

MACHINE LEARNING

Chapitre 2 : Apprentissage supervisé

Ali Ben Mrad

Classification

2

- Elle permet de **prédire** si un élément est membre d'un groupe ou d'une catégorie donnée.

- **Classes**
 - ▣ Identification de groupes avec des profils particulier
 - ▣ 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

3

- Accord de crédit
- Marketing ciblé
- Diagnostic médical
- Analyse de l'effet d'un traitement
- Détection de fraudes fiscales
- etc

Classification - Applications

4

Classification de documents

E-mails en spam, shopping, travail, ...

[Supprimer tous les spams maintenant](#) (les messages se trouvant dans le dossier Spam depuis plus de 30 jours sont automatiquement supprimés)

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Tatiana	Re: Para os homens - Val lhe interessar muito!	01:50
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	comebuy	Téléphones les plus compétitifs de Comebuy	22:38
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Francois	100 raisons de jouer sur Majestic	27 janv.
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Fund Investigation Bureau	TREAT AS URGENT RIGHT AWAY	27 janv.
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Mrs Elizabeth Johnson	Hello My Beloved One.	27 janv.
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Evellyn	Re: Amigo, não está satisfeito com o tamanho? Isto pode te ajudar!	27 janv.
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Amanda, Amanda (2)	Re: Amigo, o que vc faria com 10cm a mais?	26 janv.
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Groupe Partouche	Et encore un gagnant au Megapot !	26 janv.
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Carli, Joshua Daniel	N/A	26 janv.
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	RCH Tournai	Votre Semaine avec 100000 en Tout	26 janv.
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Jemmy Klamet	Nicolas Baskiotis F-E..L-L..N G.._H O..R N-Y?-_-G-E-T _L_A_I_D_-_N_O_W !	26 janv.
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Jean-Pierre	Les meilleurs casinos pour les joueurs français	25 janv.

Principale Réseaux sociaux Promotions +

<input checked="" type="checkbox"/>	<input type="checkbox"/>	CollierPrenom	Annance ⓘ	Spécial St Valentin - 3 Jours Seulement - 15% de Réduction !	×
<input checked="" type="checkbox"/>	<input type="checkbox"/>	SoftLayer.com	Annance ⓘ	Get a Secure Cloud - We've secured the public cloud with private servers, private networks, and full private clouds.	×
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Booking.com	Last-minute deals for Montréal and London. Get them before they're gone!	28/12/2014
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Voyages-sncf.com	DERNIERE MINUTE NOUVEL AN : profitez des meilleurs prix !	26/12/2014
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Impossible	Year's End Clearance - Up to 20% off Film and Accessories	26/12/2014
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Booking.com	Nicolas - you qualify for at least 20% off places to stay	26/12/2014
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Communauté d'entraide Gr.	Nicolas, des questions sur vos produits ?	25/12/2014
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Polymarket	Secure Mail : votre sécurité et la nôtre	25/12/2014

gmail.com

Classification - Applications














5

Reconnaissance de chiffres

8 2 9 4 4 6 4 9 7 0 9 2 9 5 1 5 9 1 0 3
 1 3 5 9 1 7 6 2 8 2 2 5 0 7 4 9 7 8 3 2
 1 1 8 3 6 1 0 3 1 0 0 1 1 2 7 3 0 4 6 5
 2 6 4 7 1 8 9 9 3 0 7 1 0 2 0 3 5 4 6 5

Ou de captcha

[Yann et al. 08], Newcastle University

Characters under typical distortions					Recognition rate
					~100%
					96+%
					100%
					98%
					~100%
					95+%

Classification - Applications

6

□ En image

Détection de visages

(opencv)



Classification - Applications

7

Classification et organisation automatique



Classification - Applications

8

Détection d'objets

teradeep.com, Purdue University



Tracking

[Fragkiadaki et al. 12], Pennsylvania University



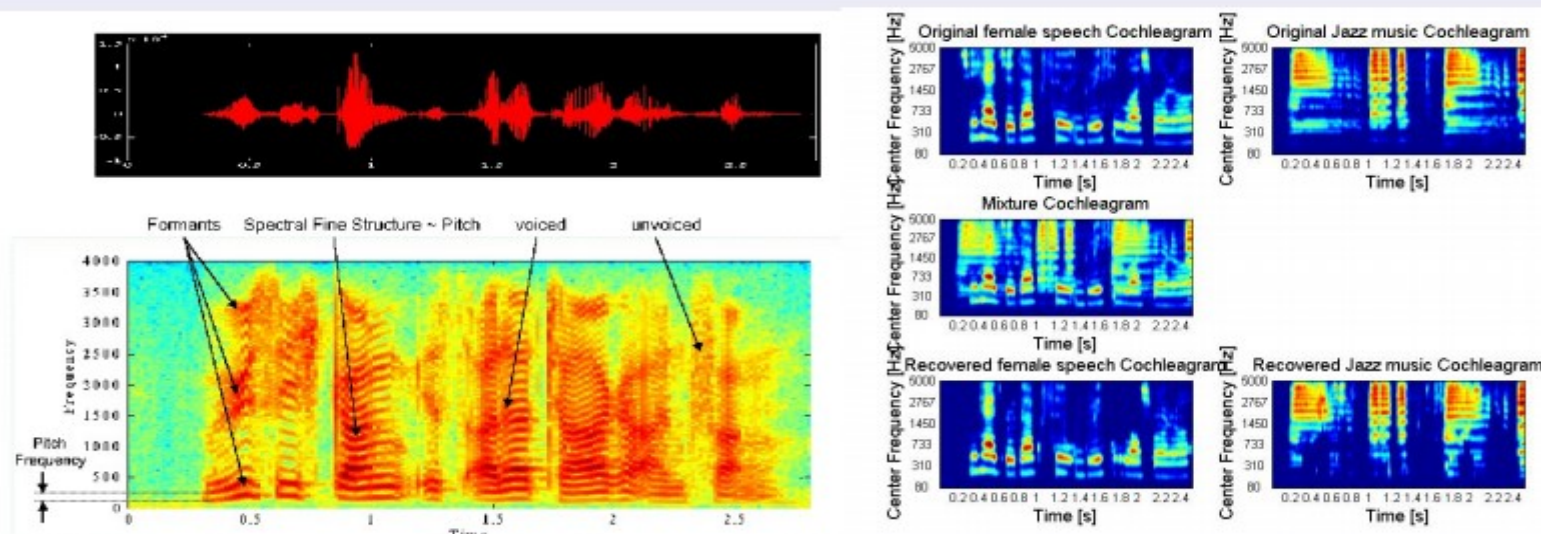
Classification - Applications

9

□ En audio

Reconnaissance de la parole, séparation de sources

<http://markus-hauenstein.de>



- Mais aussi débruitage, transcription musicale, reconnaissance du locuteur, ' classification/identification de musiques. . .

Classification - Applications

10

□ Systemes de recommandation

De musiques, de films, de produits, d'amis

The collage consists of four distinct screenshots demonstrating various recommendation systems:

- Similar Artists:** A web interface showing a list of artists recommended for Bob Dylan. The list includes: 1 Bob Dylan, 2 Radiohead (highlighted with a red arrow), 3 Led Zeppelin, 3 The Rolling Stones, 5 Pink Floyd, 6 David Bowie, 7 The Who, and 8 John Lennon. To the left of the list are small album covers for each artist.
- Amazon.com:** A screenshot of the Amazon website's 'Recommended for You' section. It features a header with the Amazon logo and the text 'Recommended for You'. Below this, a message states: 'Amazon.com has new recommendations for you based on books you purchased or told us you own.' Four book covers are displayed: 'The Little Big Things: 163 Ways to Pursue EXCELLENCE', 'Fascinate! Your 7 Triggers to Persuasion and Captivation', 'Sherlock Holmes [Blu-ray]', and 'Alice in Wonderland [Blu-ray]'.
- Facebook:** A screenshot of a Facebook social network graph. It shows a dense cluster of numerous small circular profile pictures of users, connected by a network of lines, representing social connections.
- Movie Recommendation Engine:** A screenshot of a web interface titled 'Recommendation Engine'. It features a search bar at the top and a grid of movie posters below. The movies shown include 'Umbrellas of Cherbourg, The', 'Brokedown Palace', 'West Beirut', 'Suspect', 'Heights', and 'Babylon A.D.'. Each poster has a 'Select' button underneath it.

Processus à deux étapes

11

- **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

Etape 1 : Construction du modèle

12

- **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, ...

Etape 2 : Utilisation du modèle

13

- **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)

14

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):

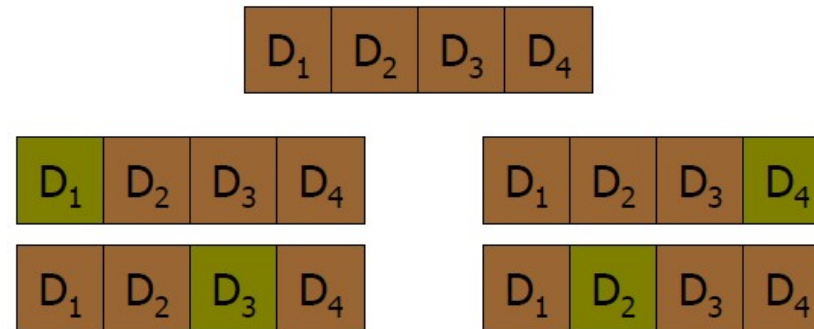
Apprentissage D_t

Validation D/D_t

Validation de la Classification (accuracy)

