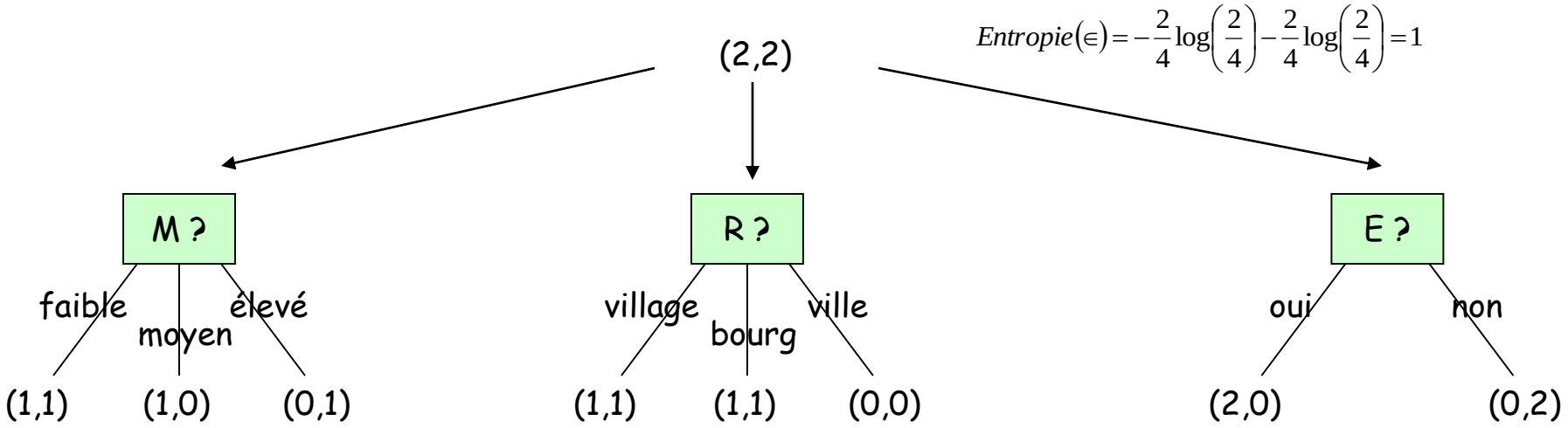


Arbres de Décision : Exemple



	M	R	E	I
1	moyen	village	oui	oui
2	élevé	bourg	non	non
4	faible	bourg	oui	oui
8	faible	village	non	non

Arbres de Décision : Exemple



$$Entropie(\in) = -\frac{2}{4}\log\left(\frac{2}{4}\right) - \frac{2}{4}\log\left(\frac{2}{4}\right) = 1$$

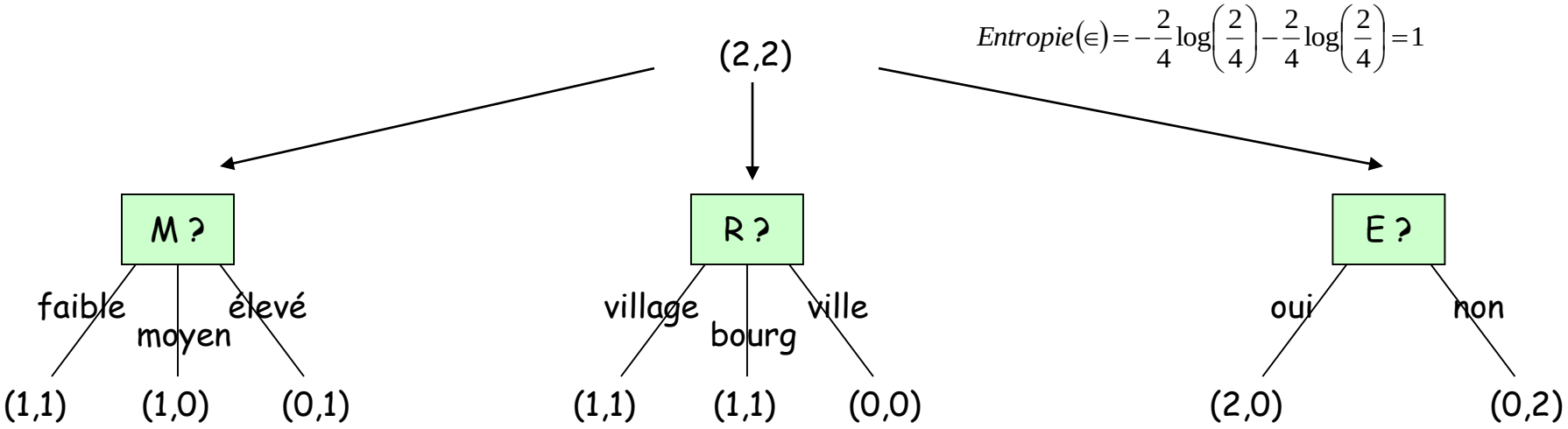
$$Entropie(1) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) = 1$$

