

# Cours RF

Hubert CARDOT

Département Informatique de Polytech TOURS  
Laboratoire d'Informatique – RFAI  
Bureau 211

## Plan

- Cours : 8h
  - Intro – méthodes statistiques – kPPV
  - Arbres de décision – méthodes structurales
  - Sélection/extraction de caractéristiques
  - SVM
- TD : 4h
- TP : 6h
  - kPPV
  - libSVM
  - WEKA

2

Cours RF - Di 5

## Introduction

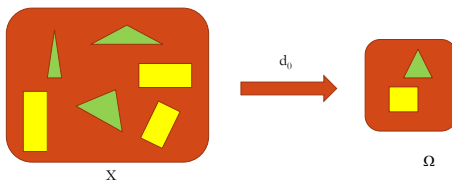
## Reconnaissance des Formes

- Formes : description d'un objet ou d'un concept (ex. une voiture, la grippe)
- Reconnaissance : à partir d'exemples de formes, on cherche à les classifier ou regrouper

4

Cours RF - Di 5

## Espace de représentation – espace des classes

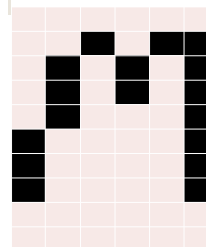


- $d_0$  : décision idéale

5

Cours RF - Di 5

## Exemple



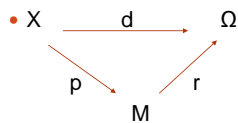
$X$  : ensemble des  $2^{60}$  distributions N&B

$\Omega$  : alphabet

6

Cours RF - Di 5

## Extraction de caractéristiques



$$r(p(x)) = w$$

7

Cours RF - Di 5

## Invariance et prétraitement



Forme canonique

8

Cours RF - Di 5

## Applications

- Signal 1D
  - Reco parole
  - Reco locuteur
  - Electrocardiogrammes
  - Radar
- Images
  - Reco textes (chiffres, manuscrits...)
  - Radiographies
  - Identification d'objets
  - Détection de défauts
  - Images aériennes, satellitaires
- Formes multidimensionnelles
  - Vidéo
  - Diagnostic de pannes
  - Préviction (météo, bourse...)
  - Fouille de données

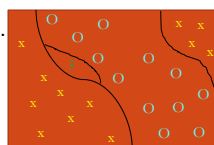
9

Cours RF - Di 5

## Fonction de décision

## Fonction de décision

- Problème :
  - $K$  classes  $w_1 \dots w_K$
  - $x \in \mathbb{R}^n$  : représentation d'une forme
  - On cherche une fct de décision  $d(x) = w_i$
  - Non unicité de la frontière à partir de la base d'apprent.
  - Zones d'indétermination (?)

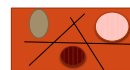


11

Cours RF - Di 5

## Séparation linéaire

- $d(x) = c_1 x_1 + \dots + c_n x_n + \text{cte}$   
 $d(x) = 'C.X + \text{cte} : \text{hyperplan}$
- Absolument séparable : chaque classe peut être séparée des autres par un hyperplan
- Séparable par paire : chaque paire de classes peut être séparée par un hyperplan

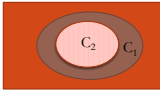


12

Cours RF - Di 5

## Fonction de décision généralisée

- Classes non linéairement séparables



- Exemple : fct polynomiale de d° 2  
 $d(x) = {}^t x.A.x + {}^t b.x + cte$   
 avec A une matrice donnant la forme de la frontière  
 - si 2 val. propres positives : ellipse (hyperellipsoïde)  
 - si A matrice identité : hypersphère  
 - si 2 val. propres négatives : hyperhyperboloïde

13

Cours RF - Di 5

## Evaluation

## Evaluation – bases de données

- $d()$  est déterminée à partir d'une base d'apprent. L'objectif est que  $d()$  puisse généraliser aux autres formes. Evaluation sur base de test.
- Pour déterminer certains paramètres, on peut garder certains exemples de la base d'apprent. en une base de validation.