15

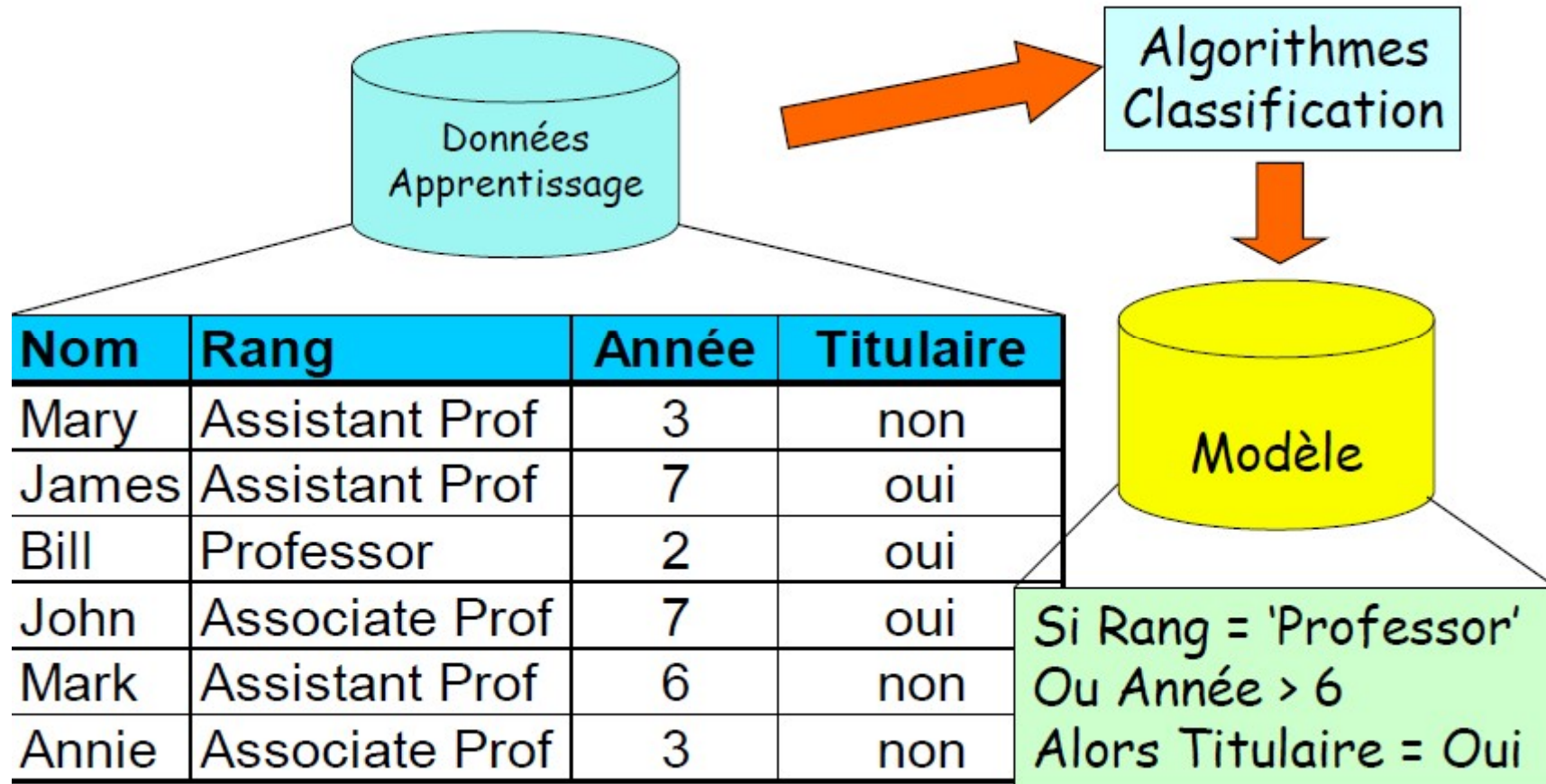
- **Validation croisée** (ensemble de données modéré)
 - Diviser les données en k sous-ensembles
 - Utiliser $k-1$ sous-ensembles comme données d'apprentissage et un sous-ensemble comme données test



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

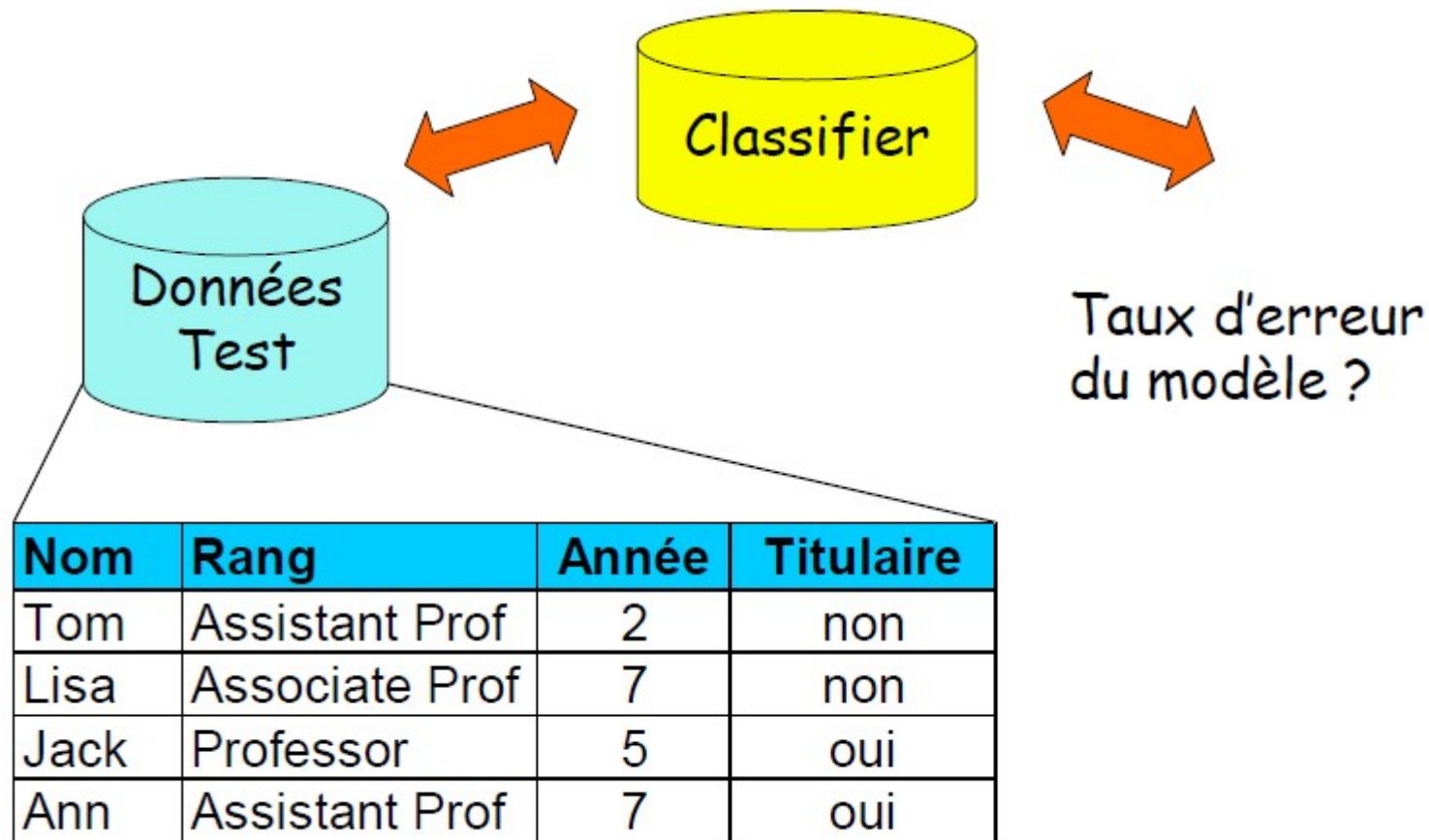
Exemple : Construction du modèle

16



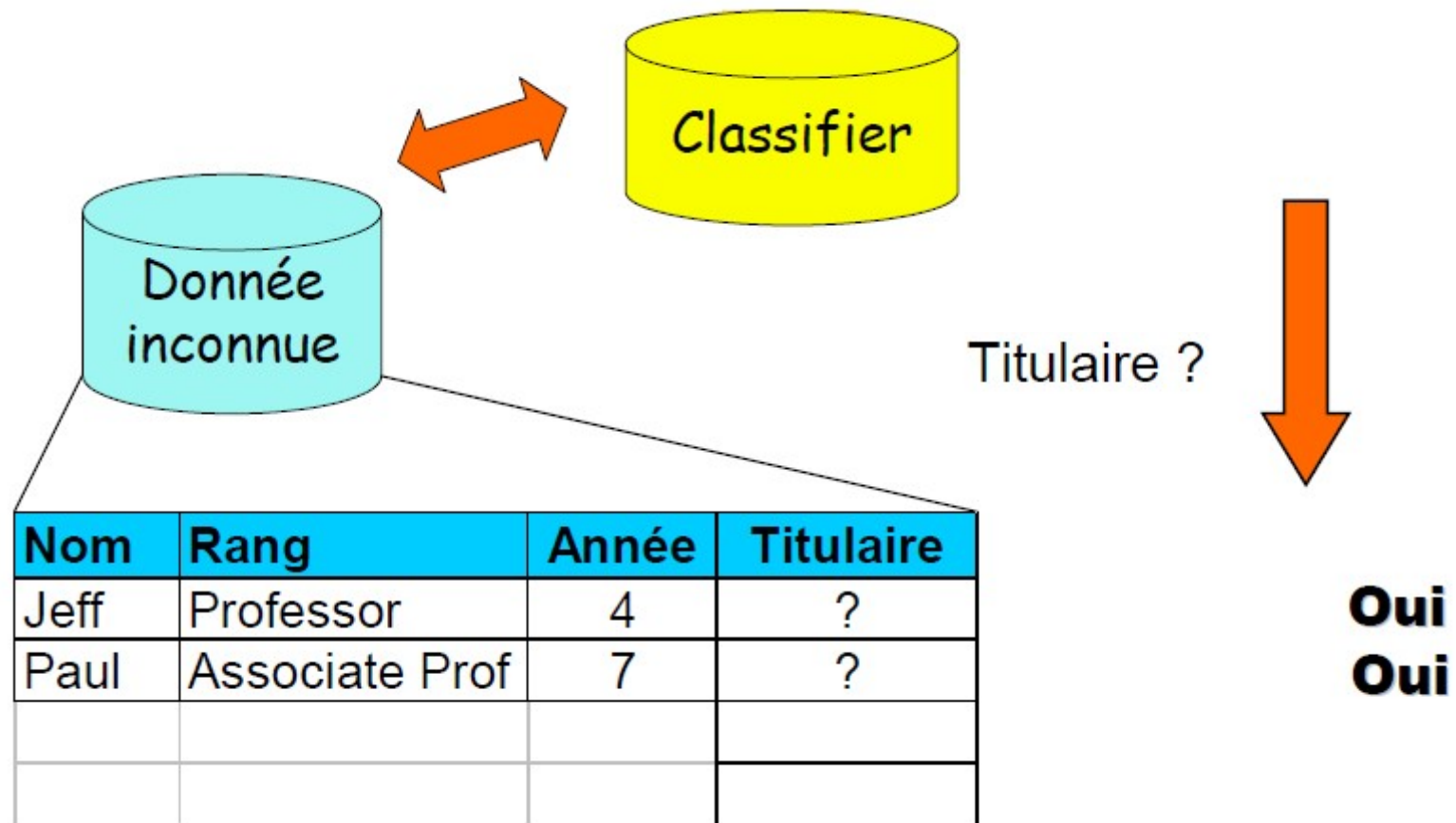
Exemple : Utilisation du modèle

17



Exemple : Utilisation du modèle

18



Evaluation des méthodes de classification

19

- **Taux d'erreur** (Accuracy)
- **Temps d'exécution** (construction, utilisation)
- **Robustesse** (bruit, données manquantes,...)
- **Extensibilité**
- **Interprétabilité**
- **Simplicité**

Méthodes de Classification

20

- **Méthode K-NN** (plus proche voisin)
- **Arbres de décision**
- **Machines à vecteurs supports** (SVM)
- **Classification bayésienne**
- **Réseaux de neurones**
- ...