$$Entropie(2) = -\frac{1}{1}\log\left(\frac{1}{1}\right) - \frac{0}{1}\log\left(\frac{0}{1}\right) = 0$$

$$Entropie(3) = -\frac{0}{1}\log\left(\frac{0}{1}\right) - \frac{1}{1}\log\left(\frac{1}{1}\right) = 0$$

$$Gain(\in, M) = Entropie(\in) - \left(\frac{2}{4}Entropie(1) + \frac{1}{4}Entropie(2) + \frac{1}{4}Entropie(3)\right) = Entropie(\in) - 0{,}5$$

Arbres de Décision : Exemple



$$Entropie(\in) = -\frac{2}{4}\log\left(\frac{2}{4}\right) - \frac{2}{4}\log\left(\frac{2}{4}\right) = 1$$

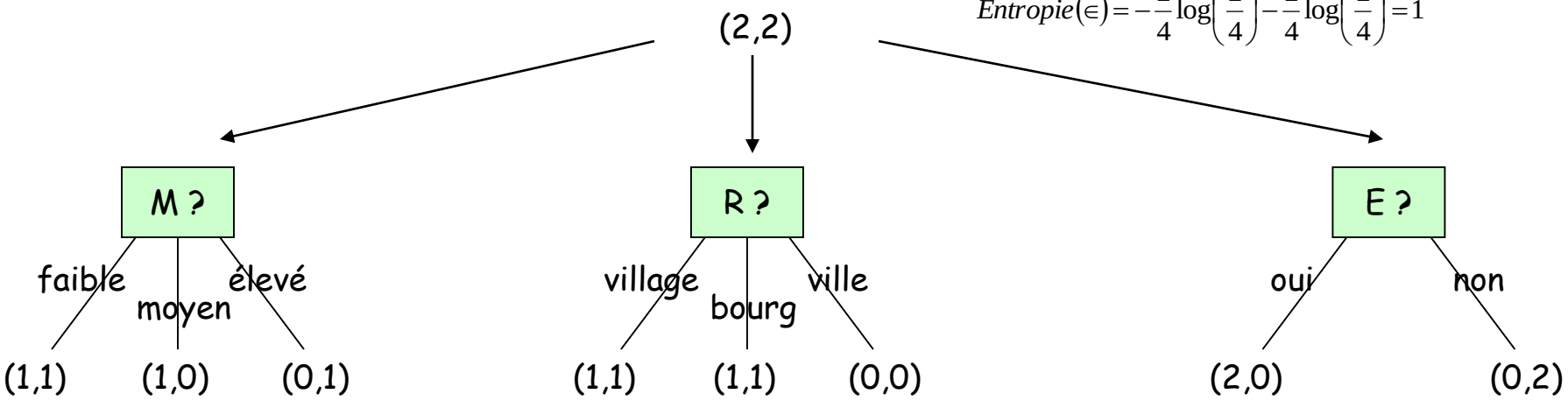
$$Entropie(1) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) = 1$$

$$Entropie(2) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) = 1$$

$$Entropie(3) = -\frac{0}{0}\log\left(\frac{0}{0}\right) - \frac{0}{0}\log\left(\frac{0}{0}\right) = 0$$

$$Gain(\in, R) = Entropie(\in) - \left(\frac{2}{4}Entropie(1) + \frac{2}{4}Entropie(2) + \frac{0}{4}Entropie(3)\right) = Entropie(\in) - 1$$

