



Apprentissage Automatique

Introduction-II

jean-francois.bonastre@univ-avignon.fr

www.lia.univ-avignon.fr



I Un cadre probabiliste

- ◆ Définitions
- ◆ Formalisation
- ◆ Notations
- ◆ Règles de décision
- ◆ Erreur de classification
- ◆ Remarques



J.F. Bonastre

2



Définitions

- ◆ Trois objets
 - Les éléments
 - Les classes
 - Les descripteurs
- ◆ On suppose qu'il existe un classement correct, soit une application qui associe une classe à tout élément
- ◆ Apprendre = Associer une classe à une liste de descripteurs de telle manière que cette association corresponde au classement défini ci-dessus.



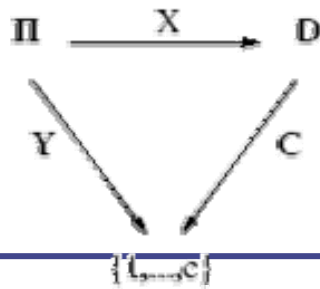
Formalisation (1)

- ◆ Π est la population, D est l'ensemble des descriptions, et l'ensemble des classes est $\{1, \dots, c\}$.
- ◆ $X: \Pi \rightarrow D$ est la fonction qui associe une description à tout élément de la population.
- ◆ $Y: \Pi \rightarrow \{1, \dots, c\}$ est la fonction de classement qui associe une classe à tout élément de la population.
- ◆ une fonction $C: D \rightarrow \{1, \dots, c\}$ sera appelée *fonction de classement* ou *procédure de classification*.



Formalisation (2)

- ◆ Le but de l'apprentissage est de rechercher une procédure de classification C , $C \circ X = Y$
- ◆ De manière plus réaliste, telle que $C \circ X$ soit une bonne approximation de Y



J.F. Bonastre

5



Formalisation (3)

- ◆ Ensemble d'attributs A_1, \dots, A_n logiques, symboliques ou numériques qui prennent leurs valeurs dans des domaines D_1, \dots, D_n
- ◆ En parole, les attributs sont souvent les différents coefficients des vecteurs acoustiques
- ◆ Décrire un élément de la population = attribuer une valeur à chacun de ces attributs.



J.F. Bonastre

6



Notations

- ✦ $P(d)$ la probabilité qu'un élément de Π ait d pour description, soit encore $P(d) = P(X^1(d))$
- ✦ $P(k)$ la prob. qu'un élément de Π soit de classe k , soit encore $P(k) = P(Y^1(k))$
- ✦ $P(d/k)$ la prob. qu'un élément de classe k ait d pour description, soit encore $P(d/k) = P(X^1(d) / Y^1(k))$ (définie si $P(k)$ est non nulle)
- ✦ $P(k/d)$ la prob. qu'un élément ayant d pour description soit de classe k , soit encore $P(k/d) = P(Y^1(k) / X^1(d))$. Définie si $P(d/k)$ est non nulle; par Bayes :

$$P(k/d) = P(d/k)P(k) / P(d)$$
- ✦ (Π a été « probabilisé » et D est estimé discret)



Règles de décision

- ✦ Règle Majoritaire
attribuer à chaque description la classe majoritaire
 - Cmaj associe à tout élément d de D la classe k de $\{1, \dots, c\}$ telle que $P(k)$ soit maximum
- ✦ Règle du Maximum de Vraisemblance (ML)
si j'observe d , je choisis la classe pour laquelle cette observation est la plus probable
 - Cvraisemblance associe à d la classe k telle que $P(d/k)$ soit max.
- ✦ Règle de Bayes (MAP)
 d se voit attribuer la classe k qui max. la probabilité $P(k/d)$
 - Par Bayes : choisir la classe k qui maximise le produit $P(d/k)P(k)$
 - CBayes associe à tout élément d la classe k telle que $P(k/d)$ soit maximum, i.e. $P(d/k)P(k)$ soit maximum



Erreur de classification

- ◆ $E(d)$, la probabilité qu'un élément de la population Π de description d soit mal classé par C
 - $E(d) = P(Y \neq X / X=d)$
- ◆ $E(C)$ est la moyenne pondérée des erreurs sur les descriptions d
 - $E(C) = \sum_{d \in D} E(d) \cdot P(X=d)$
- ◆ CBayes -> Erreur de classification minimale

$E(d) = P(Y \neq X / X=d) = 1 - P(Y = X / X=d)$
CBayes maximise $P(Y = k / X=d)$



Remarques

- ◆ $E(C)=0 \rightarrow E(CBayes)=0$
 - La probabilité que des individus appartenant à des classes différentes aient des descriptions identiques est nulle.
 - Problème déterministe
 - Rare !
- ◆ En parole, bruits, variabilités...



