

Combinaison de Classifieurs

Rappels sur la Classification

Laurent HEUTTE

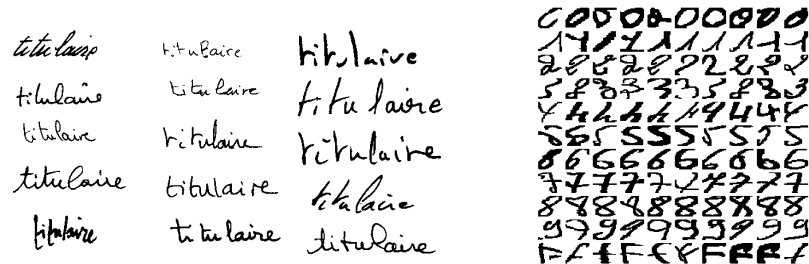
Laurent.Heutte@univ-rouen.fr
<http://www.univ-rouen.fr/psi/heutte>

Plan du Cours

1. Qu'est-ce qu'un classifieur?
2. Pourquoi et comment les combiner?
3. Combinaison parallèle de classifieurs
4. Méthodes pour la construction d'ensembles de classifieurs

Reconnaissance de Formes

But: Apprendre à une machine à reconnaître des formes malgré des modifications par rapport aux modèles



Classifieur: application de l'espace des formes (espace de représentation) vers l'espace de décision (espace des classes)

Processus de RdF

Le processus de RdF est un processus de réduction *progressive* et *sélective* de l'information :

1- Prétraitements

binarisation, localisation, segmentation, élimination du bruit, normalisation, ...

2- Extraction des informations pertinentes

parole: fréquence fondamentale, harmoniques, énergie, ...
image: occlusions, concavités, contours, fins de trait, ...

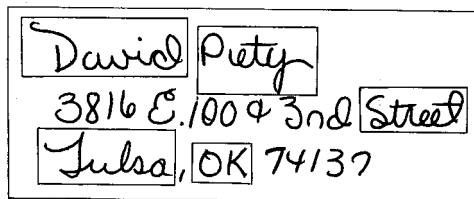
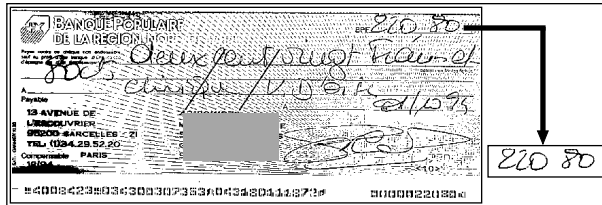
3- Classification de la forme

apprentissage et décision



Prétraitements (1)

Binarisation, localisation, segmentation



DECLARATION DE SITUATION MENSUELLE

Notifiez () les cases correspondant à votre cas

Mme. **BQUART** NOM : **BQUART** Prénom : **Catherine**

- ☐ Ce mois-ci, avez-vous travaillé ?
☐ OUI, du **03-04-01** au **27-04-01**
 Heures travaillées dans le mois :
☒ NON, je n'ai pas travaillé ce mois-ci.
- ☐ Ce mois-ci, avez-vous été en stage ?
☐ OUI, du **03-04-01** au **27-04-01**
☐ NON
- ☐ Ce mois-ci, avez-vous été en arrêt maladie ?
☐ OUI, du **11-04-01** au **20-04-01**
☐ NON
- ☐ Ce mois-ci, avez-vous été en congé maternité ?
☐ OUI, depuis le **11-04-01**
☐ NON
- Percevez-vous une retraite ?
☐ OUI, depuis le **11-04-01**
☐ NON
- Percevez-vous une pension d'invalidité de 2^e ou 3^e catégorie ?
☐ OUI, depuis le **11-04-01**
☐ NON
- Êtes-vous toujours à la recherche d'un emploi ?
☐ OUI
- NON, depuis le **11-04-01**
 Pour quel motif ?
☐ Reprise du travail
☐ Retraite
☐ Service national
☐ Autre

Prétraitements (2)

Amélioration, redressement, normalisation

faire faire faire

Réduction de l'information à traiter

2 2 2

Image brute

Contours

4 4

Squelette

2 2

Image Normalisée

Extraction de Caractéristiques (1)

But:

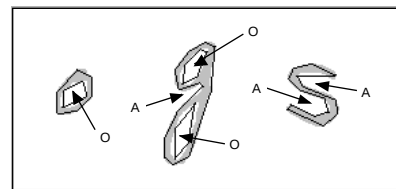
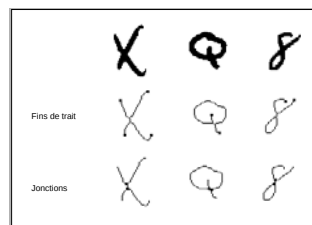
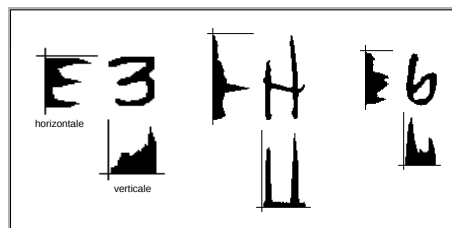
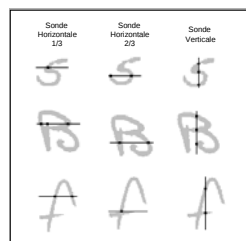
décrire la forme bidimensionnelle au moyen d'un ensemble de paramètres adéquats (attributs)

Contraintes:

- faible variance intra-classe:
qu'est-ce qui décrit une classe?
- grande variance inter-classe:
qu'est-ce qui fait la différence entre les classes?
- indépendance en translation, rotation et facteur d'échelle
en fonction de l'application
- faible nombre d'attributs:
attention au temps de calcul!!!

Extraction de Caractéristiques (2)

Exemples en reconnaissance de caractères:

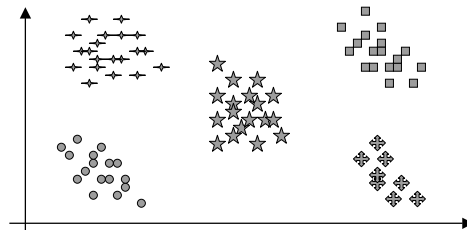


➔ Faire preuve d'imagination mais faire aussi un peu de biblio!!!

Classification Statistique (1)

Prendre une décision dans l'espace des caractéristiques:

- ➔ Une forme est représentée par un vecteur de $\mathbb{R}^d$ (ensemble des d mesures effectuées sur la forme)



⇒ une forme est un point dans l'espace $\mathbb{R}^d$

- ➔ Nécessité de définir une métrique pour pouvoir décider
 - distance d'un point à un autre, à une classe, à une frontière entre classes
 - probabilité $p(\text{classe}/\text{vecteur})$

Apprentissage / Décision Statistique

Apprentissage:

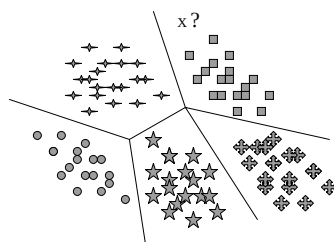
Etablissement par le système des frontières de séparation entre classes
classes ⇒ régionnement dans $\mathbb{R}^N$

Décision:

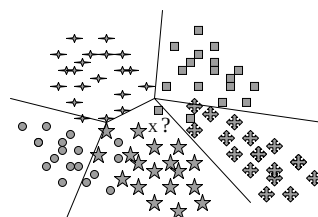
Affectation d'une classe d'appartenance en fonction de la position par rapport aux frontières

régionnement dans $\mathbb{R}^N$ ⇒ classe

Cas facile



Cas difficile



Théorie Bayésienne de la Décision (1)

Notations:

$w \in \Omega$: ensemble des classes à reconnaître

$x \in \mathbb{R}^N$: espace de représentation

$p(x)$: probabilité d'observer x

$p(w/x)$: probabilité d'appartenance à la classe w
conditionnellement au fait d'être en x

$d : \mathbb{R}^N \rightarrow \Omega$: règle de décision

$x \rightarrow d(x)=w_k$

(avec $\int_{\mathbb{R}^N} p(x)dx = 1$ et $\forall x, \sum_k p(w_k/x) = 1$)

Règle de décision bayésienne (d_1):

fonction qui à chaque objet x associe la classe la plus probable au point x :

$d_1(x)=w \Rightarrow \forall w' \in \Omega, p(w/x) \geq p(w'/x)$

Rmq: si les caractéristiques sont suffisamment descriptives alors:

$p(w/x)=1$ ou $p(w/x)=0$

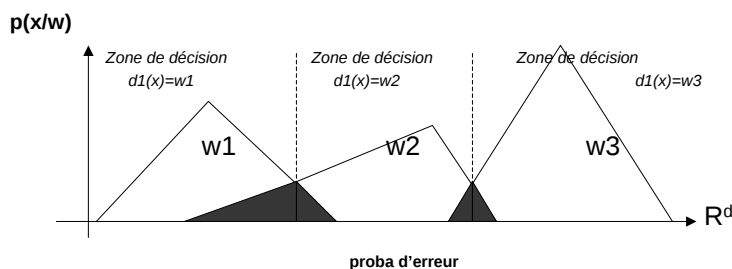
Théorie Bayésienne de la Décision (2)

Propriété: La fonction de décision bayésienne minimise la probabilité globale d'erreur du système

Rappel: La probabilité d'erreur d'une règle d est :

$$\text{Err}(d) = \int_{\mathbb{R}^N} [1-p(d(x)/x)]p(x)dx$$

Règle: A chaque point x , associer la classe dont la densité de probabilité en x est la plus forte (maximiser la proba de ce qu'on observe)



Théorie Bayésienne de la Décision (3)

Théorème de Bayes:

Utilité: Expérience (apprentissage) $\Rightarrow p(x/w)$ ou $p(x,w)$

Décision $\Leftarrow p(w/x)$

Enoncé: $p(w/x) = [p(x/w)p(w)] / p(x)$

car: $p(x,w)=p(x/w)p(w)=p(w/x)p(x)$

Exemple: $n=100$ observations

5 occurrences possibles pour x : $x_1, x_2, \dots, x_5$

4 classes possibles: w_1, w_2, w_3, w_4

	x_1	x_2	x_3	x_4	x_5
w_1	2	20	0	1	0
w_2	0	2	0	12	13
w_3	6	4	8	0	2
w_4	20	0	8	0	2

$p(w_1)=23/100$; $p(x_3)=16/100$; $p(x_5,w_4)=2/100$

$p(w_1/x_1)=2/28$; $p(w_2/x_1)=0$;
 $p(w_3/x_1)=6/28$; $p(w_4/x_1)=20/28$;
 $\Rightarrow$ si j'observe x_1 , je décide w_4

Méthodes Paramétriques / Non Paramétriques

Problème du modèle de base (TBD): difficile de mesurer les $p(x/w_i)$!!!

$\Rightarrow$ 2 types d'approches suivant les hypothèses retenues:

- méthodes paramétriques:

on fait appel à une famille paramétrable de densités ou de surfaces et on optimise le(s) paramètre(s) pour minimiser la probabilité d'erreur

- méthodes non paramétriques:

on développe un algorithme qui converge vers la densité ou la surface idéale quelle qu'elle soit (moyennant certaines hypothèses)

	Méthodes bayésiennes	Méthodes non bayésiennes
Méthodes paramétriques	Estimation de gaussiennes	Séparation linéaire
Méthodes non paramétriques	Fenêtres de Parzen	K plus proches voisins

Estimation de Gaussiennes (1)

Méthode paramétrique bayésienne

$$p(x/w_k) = (2\pi)^{-d/2} (\det \Sigma_k)^{-1/2} \exp[-0.5(x-m_k)^t \Sigma_k^{-1}(x-m_k)]$$

où: d dimension de l'espace de représentation
 Σ_k matrice de variance-covariance de la classe w_k
 m_k centre (moyenne) de la classe w_k

Apprentissage: estimer pour chaque classe w_k : m_k et Σ_k

Estimateur du Maximum de Vraisemblance

$$m_k = (1/n_k) \sum x_i$$

$$\Sigma_k = (1/n_k - 1) \sum (x_i - m_k)(x_i - m_k)^t \quad (\text{non biaisé})$$

Décision: appliquer la règle de Bayes

Choisir w_k qui maximise $p(w_k/x) = p(x/w_k)p(w_k)/p(x)$

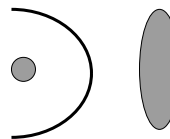
Estimation de Gaussiennes (2)

Choisir w_k qui maximise $p(w_k/x)$ revient à minimiser la distance du point au centre m_k (il suffit de changer de fonction discriminante!)