Arbres de Décision : Exemple



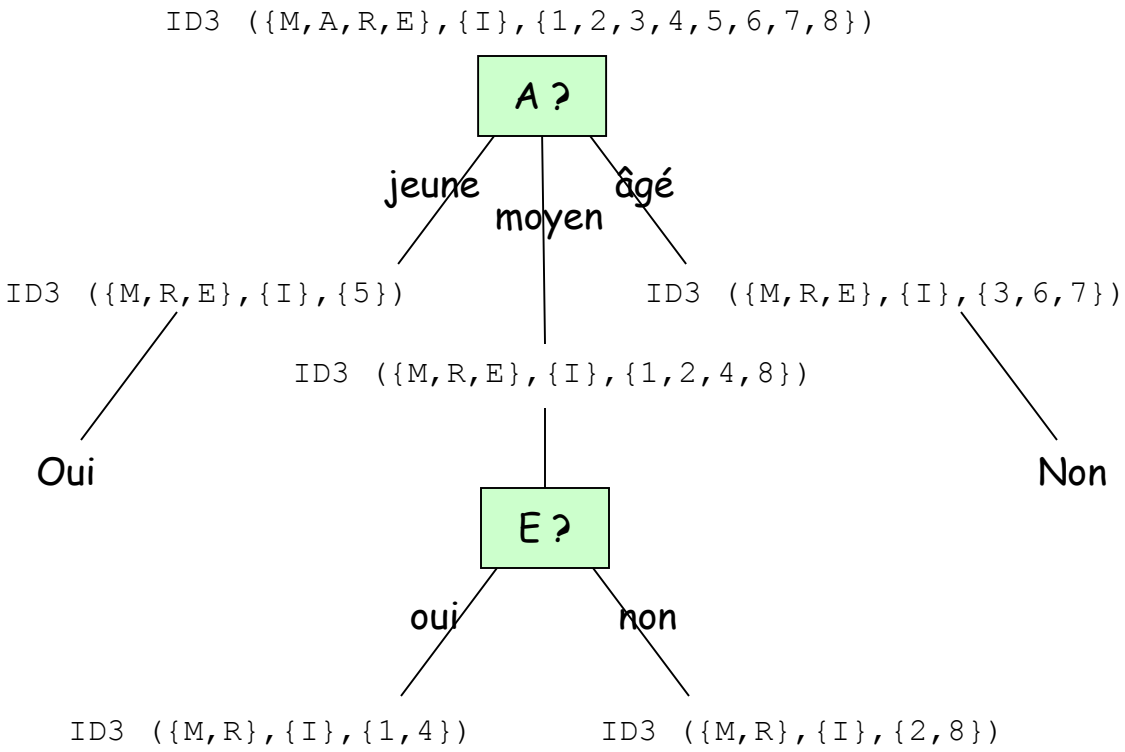
$$Entropie(\in) = -\frac{2}{4}\log\left(\frac{2}{4}\right) - \frac{2}{4}\log\left(\frac{2}{4}\right) = 1$$

$$Entropie(1) = -\frac{2}{2}\log\left(\frac{2}{2}\right) - \frac{0}{2}\log\left(\frac{0}{2}\right) = 0$$