15

Cours RF - Di 5

## Matrice de confusion

Classe réelle \ Décision du syst	$W_1$	$W_2$	$W_j$	$W_K$	$W_0$ (rejet)
$W_1$	$n_{11}$	$n_{12}$		$n_{1K}$	$n_{10}$
$W_2$	$n_{21}$	$n_{22}$		$n_{2K}$	$n_{20}$
$W_i$			$n_{ij}$		
$W_K$	$n_{K1}$	$n_{K2}$		$n_{KK}$	$n_{K0}$

$n_{ij}$  : nombre d'ex. de la classe  $i$  affectés à la classe  $j$

16

Cours RF - Di 5

## Taux de reconnaissance

- Si classes équilibrées

$$\frac{\text{somme diagonale}}{\text{nb total d'exemples}} = \frac{\sum_{i=1}^K n_{ii}}{\sum_{i=1}^K \sum_{j=0}^K n_{ij}}$$

- Sinon : moyenne des taux de reco par classe 99,9 % ?

	$W_1$	$W_2$
$W_1$	10 000	0
$W_2$	10	0

17

Cours RF - Di 5

## Autres notions d'évaluation

- Sensibilité, spécificité, courbe de ROC (médical)
- Précision et rappel (document)
- Taux de Fausse Acceptation et Taux de Faux Rejet, Taux d'Égale Erreur (biométrie)

18

Cours RF - Di 5

## Approche statistique

## Approche statistique

- $d : \mathbb{R}^n \rightarrow \Omega$   
 $x \mapsto d(x)$
- $f$  : densité de probabilité de  $x$  dans  $\mathbb{R}^n$

$\int_A f(x)dx$  : proba de trouver  $x$  dans la zone  $A \in \mathbb{R}^n$

$f(x / w)$  : densité de proba au sein de la classe  $w$

$p(w)$  : proba de la classe  $w$

$p(w / x)$  : proba qu'une forme  $x \in w$

20

Cours RF - Di 5

## Théorie bayésienne de la décision

- Fonction de décision bayésienne : on affecte  $x$  à la classe pour laquelle  $p(w / x)$  est maximum
- $d_1(x) = w_i$  avec  $i = \arg_{j=1 \text{ à } K} \max p(w_j / x)$
- $d_1()$  n'est pas unique car il peut y avoir plusieurs classes avec la même proba
- Toute fonction de décision bayésienne minimise la proba d'erreur globale

21

Cours RF - Di 5

## Règle pratique de décision bayésienne

- Pb : il est difficile de connaître tous les  $p(w/x)$
- En appliquant le théorème de Bayes  
 $P(A / B) = P(A).P(B / A) / P(B)$
- $\forall w_i \in \Omega \quad p(w)f(x/w) \geq p(w_i)f(x/w_i) \Leftrightarrow d_1(x) = w$   
c'est la règle pratique de décision bayésienne
- Cela revient à choisir l'hypothèse maximisant la proba de ce que nous observons
- Il reste à connaître les  $K$  valeurs  $p(w_i)$  et les  $K$  fonctions  $f(x / w_i)$

22

Cours RF - Di 5

## Apprentissage

- Déterminer  $p(w_i)$  et  $f(x / w_i)$
- $p(w_i)$ 
  - Fréquence sur une base d'apprent.
  - Connaissance extérieure
  - Equiprobabilité
- $f(x / w_i)$ 
  - Méthodes paramétriques : on choisit la forme de  $f$  puis on calcule au mieux, sur les données d'apprent., les paramètres
  - Méthodes non paramétriques

23

Cours RF - Di 5

## Coût d'une décision

C	$w_1$	$w_2$	$w_0$
$w_1$	0	5	2
$w_2$	20	0	2

- On cherche une fonction de décision qui minimise les coûts de mauvais classement
- $\text{coût}(d()) = \int_{\mathbb{R}^n} [\sum_{i=1}^K p(w_i|x) \cdot C(w_i, d(x))] f(x) dx$

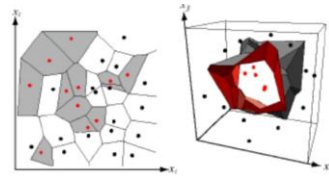
24

Cours RF - Di 5

## Plus proche voisin

## Méthode du plus proche voisin

- $d(x) = d(x_i)$  avec  $i = \arg \min_j \text{dist}(x, x_j)$



- La méthode du PPV a un taux d'erreur moyen inférieur à 2 fois celui de Bayes quand le nombre d'exemples d'apprentissage est grand

26

Cours RF - Di 5

## k-PPV

- On peut améliorer les performances du PPV quand les données d'apprentissage sont bruitées ou potentiellement mal étiquetées en prenant les k PPV puis en faisant un vote pour déterminer la classe majoritaire.
- Classe de rejet : rejet d'ambiguïté et rejet de distance

