

Classification supervisée pour données de dissimilitudes : une approche par maximum de vraisemblance




Guillaume Saint Pierre

Dépt. de mathématiques. Univ. Paris-Sud Orsay
Equipe Sélection de modèles et apprentissage statistique
INRIA – FUTURS



saint@cict.fr



La plupart des techniques de classification utilisent les coordonnées des objets, mais dans beaucoup de situations pratiques, seules les distances entre objets sont connues.

- Distances calculées entre objets de grande dimension (courbes, images ...)

⇒ distances euclidiennes

- Dissimilitudes issus d'expériences ou mesures entachées d'erreurs

⇒ distances non-euclidiennes




Développer des techniques de classification adaptées à ces différents type de situations



Travailler directement avec les distances

- Espace de représentation adéquat
- Méthodes de classification spécifiques

El Golli et al (2004), Guérin Dugué & Celeux (2001)



Se ramener à un espace de représentation euclidien et travailler avec l'estimation des coordonnées des objets

- Méthodes de classification classiques (supervisées ou non)

Approche Bayésienne : Oh & Raftery (2003),

K-PP : Fukunaga (1990)

Et bien d'autres ...

Algorithme de Multidimensional Scaling (MDS)

Reconstitue les coordonnées des objets à partir de mesures de distance (à qqes transformations près)

Mesures de
distances

$$d_{ij}$$


Coordonnées
dans un espace
Euclidien

$$\mathbf{X}_i, \mathbf{X}_j$$


La dimension de l'espace Euclidien doit être spécifiée

La dim 2 permet de visualiser les objets mais peut ne pas être assez fidèle aux données




Développer un outil de classification supervisée

- Nombre de classes connu
- Echantillon d'apprentissage



Utiliser une approche basée sur un modèle de mélanges gaussiens



Tenir compte de l'erreur de mesure



Déterminer le modèle le plus adéquat par des techniques de type maximum de vraisemblance

Un individu \mathbf{x}_i est décrit par ses coordonnées (x_{i1}, \dots, x_{ip})

Distance réelle entre \mathbf{x}_i et \mathbf{x}_j supposée euclidienne :

$$\delta_{ij} = \left(\sum_{k=1}^p (x_{ik} - x_{jk})^2 \right)^{1/2} = \|\mathbf{x}_i - \mathbf{x}_j\|$$

Hypothèse :

distance observée = distance réelle + erreur de mesure

$$d_{ij} \sim \mathcal{N}(\delta_{ij}, \sigma^2) \mathbb{1}_{(d_{ij} > 0)}, i \neq j, \quad i, j = 1, \dots, n$$