II Problèmes et généralités

- ◆ Supervisé vs non Supervisé
- ◆ Hiérarchique ou non
- ◆ Rescaling
- ◆ Sélection des paramètres
- ◆ Partitionnement hard ou soft
- ◆ Nombre de classes
- ◆ Qualité d'une partition/segmentation



Supervisé vs non supervisé (1)

- ◆ On observe un phénomène régit par des lois inconnues
 - Des données observées aux connaissances
- ◆ Supervisé :
 - Les données sont accompagnées de la connaissance à inférer (classe ou valeur)
 - On va vouloir généraliser à d'autres ensembles de données
- ◆ Non supervisé
 - Les données seules sont observées



Supervisé vs non supervisé (2)

- ◆ Approche orientée connaissance
 - On utilise les connaissances d'un expert
- ◆ Approche à apprentissage à partir d'exemples
 - La procédure de classification est extraite automatiquement à partir d'un ensemble de couples (exemple, classe)
 - Problème de généralisation et de sur apprentissage
- ◆ Approche à apprentissage non supervisé (clustering/ partitionnement)
 - Séparer en classes un ensemble de données
 - Métrique, Nombre de classes
 - A priori sur les classes (méthodes paramétriques) ou non (méthodes non paramétriques)

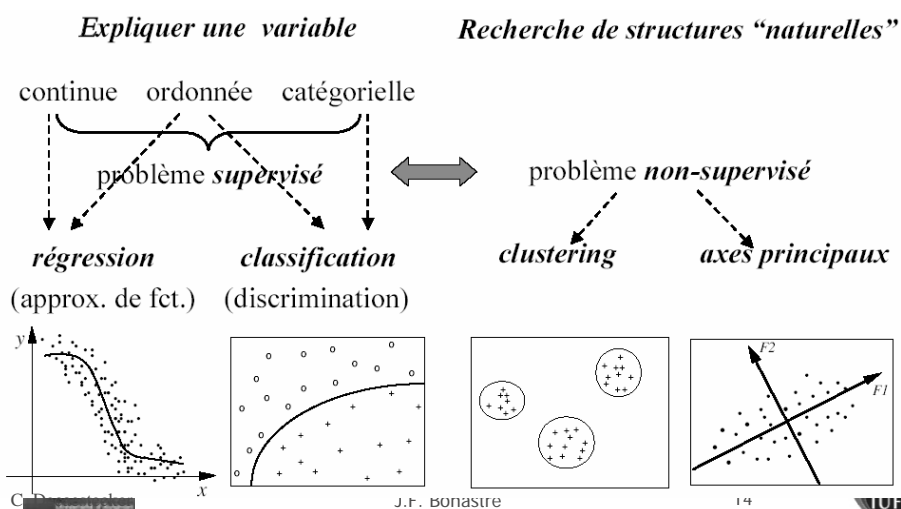


J.F. Bonastre

13



Supervisé vs non supervisé (3)



Hierarchique ou non ?

◆ Algorithmes hiérarchiques

- itératif
- Ascendant : on regroupe des classes à chaque étape
- Descendant : on coupe des classes
- Efficace (toutes les part.)
- Mais
 - ◆ un a priori non remis en question
 - ◆ partitionnement

◆ Algorithmes non hiérarchiques

- Toutes les classes sont calculées/optimisées simultanément
- Peu efficaces car tous les éléments sont utilisés
- Doit être recalculé pour chaque nombre de classe
- Une erreur peut être corrigée
- Décision Soft ou Hard



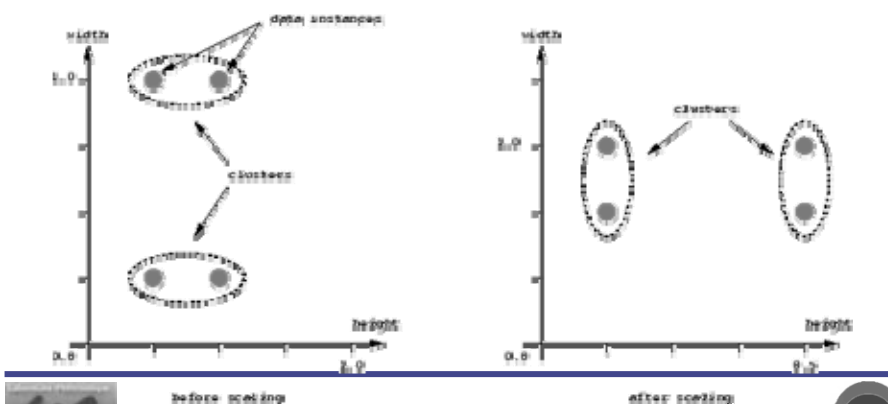
J.F. Bonastre

15



Rescaling

La proximité entre deux éléments dépend du point de vue...