- **Caractéristiques**
 - ▣ **Apprentissage supervisé** (classes connues)

21

La méthode des plus proches voisins

(KNN : K-nearest neighbors))

Méthode des plus proches voisins

22

- **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.
- **Modèle** = échantillon d'apprentissage + fonction de distance + fonction de choix de la classe en fonction des classes des voisins les plus proches.

Algorithme KNN (K-nearest neighbors)

23

- **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
(faire voter les voisins de y)

- **Sortie** : la classe de y est $c(y) = C$

Algorithme KNN : sélection de la classe

24

- **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.

Principe de l'algo. KNN

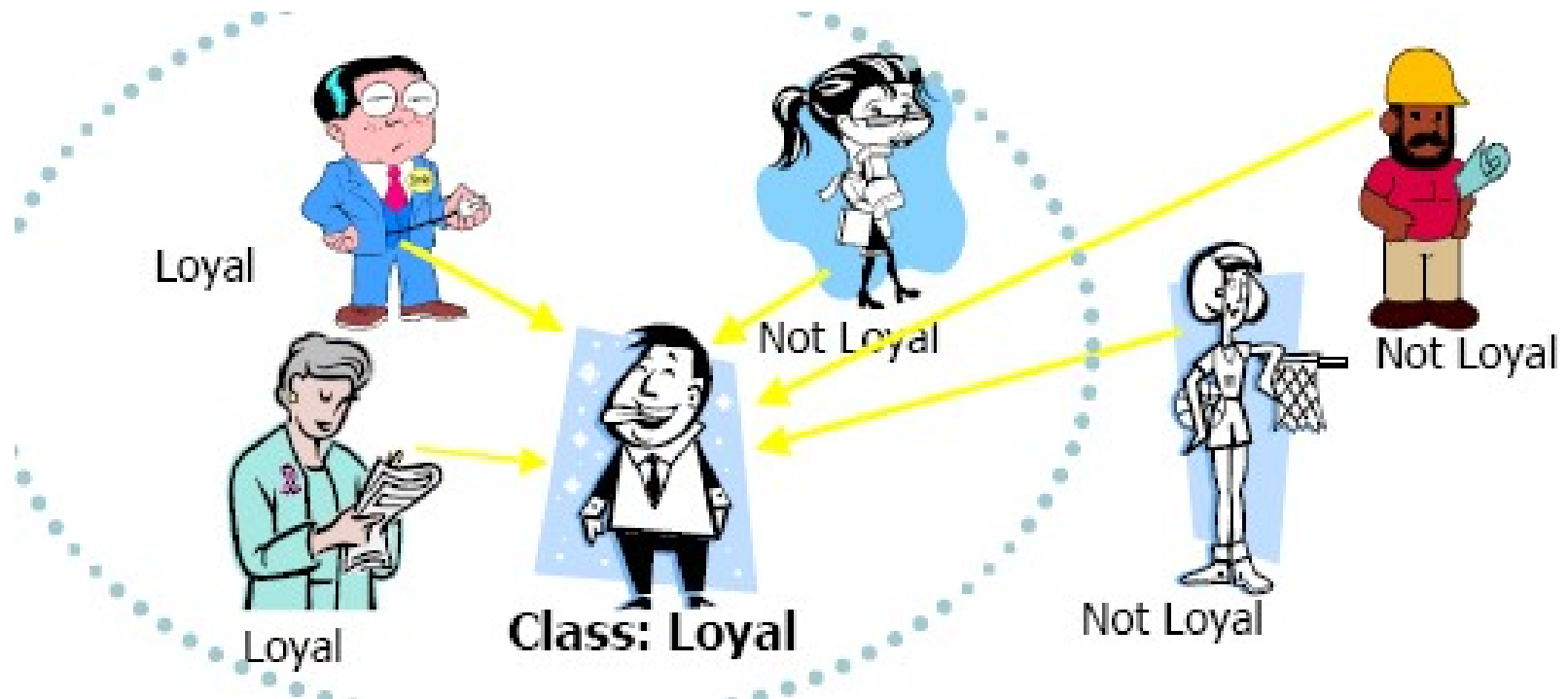
25

1. on stocke les exemples tels quels dans une table ;
2. pour prédire la classe d'une donnée, on détermine les exemples qui en sont le plus proche ;
3. de ces exemples, on déduit la classe ou on estime l'attribut manquant de la donnée considérée.

Exemple: Client loyal ou non

26

$K = 3$



Distance

27

- Le **choix** de la distance est primordial au bon fonctionnement de la méthode
- Les distances les plus simples permettent d'obtenir des résultats satisfaisants (lorsque c'est possible)
- **Propriétés de la distance:**
 - ▣ $d(A, A) = 0$
 - ▣ $d(A, B) = d(B, A)$
 - ▣ $d(A, B) \leq d(A, C) + d(B, C)$

Distance entre numériques

28

$$\blacksquare d(x, y) = |x - y|$$

ou

$$\blacksquare d(x, y) = |x - y| / d_{\max}, \text{ où } d_{\max} \text{ est la distance maximale entre deux numériques du domaine considéré}$$

Distance entre nominaux

29

- **Données binaires** : 0 ou 1. On choisit $d(0, 0) = d(1, 1) = 0$ et $d(0, 1) = d(1, 0) = 1$.
- **Données énumératives** : la distance vaut 0 si les valeurs sont égales et 1 sinon.
- **Données énumératives ordonnées** : elles peuvent être considérées comme des valeurs énumératives mais on peut également définir une distance utilisant la relation d'ordre.
 - Exemple: Si un champ prend les valeurs A, B, C, D et E, on peut définir la distance en considérant 5 points de l'intervalle $[0, 1]$ avec une distance de 0,2 entre deux points successifs, on a alors $d(A, B) = 0,2$; $d(A, C) = 0,4$; ... ; $d(D, E) = 0,2$.

Distance Euclidienne entre 2 exemples

30

- Soit $X = (x_1, \dots, x_n)$ et $Y = (y_1, \dots, y_n)$ deux exemples, la distance euclidienne entre X et Y est:

$$D(X, Y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

Autres distances

31

Sommation:

$$D(X, Y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

Distance euclidienne pondérée:

$$D(X, Y) = \sqrt{\sum_{i=1}^n w_i (x_i - y_i)^2}$$

Pourquoi pondérer les attributs?

32

- Certains attributs peuvent dominer le calcul de la distance
- Exemple:

$$\text{Distance (John, Rachel)} = \text{sqrt} [(35-22)^2 + (35,000-50,000)^2 + (3-2)^2]$$

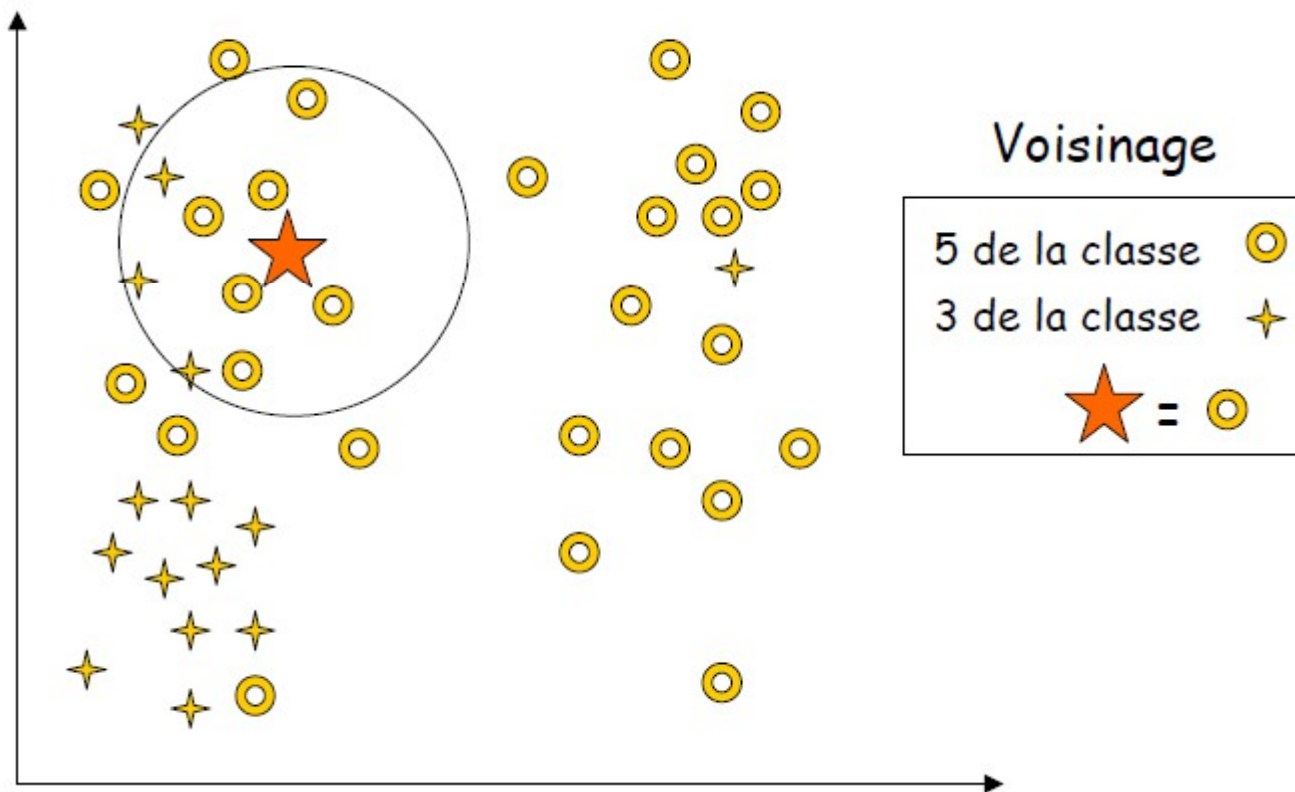
Choix de la classe

33

- Choix de la classe majoritaire
- Choix de la classe majoritaire pondérée
 - ▣ Chaque classe d'un des k voisins sélectionnés est pondéré
 - ▣ Soit V le voisin considéré. Le poids de $c(V)$ est inversement proportionnel à la distance entre l'enregistrement Y à classer et V
- Calculs d'erreur