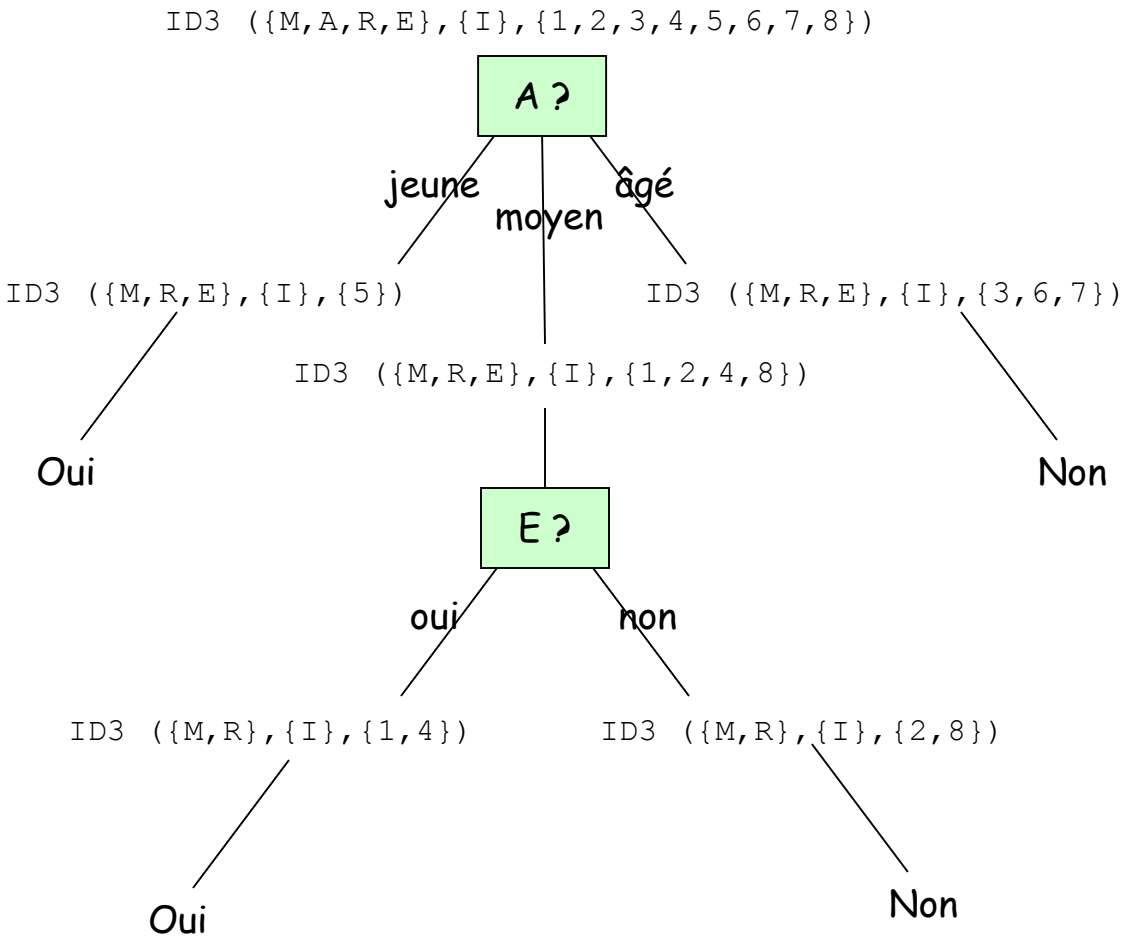
$$Entropie(2) = -\frac{0}{2}\log\left(\frac{0}{2}\right) - \frac{2}{2}\log\left(\frac{2}{2}\right) = 0$$

$$Gain(\in, E) = Entropie(\in) - \left(\frac{2}{4}Entropie(1) + \frac{2}{4}Entropie(2)\right) = Entropie(\in) - 0$$

Arbres de Décision : Exemple

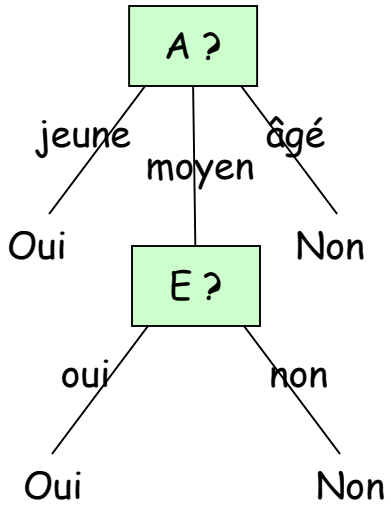


Arbres de Décision : Exemple



Arbres de Décision : Exemple

- Arbre final



Arbres de Décision : Exercice

- Peut-on jouer au tennis ?

	Outlook	Temperature	Humidity	Windy	Play
1	sunny	hot	high	false	no
2	sunny	hot	high	true	no
3	overcast	hot	high	false	yes
4	rainy	mild	high	false	yes
5	rainy	cool	normal	false	yes
6	rainy	cool	normal	true	no
7	overcast	cool	normal	true	yes
8	sunny	mild	high	false	no
9	sunny	cool	normal	false	yes
10	rainy	mild	normal	false	yes
11	sunny	mild	normal	true	yes
12	overcast	mild	high	true	yes
13	overcast	hot	normal	false	yes
14	rainy	mild	high	true	no

Sources

- Denis, François et Gilleron, Rémi.
  http://www.grappa.univ-lille3.fr/polys/apprentissage/index.html.

- Tommassi, Marc et Gilleron, Rémi.
  http://www.grappa.univ-lille3.fr/polys/fouille/index.html.

Bibliographie

- Statistiques et analyses de données
  - « Probabilité, analyse des données et statistique », G. Saporta, éditions Technip
  - « Analyse des données », M. Volle, Economica
  - « Le data mining », René Lefébure et Gilles Venturi, Eurolles
  - Cornuéjols A. & Miclet L. (2010) Apprentissage artificiel Concepts et algorithmes. - Editeur : Eyrolles - Collection : Algorithmes - Nombre de pages : 804 pages - Date de parution : 03/06/2010 (2e édition) - EAN13 : 9782212124712