- cas général:

distance de Mahalanobis généralisée

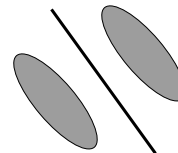
$$d_k^M(x) = (x-m_k)^t \Sigma_k^{-1}(x-m_k) + [\ln(\det \Sigma_k) - 2 \ln p(w_k)]$$



- même covariance Σ :

distance de Mahalanobis

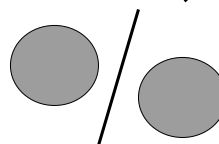
$$d_k^m(x) = (x-m_k)^t \Sigma^{-1}(x-m_k)$$



- gaussiennes isotropes:

distance euclidienne

$$d_k^e(x) = (x-m_k)^t (x-m_k)$$



Séparation Linéaire (1)

Méthode paramétrique non bayésienne

But : rechercher la meilleure surface séparatrice entre classes parmi une famille paramétrable de surfaces (ici: hyperplans)

Hypothèse : séparabilité entre les classes

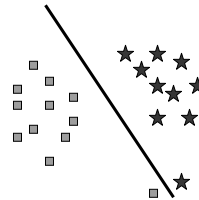
Apprentissage: Déterminer un hyperplan défini par un vecteur normal a et un scalaire a_0 tel que:

$$\forall x \in w_1, \sum_i a_i x_i + a_0 > 0$$

$$\forall x \in w_2, \sum_i a_i x_i + a_0 < 0$$

$(a_0, \dots, a_n)$ vecteur de paramètres à optimiser

Décision : positionnement du point inconnu par rapport aux surfaces



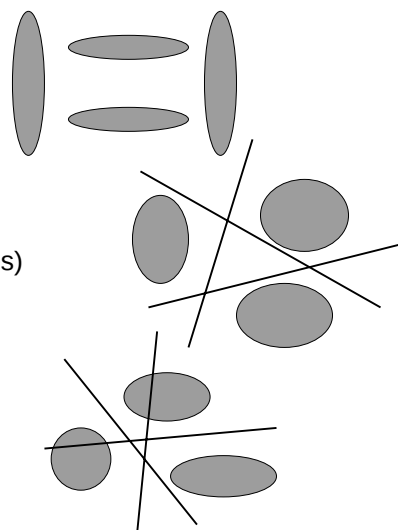
Séparation Linéaire (2)

Théoriquement, s'il y a $K=2^p$ classes, il faut p hyperplans pour les discriminer!

En pratique: difficile!!!

K classes $\Rightarrow K$ hyperplans:
(une classe contre toutes les autres)

K classes $\Rightarrow K(K-1)/2$ hyperplans:
(classes deux à deux)



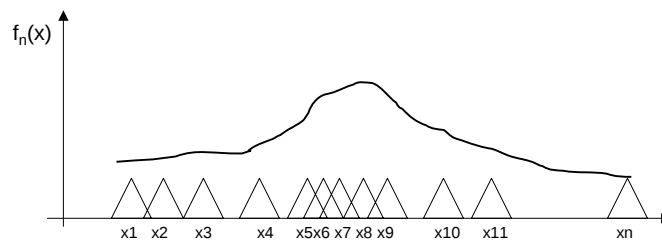
Fenêtres de Parzen

Méthode bayésienne non paramétrique

But : approximer $p(x,w)$ par une somme de fonctions ϕ centrées autour des points d'apprentissage puis appliquer la règle de Bayes

Exemples de fonctions ϕ : gaussienne, uniforme, triangle, ...

Principe: $x_1, x_2, \dots, x_n \Rightarrow f_n(x) = 1/n \sum_i \phi(x-x_i) = \hat{p}(x,w)$
 n représentations d'éléments de w (apprentissage) x : représentation d'un élément inconnu (décision)



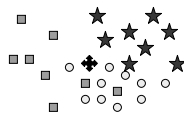
➔ Gros volume de calcul lors de la reconnaissance

K Plus Proches Voisins (1)

Méthode non bayésienne non paramétrique

Principe : faire voter les K plus proches voisins de x pour savoir à quelle classe x appartient (K fixé)

Exemple :



Propriété remarquable : $Err^* \leq \lim_{N \rightarrow \infty} Err_{1ppv} \leq 2 Err^*$

Avantages / inconvénients :

- + possibilité de faire du rejet ($K/2 \leq m \leq K$)
- + mise en œuvre très simple
- + convergence indépendante de la distance
- + possibilité de tenir compte des coûts
- complexité de l'algorithme de décision proportionnel à N
- performances mal connues quand N est fini

K Plus Proches Voisins (2)