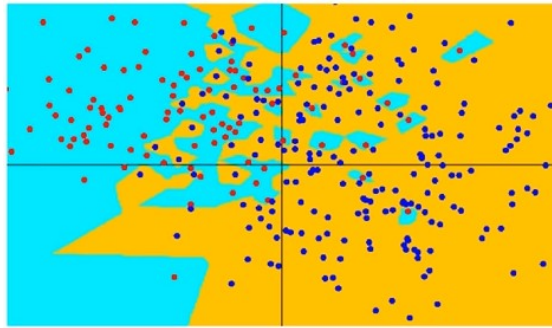
Illustration

34

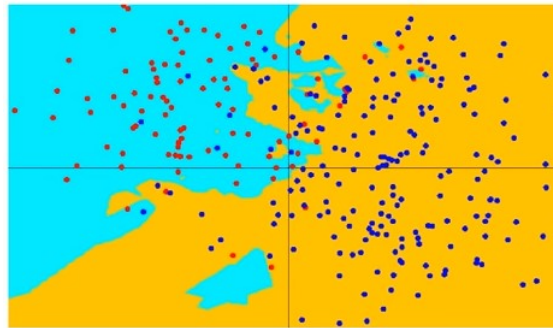


Effet de la valeur de K

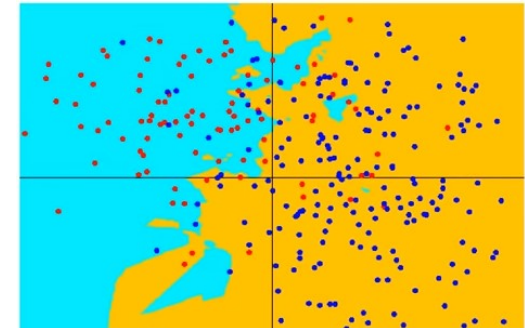
35



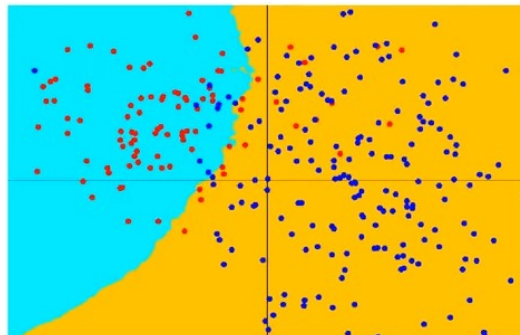
K=1



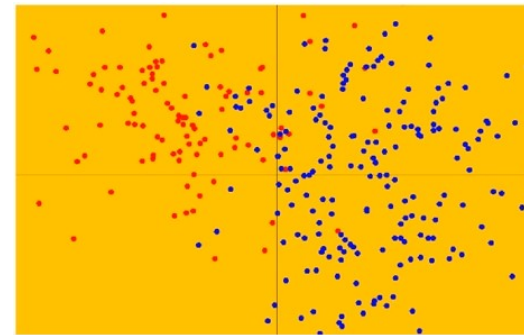
K=3



K=5



K=20



K=200

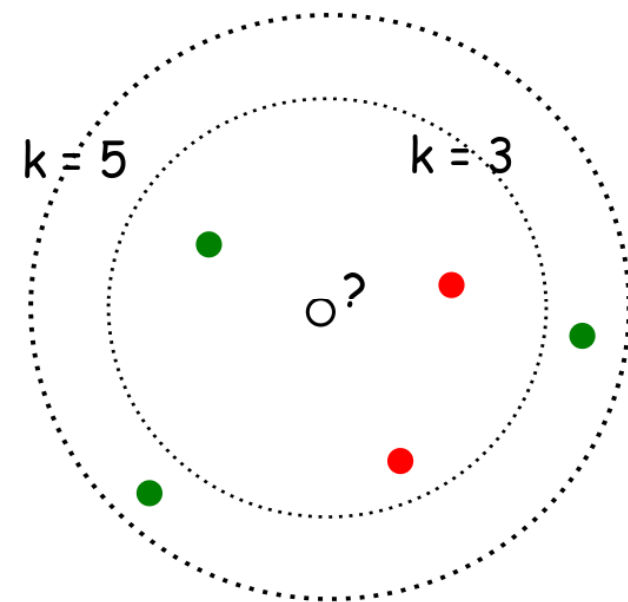
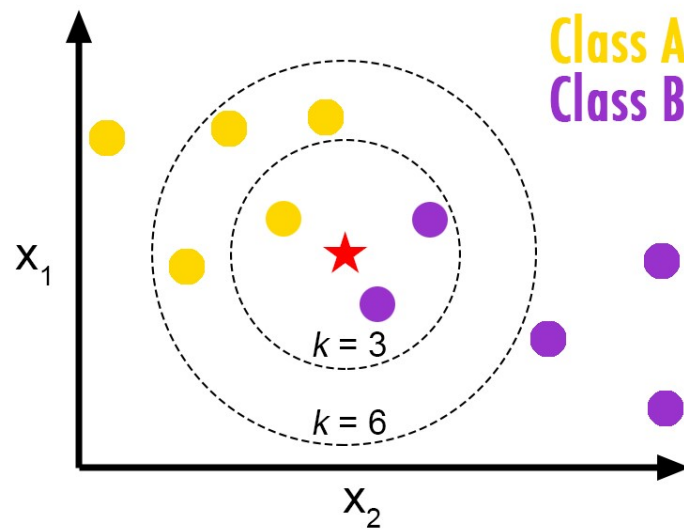
Algorithme KNN : critique

36

- ❑ **Pas d'apprentissage**: introduction de nouvelles données ne nécessite pas la reconstruction du modèle.
- ❑ Clarté des résultats
- ❑ Tout type de données
- ❑ Nombre d'attributs
- ❑ Temps de classification
- ❑ Stocker le modèle
- ❑ **Distance et nombre de voisins** : dépend de la distance, du nombre de voisins et du mode de combinaison.

Examples :

37



with $k = 3$, ●
with $k = 5$, ●

Exercice :

38

$$C_1 \Rightarrow (0,3), (0,2), (0,1), (0,0), (-1,0), (-2,0)$$

$$C_2 \Rightarrow (1,3), (1,1), (1,0), (0,-1)$$

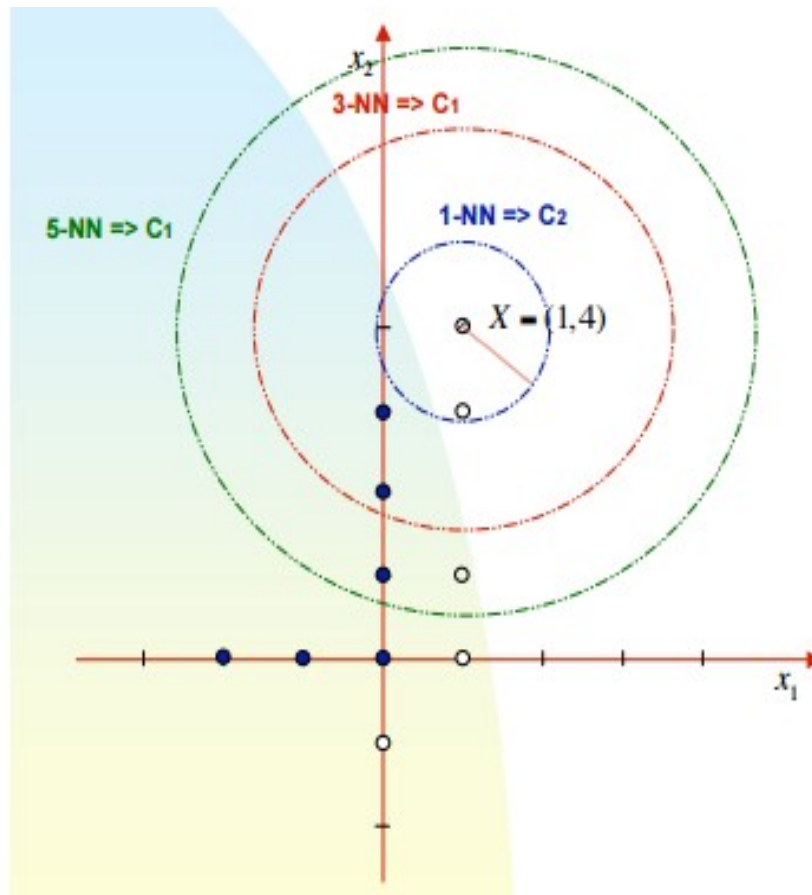
$$X = (1,4) \in ? \quad \text{avec} \quad 1 - NN, \quad 3 - NN \quad \text{et} \quad 5 - NN$$

□ La fonction de décision est :

$$g_i(X) = X^t Y_i - 1/2 Y_i^t Y_i, \quad 1 \leq i \leq M$$

Exercice : (corrigé)

39



La fonction de décision est :

$$g_i(X) = X^T Y_i - \frac{1}{2} Y_i^T Y_i, \quad 1 \leq i \leq M$$

	C_1	$g_1(X)$	C_2	$g_2(X)$	
5-NN			5-NN		
5-NN	(0, 3)	7.5	(1, 3)	8	1-NN
3-NN	(0, 2)	6	(1, 1)	4	
3-NN	(0, 1)	3.5	(1, 0)	0.5	
5-NN	(0, 0)	0	(0, -1)	-4.5	
	(-1, 0)	-1.5			
	(-2, 0)	-4			