27

Cours RF - Di 5

## Distances inter et intraclasse

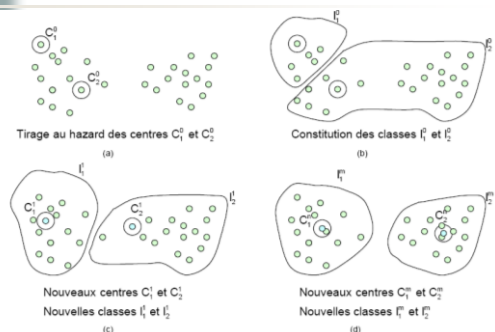
- La distance **intraclasse** est la moyenne des distances des exemples de la classe avec tous les autres exemples de la classe. Elle vaut 2 fois la somme des variances sur chaque composante des formes.
- La distance **interclasse** est la moyenne des distances des exemples d'une classe avec tous les exemples d'une autre classe.
- On cherche à maximiser  $\text{dist inter} / \text{dist intra}$ .

28

Cours RF - Di 5

## K-Means (centres mobiles, nuées dynamiques)

## K-Means



30

Cours RF - Di 5

## Classification Ascendante Hiérarchique (CAH)

## Intro

- Classier, c'est regrouper entre eux des objets similaires selon des critères  
On parle aussi de clustering ou de partitionnement
- Classif. non hiérarchique : classes disjointes
- Classif. hiérarchique : en fonction du niveau de précision deux individus peuvent ou non être dans une même classe

分类, 它是根据标准将相似的对象组合在一起的我们还讨论了聚类或分区分类。非层次结构: 不连续的类分类。层次结构: 根据选择精度的级别, 两个人可以在同一个类中, 也可以不在同一个类中

32

Cours RF - Di 5

## Données centrées réduites

	TAI	VIT	DET	PAS	LEG	STA
I1	-1.1125	1.3473	1.5025	0.9702	1.1665	0.5535
I2	-0.0056	-0.7615	-0.7446	0.9702	-1.0845	-0.9793
I3	1.1013	-0.9724	-0.6643	1.5023	-0.9514	0.0426
I4	-0.8106	1.1364	0.9407	1.2120	1.1423	0.5535
I5	-1.1125	1.3473	0.9407	-0.2392	1.2391	1.0644
I6	-0.6093	0.7146	1.1012	-0.2876	0.9124	0.8090
I7	-1.1125	0.5038	0.9407	-0.4810	1.3964	-1.2347
I8	-1.3137	1.3473	1.4222	-0.9648	1.2028	-2.0011
I9	-1.5150	1.3473	1.4222	-1.4485	1.2028	-1.7456
I10	-0.0056	-1.3941	-0.8248	1.0669	-0.4673	-1.2347
I11	-0.4075	-0.1289	-0.2630	1.2120	-0.3705	-0.2129
I12	-0.1062	0.0820	-0.6643	-0.4810	-0.2857	0.2980
I13	0.3969	-0.3398	-0.6643	-0.0941	-0.5278	1.0644
I14	1.1013	-0.7615	-0.8248	-0.7229	-1.0240	0.5535
I15	1.0007	-0.5506	-1.0656	-0.8197	-0.9877	0.2980
I16	1.1013	-0.5506	-1.2261	-1.2067	-0.6246	0.8090
I17	1.1013	-0.9724	-0.6643	-1.2067	-1.0966	0.2980
I18	1.4032	-1.3941	-0.6643	1.0185	-0.8424	1.0644

33

Cours RF - Di 5

## Indice de dissimilarité ou distance

$$d(I_i, I_j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + (x_{i3} - x_{j3})^2 + (x_{i4} - x_{j4})^2 + (x_{i5} - x_{j5})^2 + (x_{i6} - x_{j6})^2}$$

Ainsi, la distance entre les sujets I1 et I2 est donnée par :

$$d(I_1, I_2) = \sqrt{(-1.1125 + 0.0056)^2 + (1.3473 + 0.7615)^2 + \dots + (0.5535 + 0.9793)^2} = 4.2588$$

34

Cours RF - Di 5

## Distances

Dist. Euclidiennes (Basket-CR.sta)