J.F. Bonastre

16



Sélection des paramètres (1)

- ◆ Un élément est représenté par un descripteur
 - Dimension du descripteur
 - Nature du descripteur
- ◆ En parole, étape de paramétrisation acoustique : du signal vers un vecteur de paramètre par fenêtre temporelle



Sélection des paramètres (2)

- ◆ Problème
 - La dimension peut être très grande (~100 en parole, ~10000 en texte)
 - Grande dimension = coûteux
 - Souvent, il y a du « bruit »
 - ◆ Perte d'efficacité
 - ◆ Perte du potentiel d'interprétation
- ◆ La capacité du classifieur dépend des paramètres
 - > Sélectionner les meilleurs paramètres



Sélection des paramètres (3)

◆ Deux approches

- Sélectionner les meilleurs paramètres suivant un critère a priori, indépendant du problème (filtrage)
- Sélectionner le meilleur sous ensemble en fonction des résultats pour le problème visé (sélection)

◆ Dans tous les cas, le nombre de combinaison est :

$$\sum_{p=1}^n C_n^p = \sum_{p=1}^n \frac{n!}{p!(n-p)!}$$



J.F. Bonastre

19



Sélection des paramètres (4) filtrage

◆ Diminuer la corrélation

- PCA
- LDA
- ICA

◆ Meilleure corrélation avec la cible

- Corrélation linéaire
- Mutual Information



$$I(x, y) = \sum_i \sum_j p(x_i, y_j) \log \left[\frac{p(x_i, y_j)}{p(x_i)p(y_j)} \right]$$

J.F. Bonastre

20



Sélection des paramètres (5)

Sélection descendante

◆ Algorithme de knock-out

- On a un ensemble N de n paramètres
 - ◆ Construire n sets, N_i , en retirant de N le i ème élément
 - ◆ Faire n expériences (apprentissage+validation !)
 - ◆ Sélectionner le i ème paramètre, correspondant au set N_i avec les moins bonnes performances (le i ème paramètre n'était pas dans le set N_i)
 - ◆ Remplacer N par N_i
 - ◆ Répéter
- Cher !



Sélection des paramètres (6)

Sélection Ascendante

◆ Partir avec un sous ensemble Q de dimension q

- Réaliser $n-q$ sets en ajoutant un paramètre
- Mesurer les performances (app + validation)
- Sélectionner le paramètre menant à la meilleur performance et le rajouter dans Q
- Répéter

◆ Attention à l'initialisation

- Valeur pour q
- Répéter l'algorithme pour chaque sous ensemble Q initial (de dimension q)



Sélection des paramètres (7)

Autres solutions

- ◆ Pondérer (en continu) les paramètres
 - Descente de gradient
 - Algorithmes génétiques
- ◆ Arbres de décision
- ◆ ...



J.F. Bonastre

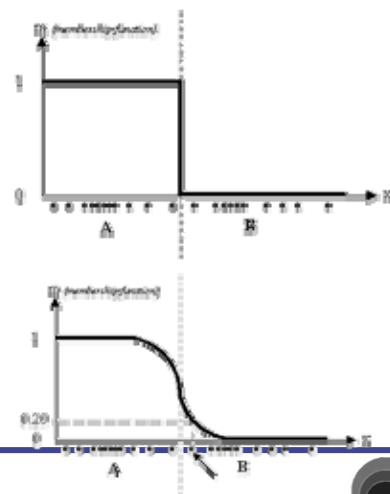
23



Affectation hard ou soft

- ◆ La décision d'affectation d'un élément à une classe peut être binaire ou continue (fuzzy)

Kmeans vs FCM



J.F. Bonastre

24



Nombre de classes (1)

- ◆ Un problème ouvert, très difficile
- ◆ Incontournable en classification non supervisée
- ◆ Mais aussi présent en supervisé, pour décider par exemple de la complexité d'un modèle (GMM)
- ◆ Approches
 - Inertie/Entropie
 - MDL
 - Pureté
 - Bayésienne (BIC)