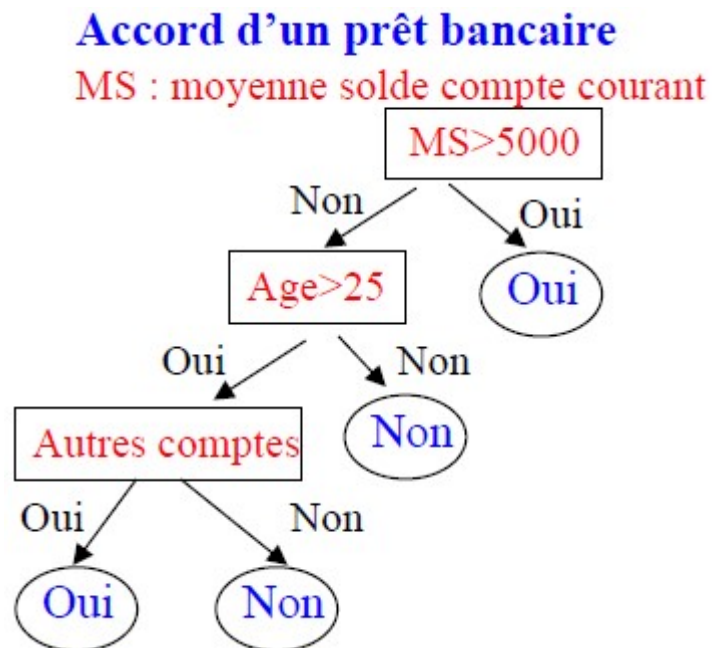
40

Les arbres de décision

Arbres de décision

41

- Génération d'**arbres de décision** à partir de données
- **Arbre** = Représentation graphique d'une procédure de classification



Un arbre de décision est un arbre où

- **Noeud interne** = un attribut
- **Branche d'un noeud** = un test sur un attribut
- **Feuilles** = classe donnée

Arbre de décision Exemple

42

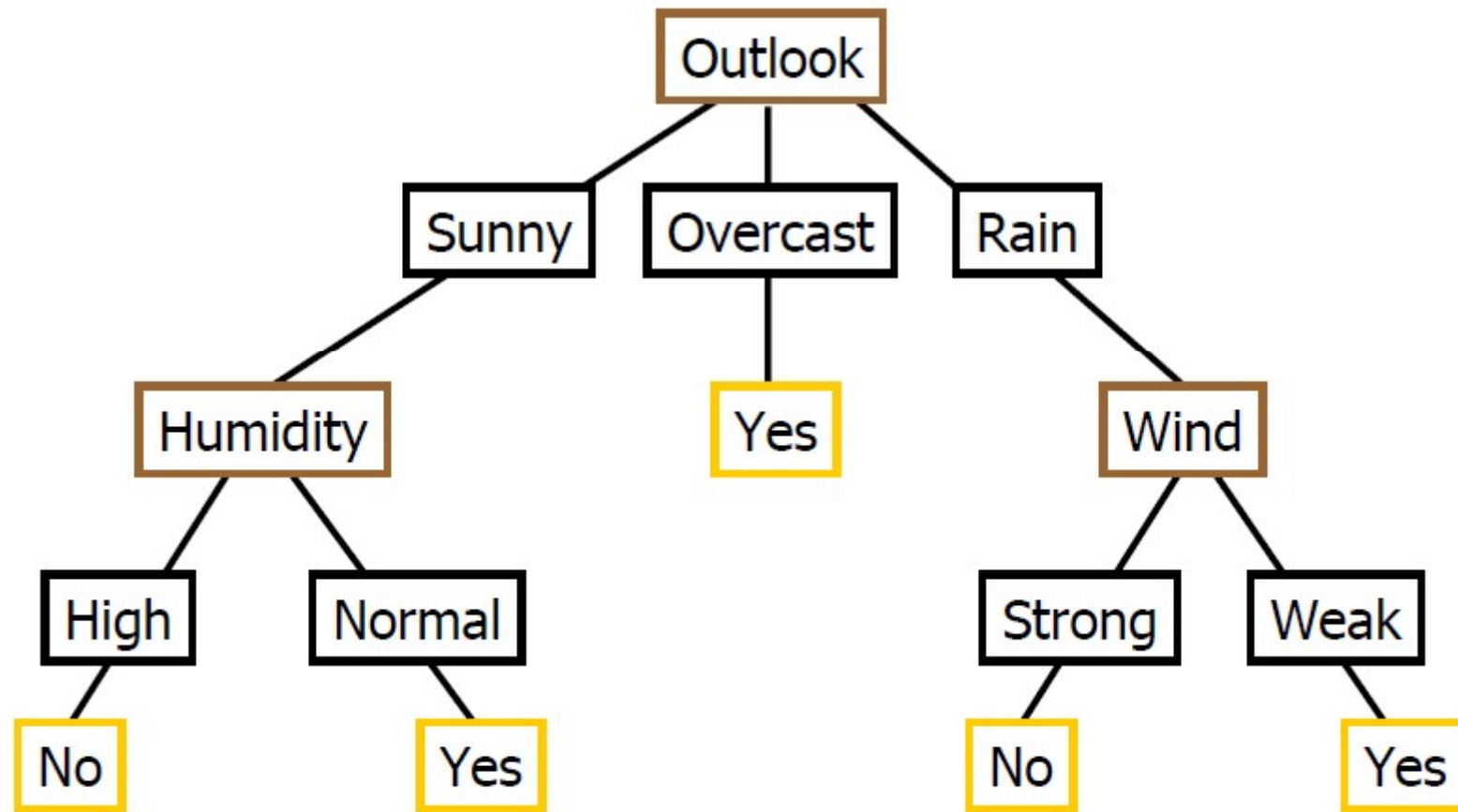
**Ensemble
d'apprentissage**

Outlook	Temperature	Humidity	Windy	Class
sunny	hot	high	false	N
sunny	hot	high	true	N
overcast	hot	high	false	P
rain	mild	high	false	P
rain	cool	normal	false	P
rain	cool	normal	true	N
overcast	cool	normal	true	P
sunny	mild	high	false	N
sunny	cool	normal	false	P
rain	mild	normal	false	P
sunny	mild	normal	true	P
overcast	mild	high	true	P
overcast	hot	normal	false	P
rain	mild	high	true	N

Jouer au tennis ?

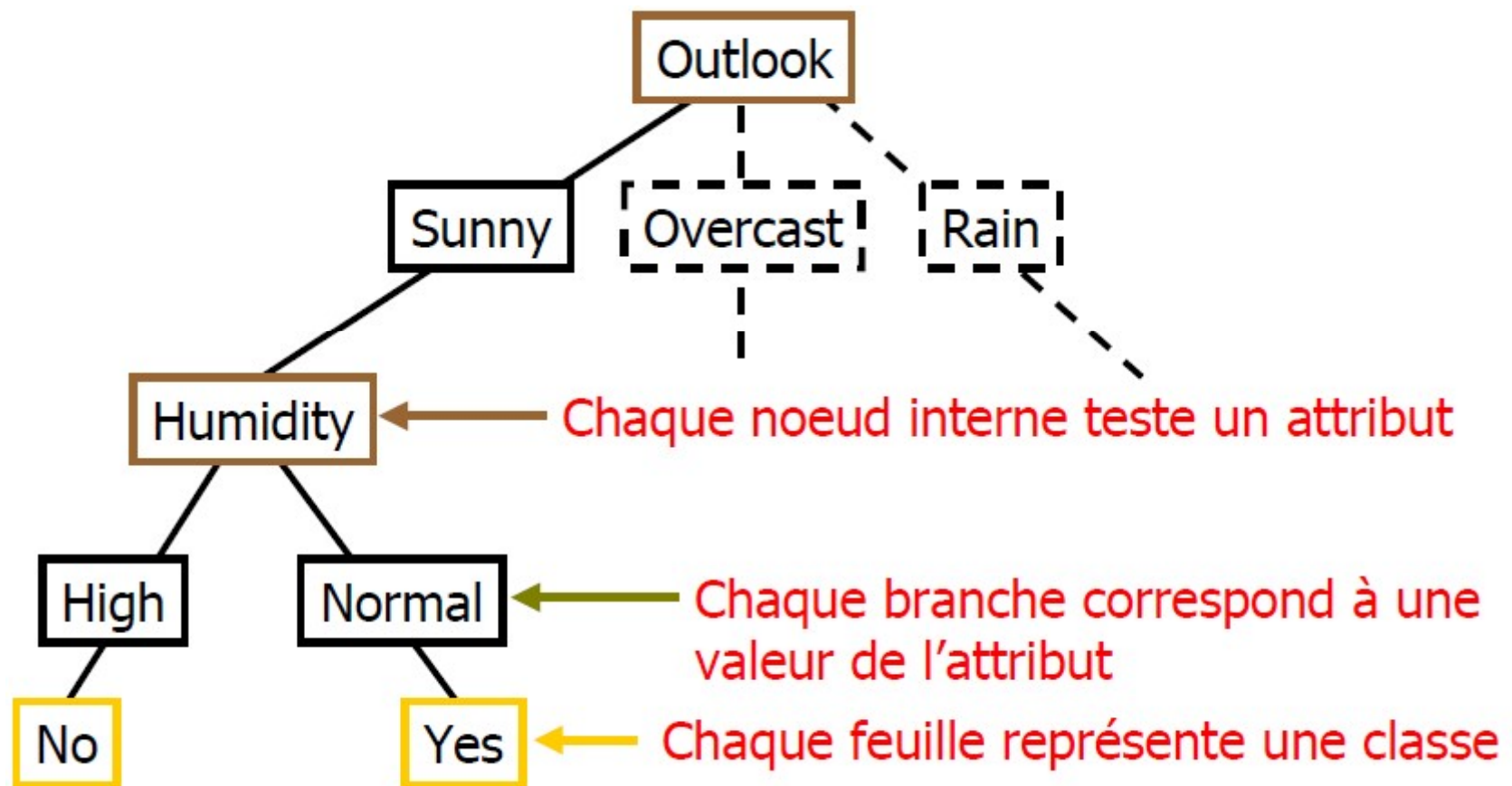
Arbre de décision Exemple

43



Exemple : Jouer au Tennis ?