Variantes accélérées sub-optimales:

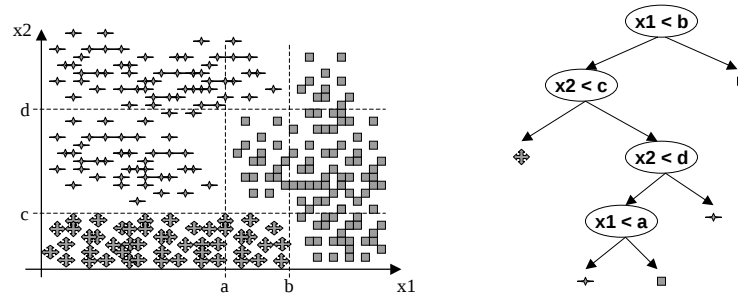
- Kppv avec élimination (édition)
prétraitement de la base d'apprentissage: élimination de tous les éléments dont le ppv appartient à une autre classe
- Condensation
réduction de l'espace de représentation à un sous-espace ayant les mêmes propriétés (par ACP par exemple)
- Pavage
l'espace est divisé en cellules et la recherche est limitée à quelques cellules: décision en 2 passes
- Tri sur ou plusieurs coordonnées
limite la recherche à quelques dimensions
- Kppv hiérarchiques
organisation des points sur plusieurs niveaux de finesse

Laquelle Choisir ???

Méthode	Champs d'application	Convergence	Vol. de calcul (décision)	Vol. à stocker (apprentissage)
Parzen	Toute distribution	Lente, sûre	Énorme $N \cdot I$	Énorme $N \cdot I$
K-ppv	Toute distribution	Lente, sûre	Énorme $N \cdot I$	Énorme $N \cdot I$
Paramétrique gaussien	Très limité	Rapide (si...)	Faible $J \cdot I^2$	Faible $J \cdot I^2$
Séparation linéaire	Limité	Rapide (si...)	Très faible $1.5 \cdot J \cdot I$	Faible $I \cdot J(J-1)/2$

I: nombre de caractéristiques; J: nombre de classes;
N: nombre d'éléments en apprentissage

Arbres de Décision

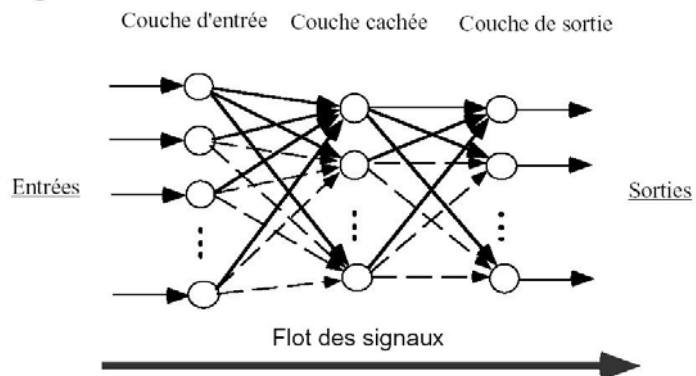


- + construction automatique / ordonnancement des tests
- + possibilité de pondérer les arcs par des probas
- + combinaison de caractéristiques
- + interprétation simple
- structure trop proche des données (sensibilité)
- difficultés à « balancer » l'arbre

Réseaux de Neurones (1)

A passe avant (e.g. Perceptron multi-couches)

- ❑ Fonctionnement en reconnaissance
- ❑ Apprentissage ?

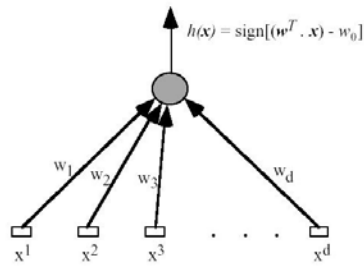
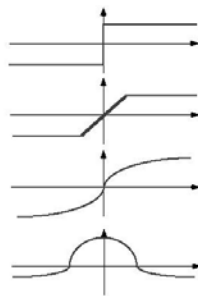


Réseaux de Neurones (2)

Neurone de base:

Calcul :
$$h(x) = g\left(\sum_{1 \leq i \leq d} w_i x^i - w_0\right) = g(a)$$

g : fonction d'activation
(Généralement non linéaire)



Réseaux de Neurones (3)