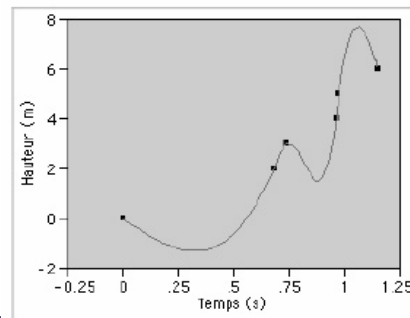
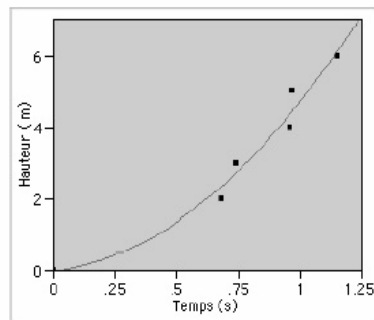
Nombre de classes (2)

- ◆ Première approche
 - essayer n solutions avec un nombre de classe différent
 - Mesurer la « qualité » de chacune des solutions et choisir la meilleure (BIC ou inertie)
- ◆ Deuxième approche (classification hiérarchique)
 - Optimiser un paramètre pour la coupure ou l'élagage (pureté estimée ou entropie)
 - Revient à la 1ère approche car tous les regroupements, de 1 à n classes, sont déjà effectués



Surapprentissage (1)

◆ Galilée – Temps de chute d'un objet - Loi quadratique



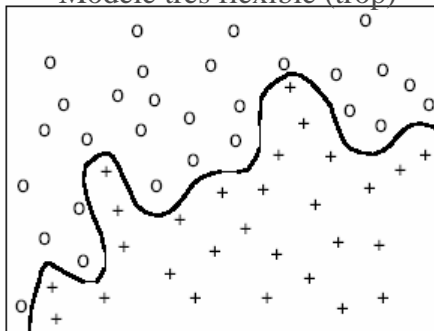
J.F. Bonastre

27

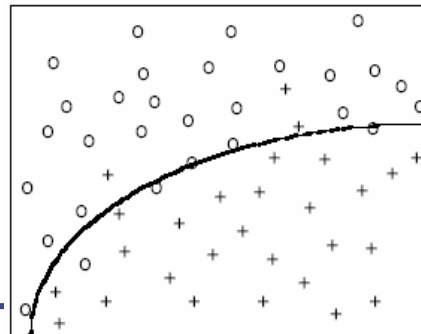


Surapprentissage (2)

Modèle très flexible (trop)



Modèle moins flexible



J.F. Bonastre

28



Qualité

◆ Validation expérimentale



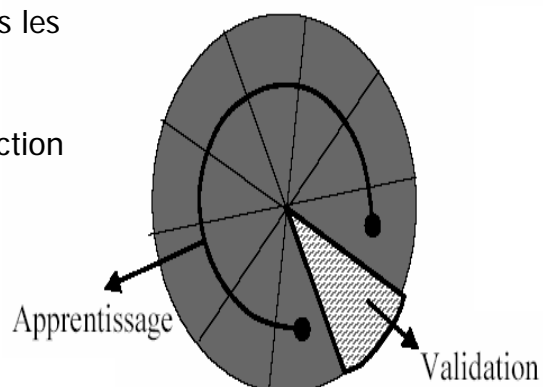
J.F. Bonastre

29



Validation expérimentale

- ◆ Kfold (Jack knifing)
- ◆ Applicable à toutes les méthodes
- ◆ Bootstrap : même principe mais sélection aléatoire



J.F. Bonastre

30



III Quelques algorithmes et approches

- ◆ Regroupement Hiérarchique
- ◆ Kmeans
- ◆ Isodata
- ◆ Fuzzy C Means
- ◆ KNN
- ◆ Les mixtures de Gaussiennes en classification non supervisée
- ◆ Le classifieur naïf de Bayes
- ◆ Les arbres de décisions
- ◆ Evaluation des performances



J.F. Bonastre

31



Regroupement Hiérarchique (1) Principes

- ◆ Non supervisé
- ◆ Au départ, chaque élément constitue une classe
- ◆ Iteratif : regroupement à chaque étape des deux classes les plus proches
- ◆ Problèmes
 - Distance inter classes
 - Distance/Similarité inter éléments
 - Nombre de classe



J.F. Bonastre

32



Regroupement Hiérarchique (2)

Algorithme

- ◆ $D = [d(i,j)]$ est la matrice de proximité inter éléments ($N \times N$)
- ◆ Les regroupements sont numérotés séquentiellement : $0, 1, \dots, (n-1)$ et $L(k)$ est le degré du kème regroupement
- ◆ $D(m)$: une classe de numéro de séquence m
- ◆ $d[(r),(s)]$ est la proximité entre les classes (r) et (s)



Regroupement Hiérarchique (3)

Algorithme

1. *Début avec une classe par élément, degré de regroupement $L(0)=0$, num. de séquence $m=0$*
2. *Trouver les 2 classes (r) et (s) telles que $d[(r),(s)] = \min d[(i),(j)]$*
3. *$m = m + 1$, on regroupe (r) et (s) , le degré de regroupement est fixé à : $L(m) = d[(r),(s)]$*
4. *Si une seule classe, stop, sinon aller en 2*



Regroupement Hiérarchique (4)

Distances et similarités

- ◆ La matrice de similarité (ou de dissimilarité) inter-éléments n'est jamais remise en cause (en général, pour des raisons d'efficacité)
- ◆ La distance inter classes correspond à
 - Single Linkage
 - Complete Linkage
 - Average Linkage
 - Ward (inertie)
 - ...