44



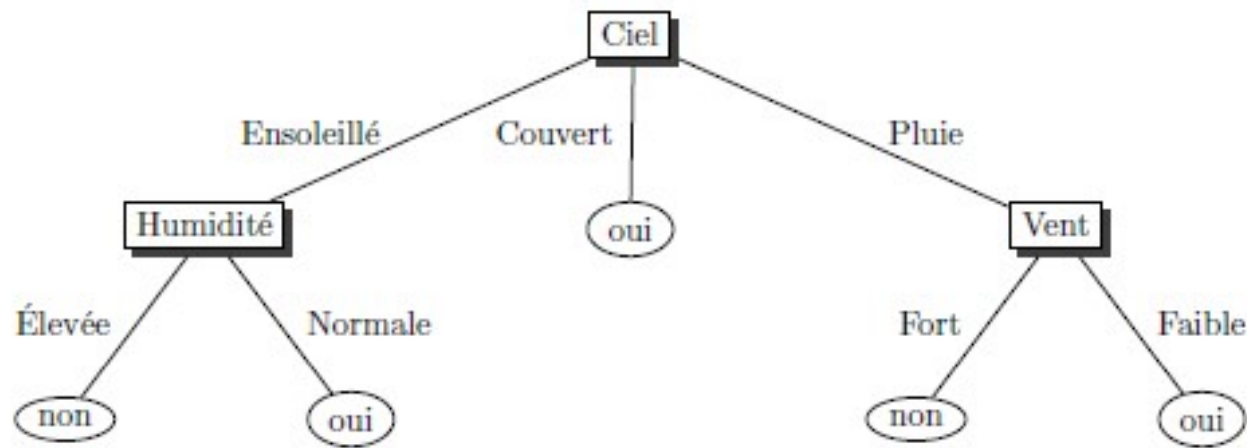
Utilisation d'un arbre de décision

45

- L'arbre de decision peut être exploité de différentes manières :
 - ▣ en y classant de nouvelles donnees;
 - ▣ en faisant de l'estimation d'attribut;
 - ▣ en extrayant un jeu de règles de classification concernant l'attribut cible;
 - ▣ en interprétant la pertinence des attributs;

Exemple : Classification de nouvelles données

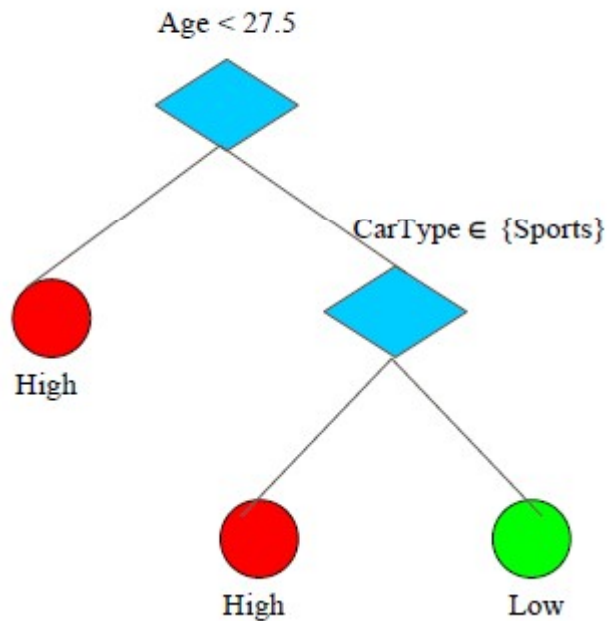
46



- (Ensoleille, Fraîche, Elevee, Fort) est classée comme **non**
- (Ensoleille, Fraîche, Normale, Fort) est classée comme **oui**
- (Pluie, Chaude, Normale, Faible) est classée comme **oui**
- (Pluie, Fraîche, Elevee, Fort) est classée comme **non**

Des arbres de décision aux règles

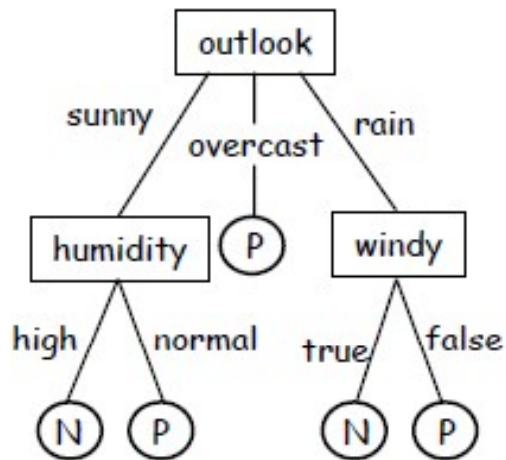
47



1. $\text{Age} < 27.5 \Rightarrow \text{High}$
2. $\text{Age} \geq 27.5$ and $\text{CarType} = \text{Sports} \Rightarrow \text{High}$
3. $\text{Age} \geq 27.5$ and $\text{CarType} \neq \text{Sports} \Rightarrow \text{Low}$

De l'arbre de décision aux règles de classification

48

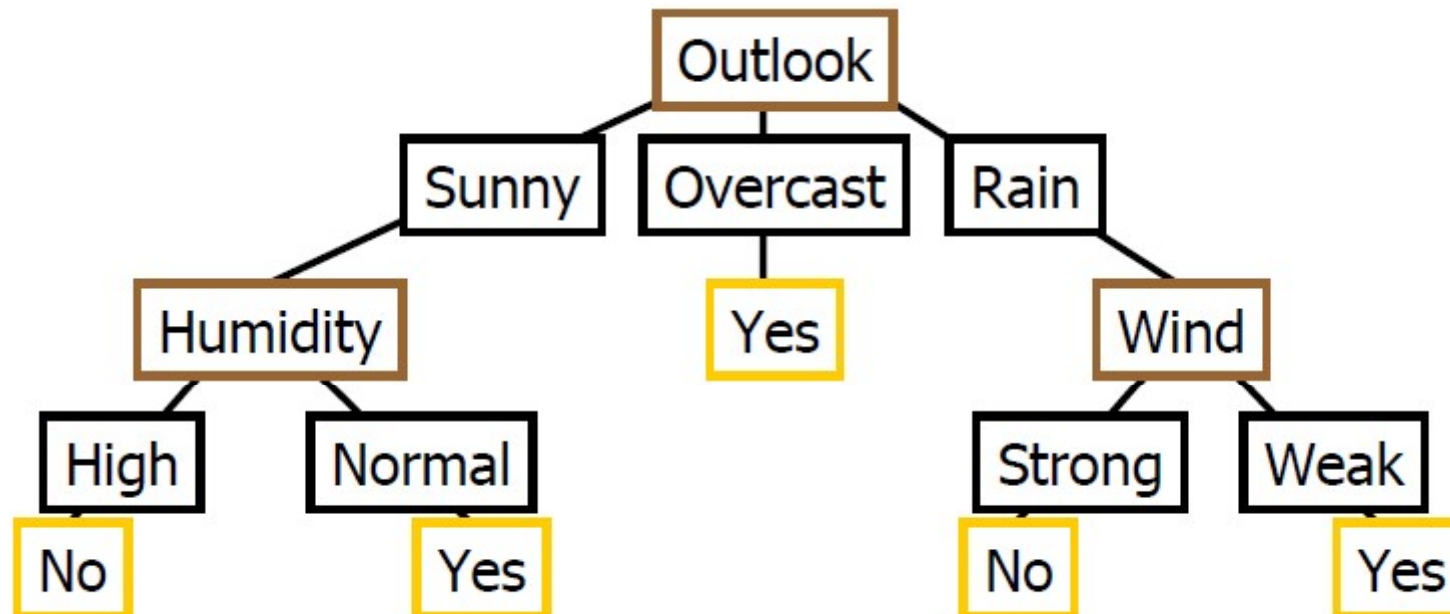


Si outlook=sunny
Et humidity=normal
Alors play tennis

- une **règle** est générée pour chaque chemin de l'arbre (de la racine à une feuille)
- Les paires attribut-valeur d'un chemin forment une conjonction
- Le noeud terminal représente la classe prédite
- Les règles sont généralement plus faciles à comprendre que les arbres

Des arbres de décision aux règles

49



R_1 : If (Outlook=Sunny) \wedge (Humidity=High) Then PlayTennis=No

R_2 : If (Outlook=Sunny) \wedge (Humidity=Normal) Then PlayTennis=Yes

R_3 : If (Outlook=Overcast) Then PlayTennis=Yes

R_4 : If (Outlook=Rain) \wedge (Wind=Strong) Then PlayTennis=No

R_5 : If (Outlook=Rain) \wedge (Wind=Weak) Then PlayTennis=Yes

Algorithmes de classification

50

□ Construction de l'arbre

- Au départ, toutes les instances d'apprentissage sont à la **racine** de l'arbre
- **Sélectionner** un attribut et choisir un test de séparation (**split**) sur l'attribut, qui sépare le “mieux” les instances. La sélection des attributs est basée sur une heuristique ou une mesure statistique.
- **Partitionner** les instances entre les noeuds fils suivant la satisfaction des tests logiques

Algorithmes de classification

51

- ▣ Traiter chaque noeud fils de façon récursive
- ▣ Répéter jusqu'à ce que tous les noeuds soient des terminaux. Un noeud courant est terminal si:
 - Il n'ya plus d'attributs disponibles
 - Le noeud est “pur”,i.e. Toutes les instances appartiennent à une seule classe,
 - Le noeud est “presque pur”,i.e. la majorité des instances appartiennent à une seule classe (Ex : 95%)
- ▣ Etiqueter le noeud terminal par la classe majoritaire

Algorithmes pour les arbres de décision

52

□ **Algorithme de base**

- ▣ Construction récursive d'un arbre de manière “diviser-pour-régner” descendante
- ▣ Attributs considérés énumératifs
- ▣ Glouton

□ **Plusieurs variantes : ID3, C4.5, CART, CHAID**

- ▣ Différence principale : mesure de sélection d'un attribut—critère de branchement (split)

Exemple :

53

Construire l'arbre de décision à partir des données suivantes.

client	Montant	Age	Residence	Etudes	Internet
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

Exemple :

54

□ 3 oui et 5 non

□ $(3,5) \rightarrow M (1,2) (2,1) (0,2)$

□ $(3,5) \rightarrow A (1,0) (2,2) (0,3)$

□ $(3,5) \rightarrow R (1,1) (1,2) (1,2)$

□ $(3,5) \rightarrow E (3,2) (0,3)$

Mesures de sélection d'attributs

55

- **Gain d'Information (ID3, C4.5)**
- **Indice Gini (CART)**
- **Table de contingence statistique χ^2 (CHAID)**
- **G-statistic**

Gain d'information

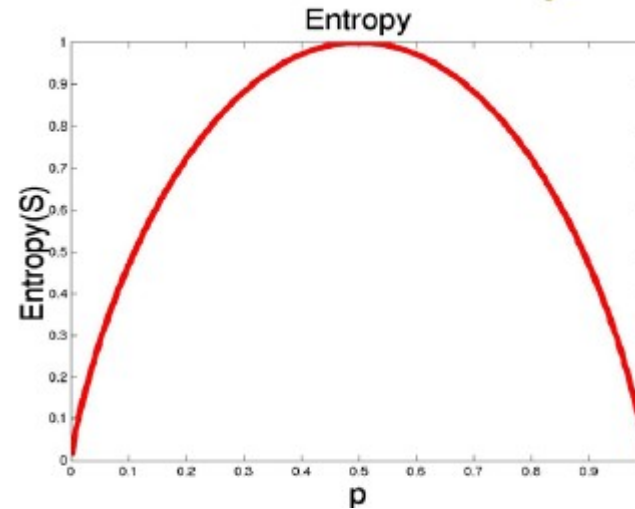
56

- **Sélectionner l'attribut avec le plus grand gain d'information**
- Soient P et N deux classes et S un ensemble d'instances avec **p** éléments de P et **n** éléments de N
- L'information nécessaire pour déterminer si une instance prise au hasard fait partie de P ou N est (entropie) :

$$I(p, n) = -\frac{p}{p+n} \log_2 \frac{p}{p+n} - \frac{n}{p+n} \log_2 \frac{n}{p+n}$$