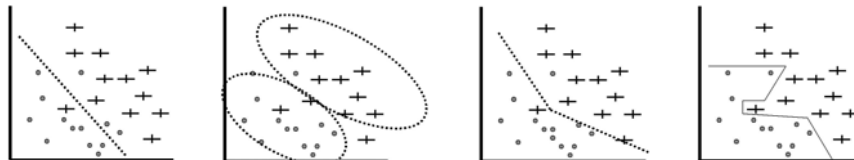
Apprentissage:

- 1. Présentation d'un exemple parmi l'ensemble d'apprentissage**
Séquentielle, aléatoire, en fonction d'un critère donné
- 2. Calcul de l'état du réseau**
- 3. Calcul de l'erreur = fct(sortie - sortie désirée)** (e.g. $= (y^k - t^k)^2$)
- 4. Calcul des gradients**
Par l'algorithme de rétro-propagation de gradient
- 5. Modification des poids synaptiques**
- 6. Critère d'arrêt**
Sur l'erreur. Nombre de présentation d'exemples, ...
- 7. Retour en 1**

Réseaux de Neurones (4)

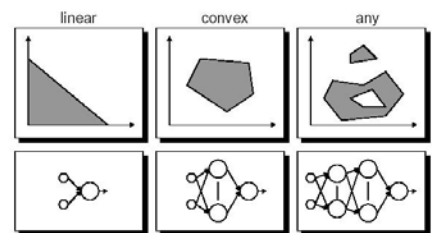
Avantages:

- Aptitude naturelle à fournir des surfaces de décision complexes (par extraction de combinaison non linéaire des caractéristiques)
- Capacités d'apprentissage fortement autonomes



Inconvénients:

Effet « boîte noire »



Université de Rouen

Combinaison de Classifieurs – M2 - L. Heutte

27

Evaluation d'un Classifieur

Bon classifieur = bonne généralisation

capacité à reconnaître des formes non déjà présentées au système

Comment évaluer un classifieur?

- il faut des données (en qualité et en quantité) pour apprendre...
- il faut des données « inconnues » pour évaluer le classifieur...
 - mesures globales: taux de reconnaissance, de confusion, de rejet
 - mesures semi-globales: taux de reconnaissance dans une liste
 - mesures locales: matrices de confusion
- il faudrait théoriquement 3 jeux de données différents...
 - apprentissage, test, évaluation
- ... mais on peut faire aussi du cross-validation, du leave-one-out, ...

Université de Rouen

Combinaison de Classifieurs – M2 - L. Heutte

28

Conclusion

- Une grande variété d'extracteurs de caractéristiques
 - Les caractéristiques peuvent être binaires, discrètes, réelles, locales, globales
 - Elles dépendent de l'application: expertise dans le choix...
- Une grande variété de classifieurs
 - Des structures différentes donc des surfaces de décision différentes
 - Des paramètres différents donc des surfaces de décision différentes
- Le choix est difficile: « choisir, c'est renoncer »
- L'optimisation d'un seul classifieur dépend des exemples d'apprentissage, de validation et de test
- Solution: Combinaison de classifieurs
 - On limite l'influence des choix initiaux
 - On utilise la complémentarité des approches
 - On fiabilise la prise de décision...

Bibliographie

Ouvrages de base (2^{ème} cycle) :

Reconnaissance des Formes : Méthodes et Applications

A. Belaïd, Y. Belaïd, InterEditions, Paris, 1992

Reconnaissance des Formes et Analyse de Scènes, vol. 3

M. Kunt et al., Presses Polytechniques et Universitaires Romandes, 2000

Reconnaissance des Formes : Méthodes Numériques et Connexionnistes

M. Milgram, Armand Colin Editeur, Paris, 1993

Méthodes Structurelles pour la Reconnaissance des Formes

L. Miclet, Eyrolles, Paris, 1984

Exercices de Reconnaissance des Formes par Ordinateur

P. Fabre, Masson, Paris, 1989

Pour approfondir (3^{ème} cycle) :

Introduction to Statistical Pattern Recognition, Second Edition

K. Fukunaga, Academic Press, 1990

Neural Networks for Pattern Recognition

C.M. Bishop, Clarendon Press, Oxford, 1998

Pattern Classification and Scene Analysis

R.O. Duda, P.E. Hart, John Wiley & Sons, 1973

Introduction to Pattern Recognition: Statistical, Structural, Neural and Fuzzy Logic Approaches

M. Friedman, A. Kandel, World Scientific, 1999