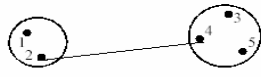

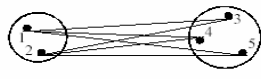

Regroupement Hiérarchique (5)

Distances et similarités

- ◆ La mesure de similarité respecte
 - $\text{sim}(u,v) = \text{sim}(v,u)$
 - $\text{sim}(u,v) > 0$ si u différent de v
 - $\text{sim}(u,u) = 0$
- ◆ En parole
 - GLR/BIC – cher
 - Cross Likelihood ratio
 - Cross entropy



Regroupement Hiérarchique (6) Linkage

élément A	élément B	$d(A, B)$	critère
		$d(2, 4)$	saut minimal
		$d(1, 5)$	saut maximal
		$\frac{1}{6}(d_{13} + d_{14} + d_{15} + d_{23} + d_{24} + d_{25})$	distance moyenne
		$I_{AB} - (I_A + I_B)$	inertie

⇒ Variation importante des résultats (dendrogramme) en fonction du critère choisi ! (sauf si les groupes sont bien distincts)



J.F. Bonastre

37



Regroupement Hiérarchique (7) Représentation

- ◆ Représentation sous forme d'un arbre de regroupement
- ◆ Si la longueur des branches est proportionnelle au degré de regroupement : Dendrogramme

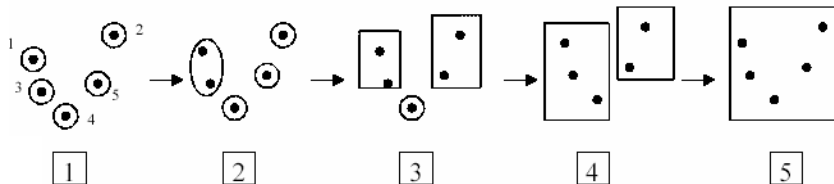


J.F. Bonastre

38



Regroupement Hiérarchique (8) Exemple



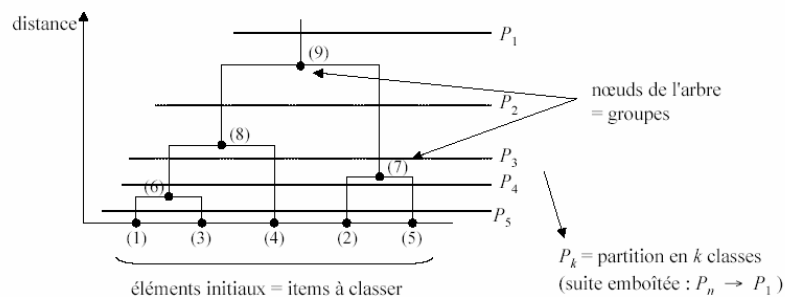
J.F. Bonastre

39



Regroupement Hiérarchique (9) Exemple

Représentation sous forme de $\left\{ \begin{array}{l} \text{arbre hiérarchique} \\ \text{ou} \\ \text{"dendrogramme"} \end{array} \right.$



\Rightarrow hauteur des branches = distances entre les 2 éléments regroupés.

\Rightarrow de + en + hautes en montant dans la hiérarchie.



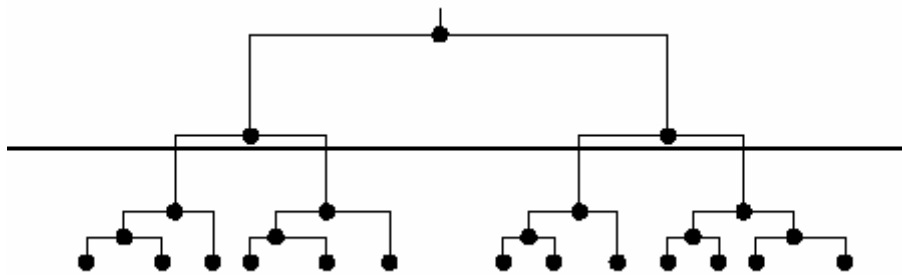
J.F. Bonastre

40



Regroupement Hiérarchique (10)

Coupure



- ◆ Avant les gros !
- ◆ Ou élagage ?