	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	I11	I12	I13	I14	I15	I16	I17	I18
I1	0.00	4.26	4.47	0.71	1.43	1.59	2.53	3.21	3.30	4.48	3.30	3.40	3.75	4.74	4.75	4.89	4.99	4.78
I2	4.26	0.00	1.62	3.80	4.42	3.84	3.74	4.57	4.81	0.93	1.43	2.26	2.44	2.54	2.45	3.11	2.77	2.57
I3	4.47	1.62	0.00	3.93	4.66	4.02	4.35	5.52	5.76	1.87	1.31	2.65	2.16	2.30	2.41	2.92	2.72	1.25
I4	0.71	3.80	3.93	0.00	1.58	1.62	2.97	3.43	3.63	4.00	2.76	3.03	3.31	4.34	4.35	4.50	4.65	4.25
I5	1.43	4.42	4.66	1.58	0.00	0.92	2.47	3.19	3.12	4.66	3.54	2.86	3.29	4.25	4.24	4.30	4.45	4.73
I6	1.59	3.84	4.02	1.62	0.92	0.00	2.18	3.07	3.05	4.05	2.96	2.35	2.72	3.58	3.61	3.63	3.75	4.06
I7	2.53	3.74	4.35	2.97	2.47	2.18	0.00	1.36	1.53	3.72	3.39	2.99	3.83	4.33	4.21	4.42	4.33	5.01
I8	3.21	4.57	5.52	3.43	3.19	3.07	1.36	0.00	0.58	4.67	4.33	3.89	4.82	5.18	5.01	5.27	5.12	6.06
I9	3.30	4.81	5.76	3.63	3.12	3.05	1.53	0.58	0.00	4.92	4.58	3.91	4.86	5.21	5.05	5.23	5.11	6.21
I10	4.48	0.93	1.87	4.00	4.66	4.05	3.72	4.67	4.92	0.00	1.80	2.64	2.82	2.89	2.82	3.39	3.06	2.73
I11	3.30	1.43	1.31	2.76	3.54	2.96	3.39	4.33	4.58	1.80	0.00	1.92	1.89	2.42	2.42	2.90	2.81	3.12
I12	3.40	2.26	2.65	3.03	2.86	2.35	2.99	3.89	3.91	2.64	1.92	0.00	1.11	1.69	1.55	1.75	1.94	2.76
I13	3.75	2.44	2.16	3.31	3.29	2.72	3.83	4.82	4.86	2.82	1.89	1.11	0.00	1.27	1.38	1.47	1.75	1.86
I14	4.74	2.54	2.30	4.34	4.25	3.58	4.33	5.18	5.21	2.89	2.42	1.69	1.27	0.00	0.43	0.82	0.61	1.96
I15	4.75	2.45	2.41	4.35	4.25	3.61	4.21	5.03	5.05	2.82	2.42	1.55	1.38	0.43	0.00	0.76	0.71	2.24
I16	4.89	3.11	2.92	4.30	4.20	3.63	4.42	5.27	5.23	3.39	2.90	1.75	1.47	0.82	0.76	0.00	0.99	2.49
I17	4.99	2.77	2.72	4.65	4.45	3.75	4.33	5.12	5.11	3.06	2.81	1.94	1.75	0.61	0.71	0.99	0.00	2.42
I18	4.78	2.57	1.25	4.26	4.73	4.06	5.01	6.06	6.21	2.73	3.12	2.76	1.86	1.96	2.24	2.49	2.42	0.00

35

Cours RF - Di 5

## Agrégation

非零最小值介于 I14 和 I15 之间

- Minimum non nul est entre **I14** et **I15** : 0,43
- Indice d'agrégation : « saut minimum » 聚合指数: 最小跳转

La distance D entre deux groupes A et B est alors définie par :

$$D(A, B) = \min_{i \in A, j \in B} d(I_i, I_j)$$

- On regroupe donc **I14** et **I15**
- Il existe d'autres choix possibles pour l'agrégation
  - Diamètre (saut maximal)
  - Moyenne pondérée
  - Centroïde pondéré ou non
  - Méthode Ward (inertie totale)