σ^2 paramètre d'erreur de mesure

On peut le supposer différent selon les types de distances

- Erreur intra-classes σ_{intra}^2
- Erreur inter-classes σ_{inter}^2

Indices des classes : $z_i \in \{1, \dots, K\}$

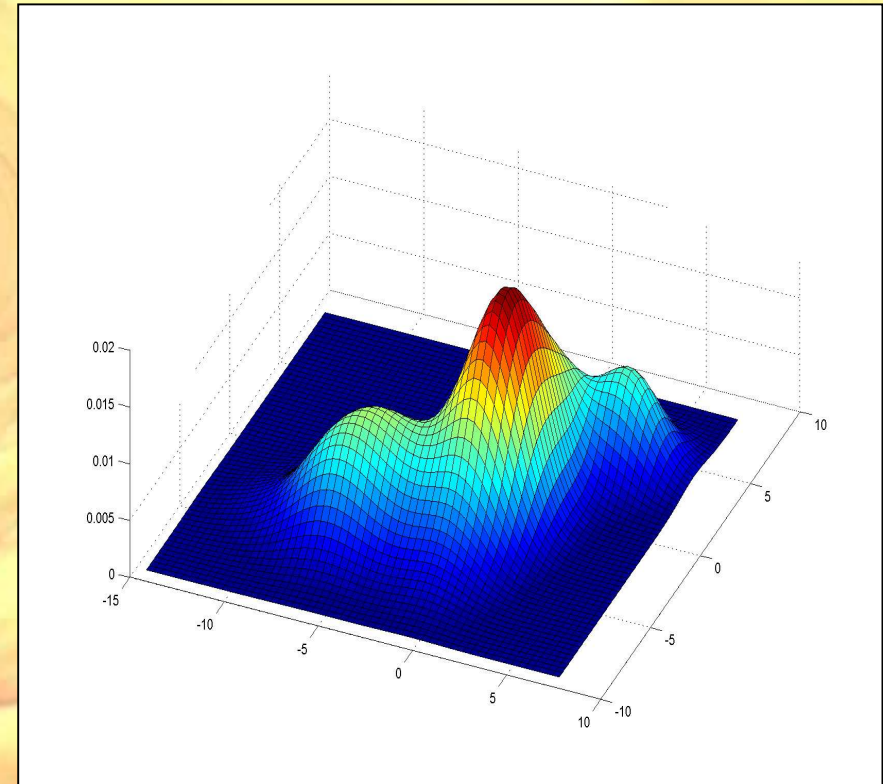
Nombre de classes (connu) : K

On suppose que les éléments x_i suivent la loi normale multivariée :

$$\mathbf{x}_i \sim \mathcal{N}(\mu_{z_i}, \Gamma_{z_i})$$

La distribution de \mathbf{x}, \mathbf{z} s'écrit :

$$f(\mathbf{x}, \mathbf{z} \mid \mathbf{p}, \mu, \Gamma) = \prod_{i=1}^n \sum_{l=1}^K \frac{p_l}{\sqrt{2\pi |\Gamma|}} \exp\left(-\frac{1}{2} \|\mathbf{x}_i - \mu_l\|_{\Gamma}^2\right) \mathbb{1}_{[z_i=l]}$$



$$\begin{aligned}
 f(\mathbf{d}, \mathbf{x}, \mathbf{z} \mid \sigma^2, \mathbf{p}, \mu, \Gamma) &= f(\mathbf{d} \mid \mathbf{x}, \mathbf{z}, \sigma^2) \times f(\mathbf{x} \mid \mathbf{z}, \mu, \Gamma) \times f(\mathbf{z} \mid \mathbf{p}) \\
 &= \prod_{i=1}^n \left\{ \left[\prod_{j>i}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (d_{ij} - \|\mathbf{x}_i - \mathbf{x}_j\|_M)^2\right) \Phi\left(\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_M}{\sigma}\right)^{-1} \right] \right. \\
 &\quad \left. \times \sum_{l=1}^K \frac{p_l}{\sqrt{(2\pi)^p |\Gamma_l|}} \exp\left(-\frac{1}{2} \|\mathbf{x}_i - \mu_l\|_{\Gamma_l^{-1}}^2\right) 1_{[z_i=l]} \right\}
 \end{aligned}$$



L'algorithme MDS nous fournit les \mathbf{x}_i

On obtient les paramètres du modèle en maximisant la vraisemblance



Ecriture similaire pour σ_{inter}^2 et σ_{intra}^2

- 1 ① Données : tableau de distances observé \mathbf{d}
Sélection des données d'apprentissage et des données à classer
- 2 ② Estimation des \mathbf{x}_i par MDS
- 3 ③ Etape d'apprentissage : Estimation des paramètres du modèle
 - Formules explicites pour \mathbf{p}, μ, Γ
 - Max. de Vraisemblance pour σ_{inter}^2 et σ_{intra}^2
- 4 ④ Etape de classement :
On affecte les données à la classe maximisant la Vraisemblance