J.F. Bonastre

41



Kmeans (1)

KMoyennes

- ◆ **entrée** le nombre k de groupes (classes) et les données, m enregistrements $x_1^{\rightarrow}, \dots, x_m^{\rightarrow}$
1. choisir k centres initiaux $c_1^{\rightarrow}, \dots, c_k^{\rightarrow}$
 2. pour chacun des m enregistrements, l'affecter au groupe i dont le centre c_i^{\rightarrow} est le plus proche
 3. si aucun élément ne change de groupe alors arrêt et sortir les groupes
 4. calculer les nouveaux centres : pour tout i , c_i^{\rightarrow} est la moyenne des éléments du groupe i
 5. aller en 2



J.F. Bonastre

42



Kmeans (2)

Problèmes

- ◆ Mesure de similarité, entre deux enregistrements
- ◆ Nombre de classes ?
- ◆ Initialisation
- ◆ Partionnement



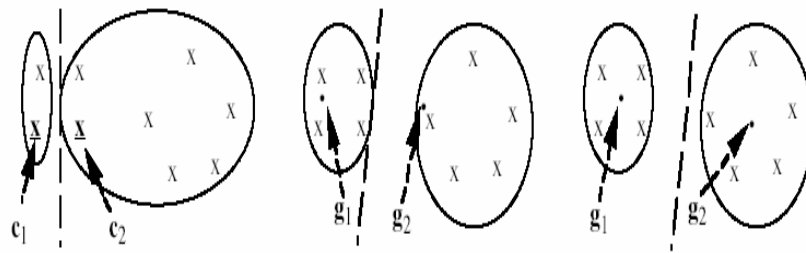
Kmeans (3)

Définir le nombre de classes

- ◆ Minimiser la distance intra groupe et maximiser la distance entre les groupes
- ◆ Distance ?
 - le rattachement simple, single linkage = la plus petite distance entre les éléments les plus proches ;
 - le rattachement complet, complete linkage = la distance entre les membres les plus éloignés
 - la distance entre les centres



Kmeans (4) Exemples



J.F. Bonastre

45



Isodata (1)

- ◆ Même algorithme que Kmeans
- ◆ Mais cherche à équilibrer les classes
- ◆ Fusion de deux groupes (diminution du nombre de classes) si la distance inter-centre est trop faible
- ◆ Éclatement d'un groupe si l'inertie du groupe est trop grande
- ◆ Seuils !!

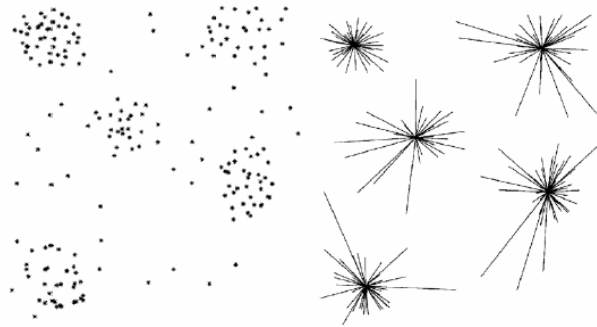


J.F. Bonastre

46



Isodata (2)



???



J.F. Bonastre

47



Fuzzy C Means

- ✦ Algorithme entre kmeans et les mixtures de gaussienne/EM. Minimise J_m

$$J_m = \sum_{i=1}^N \sum_{j=1}^C U_{i,j}^m \|x_i - c_j\|^2, 1 < m \leq \infty$$

- ✦ Avec U_{ij} l'appartenance d'une donnée x_i à la classe j

$$U_{ij} = \frac{1}{\sum_{k=1}^C \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}$$

- ✦ La mise à jour des centres est effectuée par :

$$C_j = \frac{\sum_{i=1}^N U_{i,j}^m x_i}{\sum_{i=1}^N U_{i,j}^m}$$



J.F. Bonastre

48



KNN (1)

N Plus Proches Voisins (nPPV)

- ◆ Méthode non paramétrique
 - pas de modèle
 - Les données sont l'information !
- ◆ **entrée** : $y \rightarrow$ l'élément considéré, k , le nombre de voisins, un échantillon de m enregistrements classés ($x \rightarrow, c(x \rightarrow)$)
 1. déterminer k plus proches enregistrements de $y \rightarrow$
 2. combiner les classes de ces k exemples en une classe c
- ◆ **sortie** : la classe de $y \rightarrow$ est c



KNN (2)

Usage

- ◆ Classe de sortie = étiquettes des k voisins
- ◆ Classe de sortie = vote majoritaire sur les k voisins
- ◆ Sortie = Moyenne/combinaison des k voisins
- ◆ Parole : voir thèse de F. Lefèvre