36

Cours RF - Di 5

## Tableau après regroupement

	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	I11	I12	I13	I14	I15	I16	I17	I18
I1	0	4.26	4.47	0.71	1.43	1.50	2.53	3.21	3.36	4.48	3.3	3.4	3.75	4.74	4.89	4.99	4.78	
I2	4.26	0	1.62	3.8	4.42	3.84	3.74	4.57	4.81	0.93	1.43	2.26	2.44	2.45	3.11	2.77	2.57	
I3	4.47	1.62	0	3.93	4.66	4.02	4.55	5.32	5.76	1.87	1.31	2.65	2.16	2.3	2.92	2.72	1.25	
I4	0.71	3.8	3.93	0	1.58	1.62	2.57	3.43	3.63	4	2.76	3.03	3.31	4.34	4.5	4.65	4.26	
I5	1.43	4.42	4.66	1.58	0	0.92	2.47	3.19	3.12	4.66	3.54	2.86	3.39	4.24	4.2	4.45	4.73	
I6	1.50	3.84	4.02	1.62	0.92	0	2.18	3.07	3.05	4.05	2.96	2.35	2.72	3.58	3.63	3.75	4.06	
I7	2.53	3.74	4.55	2.57	2.47	3.18	0	1.36	1.53	3.75	3.09	2.99	3.83	4.21	4.42	4.33	5.01	
I8	3.21	4.57	5.32	3.43	3.19	3.07	1.36	0	0.58	4.67	4.33	3.89	4.82	5.03	5.27	5.12	6.06	
I9	3.36	4.81	5.76	3.63	3.12	3.05	1.53	0.58	0	4.92	4.58	3.91	4.86	5.05	5.23	5.11	6.31	
I10	4.48	0.93	1.87	4	4.66	4.02	3.72	4.67	4.82	0	1.8	2.64	2.82	2.82	3.39	3.06	2.73	
I11	3.3	1.43	1.31	2.76	3.54	2.96	3.39	4.33	4.58	1.8	0	1.92	1.89	2.42	2.9	3.81	2.12	
I12	3.4	2.26	2.65	3.03	2.86	2.35	2.99	3.89	3.91	2.64	1.92	0	1.11	1.55	1.75	1.94	2.76	
I13	3.75	2.44	2.16	3.31	3.29	3.72	3.83	4.82	4.86	2.82	1.89	1.11	0	1.27	1.47	1.75	1.86	
I14	4.74	2.45	2.3	4.34	4.24	3.58	4.21	5.03	5.05	2.82	2.42	1.55	1.27	0	0.76	0.61	1.96	
I15																		
I16	4.89	3.11	2.92	4.5	4.2	3.63	4.42	5.27	5.23	3.39	2.9	1.75	1.47	0.76	0	0.99	2.49	
I17	4.99	2.77	2.72	4.65	4.43	3.75	4.33	5.12	5.11	3.06	2.81	1.94	1.75	0.61	0.99	0	2.42	
I18	4.78	2.57	1.25	4.26	4.73	4.06	5.01	6.06	6.21	2.73	2.12	2.76	1.86	1.96	2.49	2.42	0	

37

Cours RF - Di 5

## Tableau résultat

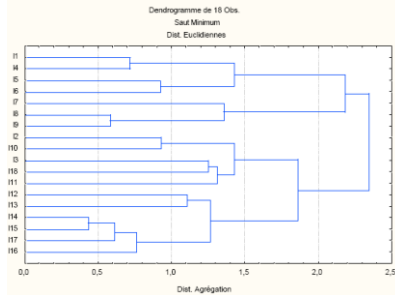
En poursuivant la méthode, on obtient la suite d'objets suivante :

	Objet #1	Objet #2	Objet #3	Objet #4	Objet #5	Objet #6	Objet #7	Objet #8	Objet #9	Objet #10	Objet #11	Objet #12	Objet #13	Objet #14	Objet #15	Objet #16	Objet #17	Objet #18
2341613	I14	I15																
2313863	I8	I9																
8121799	I14	I15	I17															
7143280	I1	I4																
7969390	I14	I15	I17	I16														
2131483	I5	I6																
9238429	I2	I10																
1107561	I13	I13																
1146807	I3	I18																
11263891	I12	I13	I14	I17	I16													
1313007	I3	I18	I11															
1337462	I7	I8	I9															
1432591	I2	I10	I3	I12	I11													
1439021	I1	I4	I5	I6														
1188421	I2	I10	I3	I10	I11	I12	I13	I14	I15	I17	I16							
2134427	I1	I4	I5	I6	I7	I8	I9	I2	I10	I3	I18	I11	I12	I13	I14	I15	I17	I16
2146117	I1	I4	I5	I6	I7	I8	I9	I2	I10	I3	I18	I11	I12	I13	I14	I15	I17	I16

38

Cours RF - Di 5

## Dendrogramme



39

Cours RF - Di 5

## Remarques

- On remarque sur le dendrogramme précédent un « saut » d'indice après la partition en 4 classes.
- Il paraît donc judicieux d'étudier cette partition en 4 classes.

40

Cours RF - Di 5