Entropie

57



- S est l'ensemble d'apprentissage
- p_+ est la proportion d'exemples positifs (P)
- p_- est la proportion d'exemples négatifs (N)
- Entropie mesure l'impureté de S
- $\text{Entropie}(S) = -p_+ \log_2 p_+ - p_- \log_2 p_-$

Gain d'information

58

- Soient les ensembles $\{S_1, S_2, \dots, S_v\}$ formant une partition de l'ensemble S , en utilisant l'attribut A
- Toute partition S_i contient p_i instances de P et n_i instances de N
- L'entropie, ou l'information nécessaire pour classifier les instances dans les sous-arbres S_i est:

$$E(A) = \sum_{i=1}^v \frac{p_i + n_i}{p + n} I(p_i, n_i)$$

- Le gain d'information par rapport au branchement sur A est

$$\text{Gain}(A) = I(p, n) - E(A)$$

- Choisir l'attribut qui maximise le gain \rightarrow besoin d'information minimal

Gain d'information Exemple

59

□ Hypothèses

Day	Outlook	Temp.	Humidit	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Weak	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cold	Normal	Weak	Yes
D10	Rain	Mild	Normal	Strong	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

- Classe P : jouer_tennis = “oui”
- Classe N : jouer_tennis = “non”
- Information nécessaire pour classer un exemple donné est :

$$I(p, n) = I(9, 5) = 0.940$$

Gain d'information Exemple

60

- Calculer l'entropie pour l'attribut outlook :

outlook	p_i	n_i	$I(p_i, n_i)$
sunny	2	3	0,971
overcast	4	0	0
rain	3	2	0,971

- $$\begin{aligned} I(2,3) &= - (2/5)\ln_2(2/5) - (3/5)\ln_2(3/5) \\ &= 0.5287712 + 0.4421796 \\ &= 0.9709508 \end{aligned}$$

Gain d'information Exemple

61

- Calculer l'entropie pour l'attribut outlook :

outlook	p_i	n_i	$I(p_i, n_i)$
sunny	2	3	0,971
overcast	4	0	0
rain	3	2	0,971

On a

$$E(outlook) = \frac{5}{14}I(2,3) + \frac{4}{14}I(4,0) + \frac{5}{14}I(3,2) = 0.694$$

Alors $Gain(outlook) = I(9,5) - E(outlook) = 0.246$

De manière similaire

$$Gain(temperature) = 0.029$$

$$Gain(humidity) = 0.151$$

$$Gain(windy) = 0.048$$

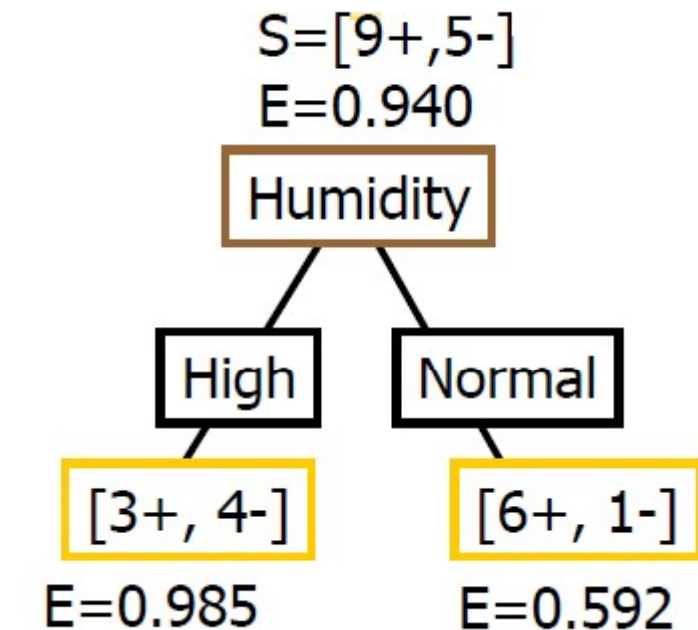
Exemple d'apprentissage

62

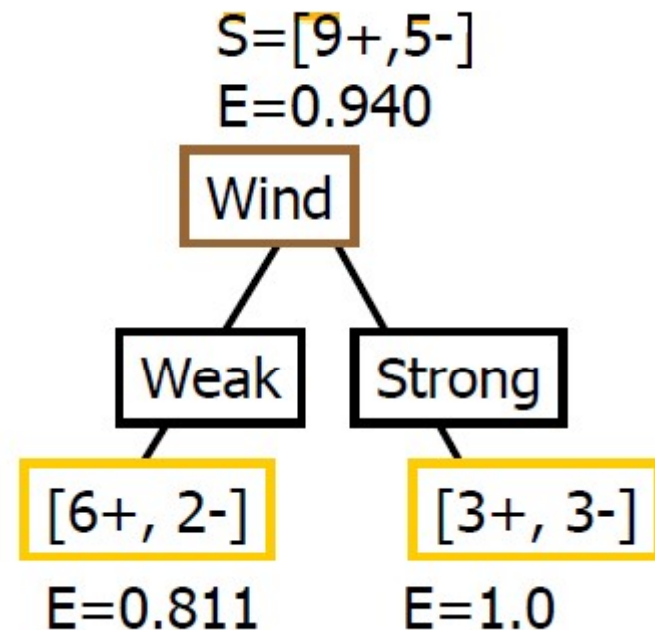
Day	Outlook	Temp.	Humidit	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Weak	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cold	Normal	Weak	Yes
D10	Rain	Mild	Normal	Strong	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Sélection de l'attribut suivant

63



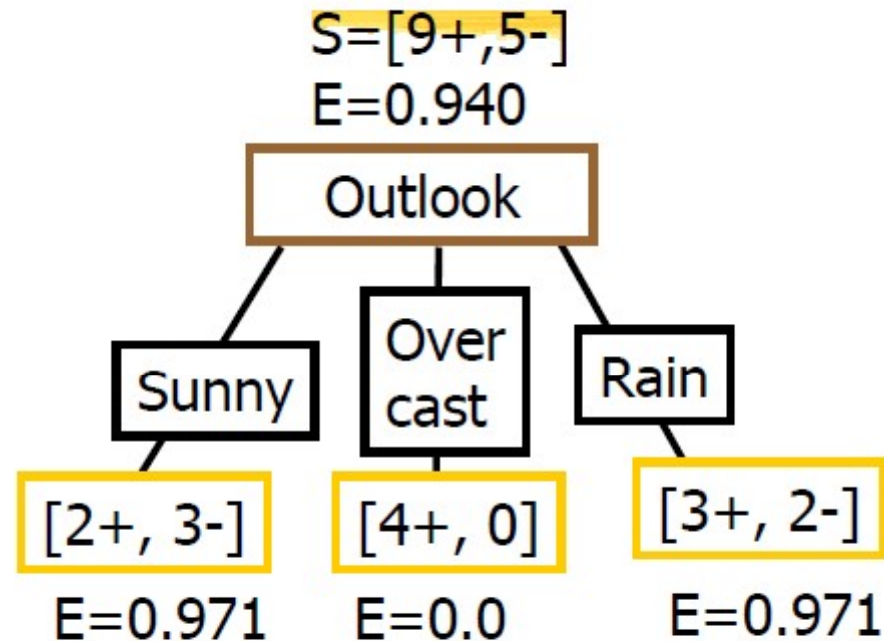
$$\begin{aligned}\text{Gain}(S, \text{Humidity}) &= 0.940 - (7/14) * 0.985 \\ &\quad - (7/14) * 0.592 \\ &= 0.151\end{aligned}$$



$$\begin{aligned}\text{Gain}(S, \text{Wind}) &= 0.940 - (8/14) * 0.811 \\ &\quad - (6/14) * 1.0 \\ &= 0.048\end{aligned}$$

Sélection de l'attribut suivant

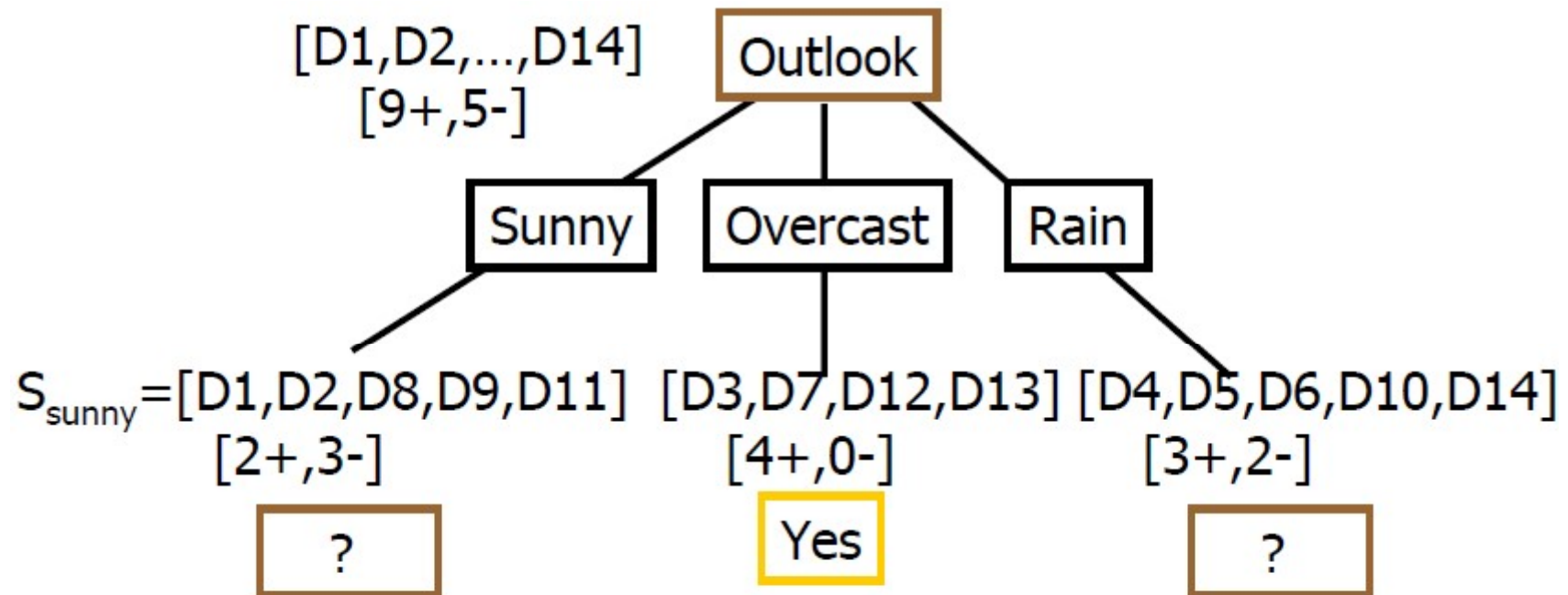
64



$$\begin{aligned} \text{Gain}(S, \text{Outlook}) &= 0.940 - (5/14) * 0.971 \\ &\quad - (4/14) * 0.0 - (5/14) * 0.0971 \\ &= 0.247 \end{aligned}$$