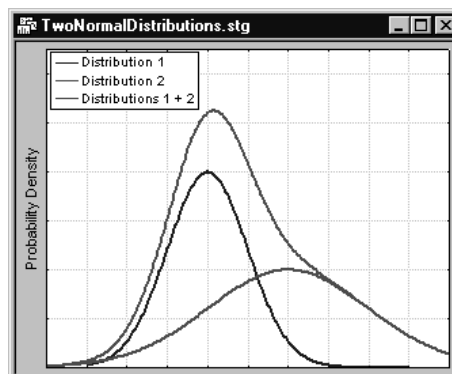
Les mixtures de Gaussiennes (1) en classification non supervisée

- ◆ Faire un classifieur à partir d'un estimateur de densité
 - Apprendre un estimateur de densité de probabilité sur l'ensemble des données
 - Cet estimateur est un mélange de densités plus simples
 - Chaque composante du mélange représente une classe
 - Un élément appartient à l'ensemble des classes, avec une probabilité d'appartenance
- ◆ Mélange de Gaussiennes -> Mixture de Gaussienne (GMM)



Les mixtures de gaussiennes (2) Exemple mono dimensionnel

- ◆ La courbe verte est la distribution des données
- ◆ Elle est approchée (ici exact !) par un mélange de deux composantes (bleu et rouge)



Les mixtures de Gaussiennes (3) en classification non supervisée

- ◆ Proche de kmeans mais
 - Une observation est rattachée à toutes les classes et non plus à la plus proche
 - Probabilité de classification vs classification
 - Utilisation de résultats formels
- ◆ On peut se ramener facilement à une classification en choisissant la classe la plus probable pour chaque obs.



Le classifieur naïf de Bayes

- ◆ $X = \{x_1, x_2, \dots, x_n\}$, un jeu de variable
- ◆ C_j appartenant à l'ensemble $C = \{c_1, c_2, \dots, c_k\}$, les classes
- ◆ On cherche la probabilité a posteriori de l'événement
- ◆ Avec Bayes :
$$p(C_j | x_1, \dots, x_n) = \frac{p(x_1, \dots, x_n | C_j) \cdot P(C_j)}{p(x_1, \dots, x_n)} \approx p(x_1, \dots, x_n | C_j) \cdot P(C_j)$$
- ◆ Hypothèse du classifieur naïf de Bayes : les variables sont statistiquement indépendantes :

$$p(C_j | x_1, \dots, x_n) = p(C_j | X) = p(C_j) \cdot \prod_{i=1}^n p(x_i | C_j)$$



GMM en classification supervisée

- ◆ Classifieur naïf des Bayes
- ◆ Estimation des probabilités par un GMM



Les arbres de décisions (1) Définition et propriétés

- ◆ Arbre de décision
 - Représentation graphique d'une procédure de classification
 - Noeud = test sur les champs ou attributs
 - Feuilles = classes (avec répétitions)
 - Classer un enregistrement = Descendre dans l'arbre selon les réponses aux différents tests
- ◆ Propriétés importantes des arbres de décision :
 - la procédure de classification associée est compréhensible et peut être justifiée
 - les attributs apparaissant dans l'arbre sont les attributs pertinents pour le problème de classification considéré
- ◆ Un arbre de décision est un système de règles exhaustives et mutuellement exclusives



Les arbres de décisions (2)

Avantages et Inconvénients

- ◆ lisibilité du résultat
- ◆ tout type de données
- ◆ sélection des variables
- ◆ classification efficace
- ◆ outils disponibles
- ◆ sensible au nombre de classes
- ◆ Algorithme non incrémental
- ◆ En parole/acoustique, voir R Blouet (reconnaissance du locuteur)



J.F. Bonastre

57



Classification automatique

Mesure des performances (1)

- ◆ Approche théorique
 - Calcul du risque théorique
- ◆ Approche expérimentale
 - Jeu de test **représentatif**
 - Métriques
 - ◆ Erreurs de type I et II
 - ◆ Précision/Rappel
 - ◆ Pureté



J.F. Bonastre

58



Classification automatique

Mesure des performances (2)

◆ Classification non supervisée – Clustering

◆ Pureté d'une classe m

$$p_m = \sum_{p=1}^m \left(\frac{n_{mp}}{N_{m*}} \right)^2$$

- n_{mp} le nombre de document dans m relevant de l'étiquette p
- N_{m*} le nombre total de document dans la classe m
- P, le nombre d'étiquettes

◆ Pureté d'une classification

$$\bar{p} = \frac{1}{N} \sum_{m=1}^M n_{m*} \cdot p_m$$



J.F. Bonastre

59



Classification automatique

Mesure des performances (3)

◆ Classification supervisée – Détection

- Erreur de Type I, False Alarm, Fausse Acceptation (FA)
- Erreur de type II, Miss, False Rejection, Faux Rejet (FR)
- Probabilité de détection = power = puissance
- Courbes FA, FR en fonction du seuil
- ROC, DET, Erreur de type II (ou power) en fonction des erreur de type I



J.F. Bonastre

60



Classification automatique Mesure des performances (4)

- ◆ Classification supervisée – Détection
 - Des points spéciaux EER, HTER, HTERmin
 - Une fonction de coût, en fonction des probabilités a priori et des coûts relatifs des erreurs

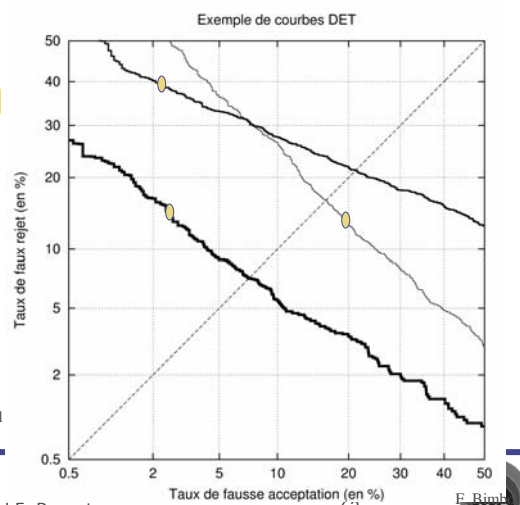
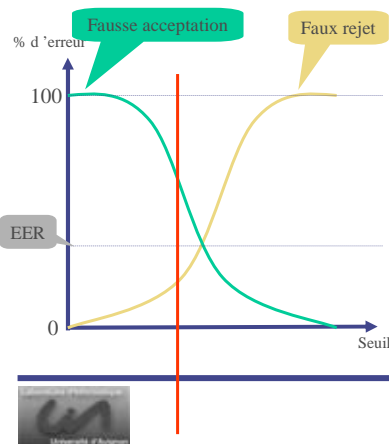


J.F. Bonastre

61



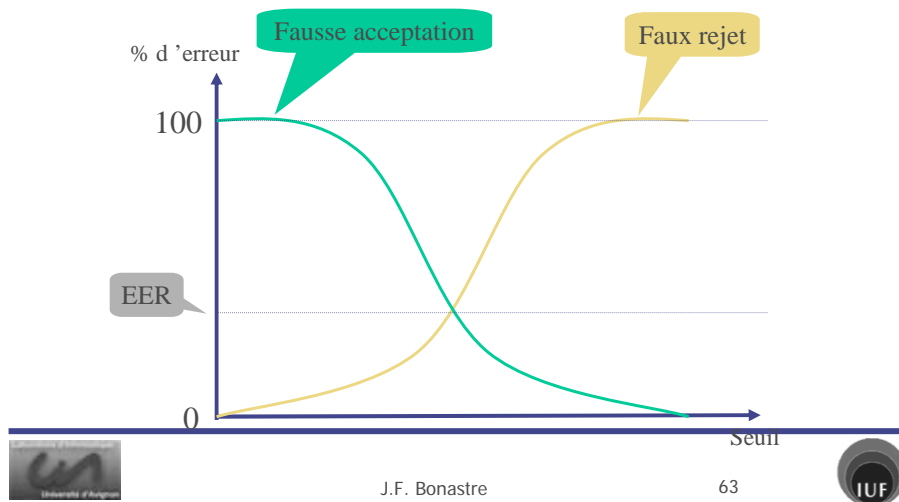
Classification automatique Mesure des performances (5)



J.F. Bonastre



Éléments d'évaluation -2



Classification automatique Mesure des performances (6)

◆ Classification supervisée – Recherche d'information

$$Rappel = \frac{\text{\#documents pertinents trouvés}}{\text{\# documents pertinents}}$$

$$Précision = \frac{\text{\#documents pertinents trouvés}}{\text{\# documents trouvés}}$$



Classification automatique

Mesure des performances (7)

- ◆ Classification supervisée – Recherche d'information
 - Rappel et Précision
 - Rappel en fonction de la précision
 - Précision à n documents
- ◆ Si détection, alors en nombre d'événements



Classification automatique

Mesure des performances (8)

- ◆ Problèmes !!!
 - Avoir une « vérité terrain »
 - Prendre en compte la longueur des documents
 - Une erreur de frontière doit coûter combien ?
 - ...