Algorithme ID3

65



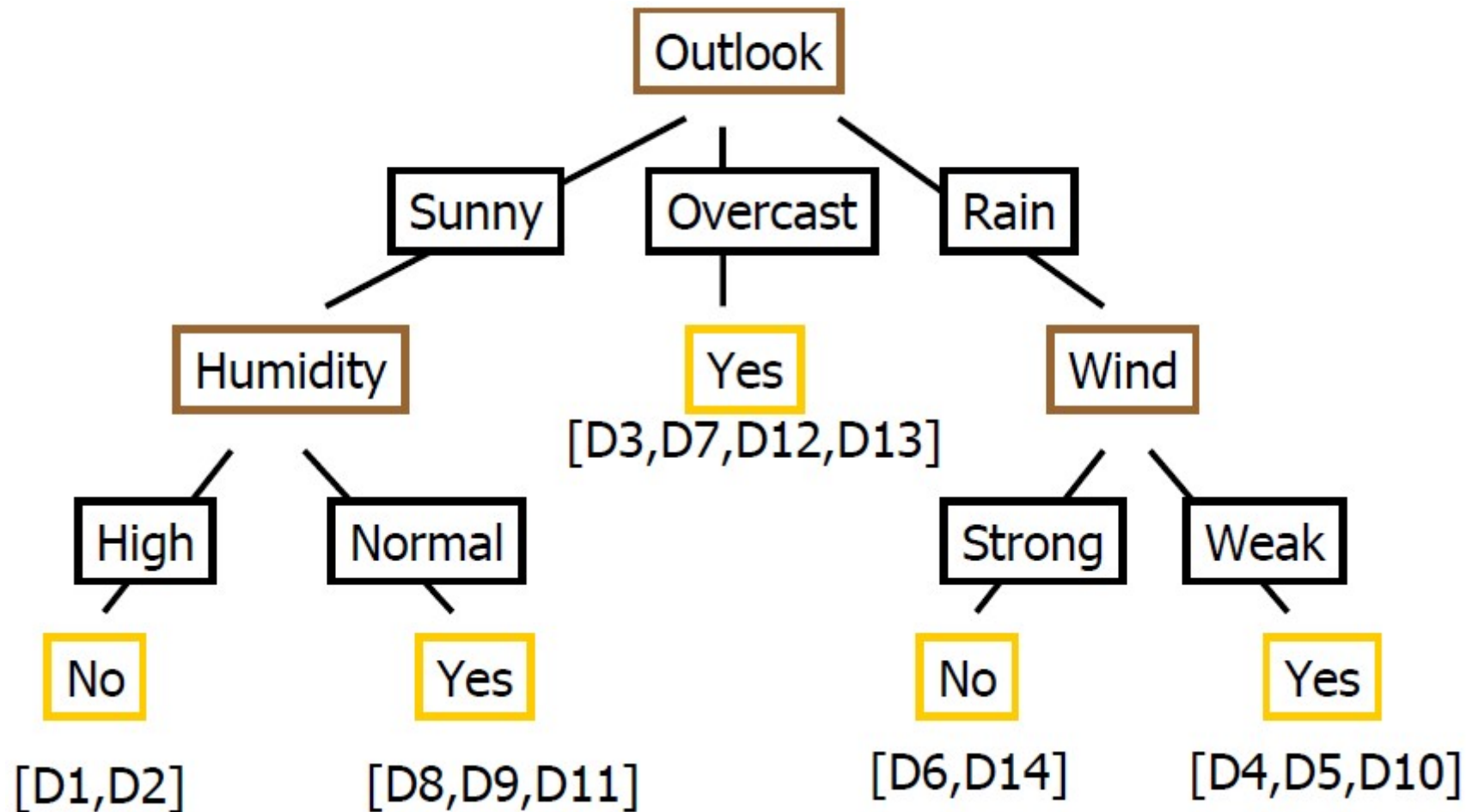
$$\text{Gain}(S_{\text{sunny}}, \text{Humidity}) = 0.970 - (3/5)0.0 - 2/5(0.0) = 0.970$$

$$\text{Gain}(S_{\text{sunny}}, \text{Temp.}) = 0.970 - (2/5)0.0 - 2/5(1.0) - (1/5)0.0 = 0.570$$

$$\text{Gain}(S_{\text{sunny}}, \text{Wind}) = 0.970 - (2/5)1.0 - 3/5(0.918) = 0.019$$

Algorithme ID3

66



Interprétation de l'arbre

67

- L'arbre de decision qui vient d'être construit nous donne des informations sur la pertinence des attributs vis-a-vis de la classe.
 - ▣ l'attribut « Temperature » n'étant pas utilisé dans l'arbre ; ceci indique que cet attribut *n'est pas pertinent* pour déterminer la classe.
 - ▣ Si l'attribut « outlook » vaut « sunny » , l'attribut « wind » n'est pas pertinent ;
 - ▣ si l'attribut « outlook » vaut « Rain » , c'est l'attribut « Humidity » qui ne l'est pas.

L'algorithme C4.5 et attributs numériques

68

- C4.5 prend en compte les attributs numériques.
- La construction d'un arbre de décision par C4.5 est identique dans son principe à la construction par ID3.
- Un noeud de l'arbre de décision peut contenir un test du fait que la valeur d'un attribut numérique est inférieure ou égale à un certain seuil.

L'algorithme C4.5 et attributs numériques

69

Jour	Ciel	Température	Humidité	Vent	Jouer au tennis ?
1	Ensoleillé	27,5	85	Faible	Non
2	Ensoleillé	25	90	Fort	Non
3	Couvert	26,5	86	Faible	Oui
4	Pluie	20	96	Faible	Oui
5	Pluie	19	80	Faible	Oui
6	Pluie	17,5	70	Fort	Non
7	Couvert	17	65	Fort	Oui
8	Ensoleillé	21	95	Faible	Non
9	Ensoleillé	19,5	70	Faible	Oui
10	Pluie	22,5	80	Faible	Oui
11	Ensoleillé	22,5	70	Fort	Oui
12	Couvert	21	90	Fort	Oui
13	Couvert	25,5	75	Faible	Oui
14	Pluie	20,5	91	Fort	Non

Test d'un attribut numérique

70

- Considerons les exemples dont l'attribut «Ciel» vaut « Ensoleille »

Jour	Température	« jouer au tennis »
1	27,5	non
2	25	non
8	21	non
9	19,5	oui
11	22,5	oui

- On commence par trier les exemples sur la valeur de leur attribut numérique. A chaque attribut, on associe le numero de son exemple associé ainsi que la valeur de l'attribut cible

Température	19,5	21	22,5	25	27,5
Jour	9	8	11	2	1
« jouer au tennis ? »	oui	non	oui	non	non

Test d'un attribut numérique

71

- On détermine le seuil S pour partitionner cet ensemble d'exemples. C4.5 utilise les règles suivantes :
 1. ne pas séparer deux exemples successifs ayant la même classe ; donc, on ne peut couper qu'entre les exemples x_9 et x_8 , x_8 et x_{11} , x_{11} et x_2
 2. si on coupe entre deux valeurs v et w ($v < w$) de l'attribut, le seuil S est fixé à v (ou encore $(v+w) / 2$);
 3. choisir S de telle manière que le gain d'information soit **maximal**

Température	19,5	21	22,5	25	27,5
Jour	9	8	11	2	1
« jouer au tennis ? »	oui	non	oui	non	non

Validation d'un arbre de decision

72

- **Valider** un arbre de décision en estimant la probabilité que la classe prédite pour une donnée quelconque soit correcte
- L'erreur de classification **E** d'un classeur est la **probabilité que ce classeur ne prédise pas correctement la classe** d'une donnée de l'espace de données.

Le taux de succès est égal à $1 - E$.

Validation d'un arbre de decision

73

- L'erreur apparente E_{app} est mesurée avec les exemples X_{app} utilisés pour la construction du classeur : c'est la proportion d'exemples dont la classe est mal prédite par le classeur.
- L'erreur de test E_{test} est mesurée avec des exemples de tests X_{test}

Mesures de qualité d'un classeur

74

- **Définitions :** (hypothèse : classification binaire)
 - **VP** : le nombre de **vrais positifs** : les exemples de classe positive et dont la classe est prédite comme positive ;
 - **VN** : le nombre de **vrais négatifs** : les exemples de classe négative et dont la classe est prédite comme négative ;
 - **FP** : le nombre de **faux positifs** : les exemples de classe négative et dont la classe est prédite comme positive ;
 - **FN** : le nombre de **faux négatifs** : les exemples de classe positive et dont la classe est prédite comme négative.

Mesures de qualité d'un classeur

75

□ Matrice de confusion

	+	− ← classe prédite
+	VP	FN
−	FP	VN
↑ classe		

S'il n'y a des nombres non nuls que sur la diagonale principale, c'est qu'aucun exemple n'est mal classe.

Mesures de qualité d'un classeur

76

□ précision pour les positifs $= \frac{VP}{VP + FP}$

□ précision pour les négatifs $= \frac{VN}{VN + FN}$

□ Autrement dit :

$$précision_i = \frac{\text{nb d'exemples correctement attribués à la classe } i}{\text{nb d'exemples attribués à la classe } i}$$

Mesures de qualité d'un classeur

77

□ rappel pour les positifs $= \frac{VP}{VP + FN}$

□ rappel pour les négatifs $= \frac{VN}{VN + FP}$

□ Autrement dit :

$$rappel_i = \frac{\text{nb d'exemples correctement attribués à la classe } i}{\text{nb d'exemples appartenant à la classe } i}$$

Mesures de qualité d'un classeur

78

□ Mesure F :

$$F = \frac{2rappel * precision}{rappel + precision} = \frac{2VP}{2VP + FP + FN}$$