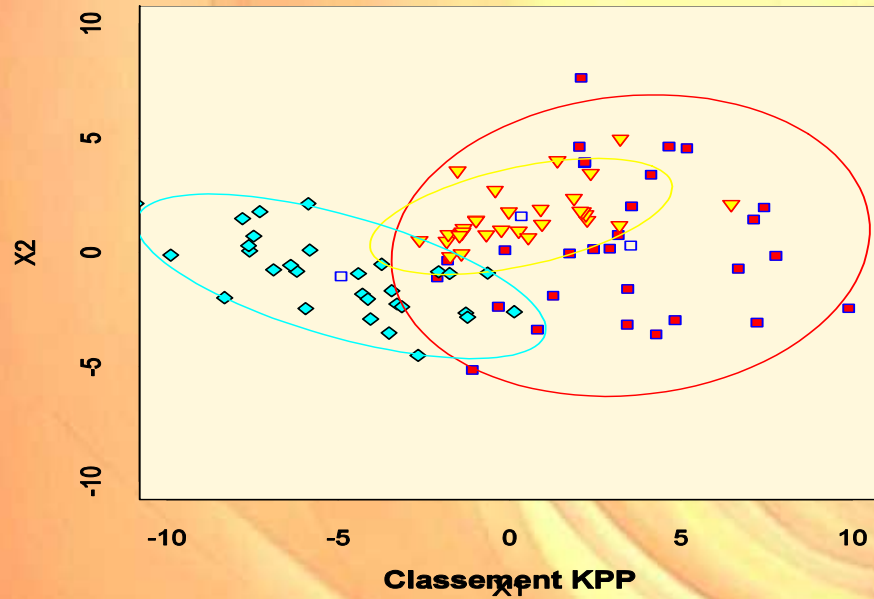
Simulations

- Génération d'un mélange gaussien à 3 composants de taille identique (200)
⇒ table de distances que l'on perturbe (tableau non-euclidien)
- Erreur de mesure intra-groupe : 0,5
- Erreur de mesure inter-groupe : 1
- KPP optimisé par validation croisée (entre 1 et 20 voisins)
- résultats moyens sur 50 échantillons d'apprentissage aléatoires

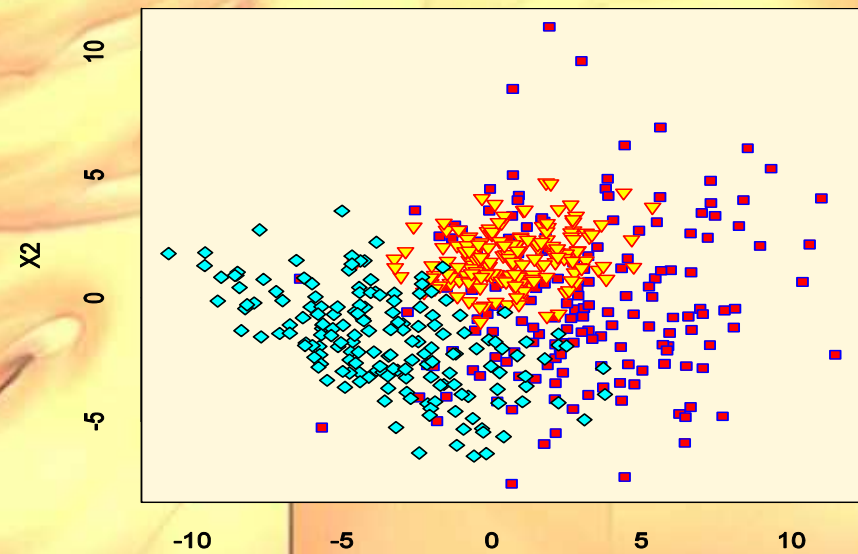
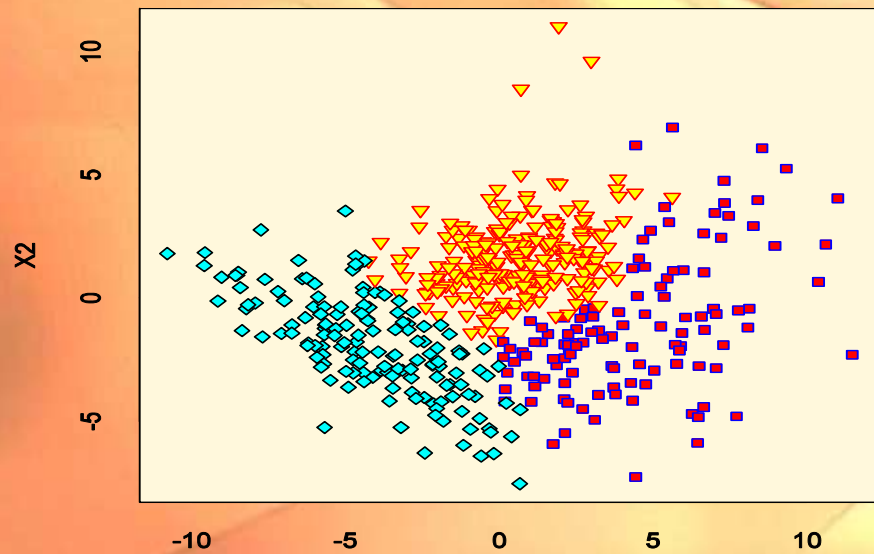
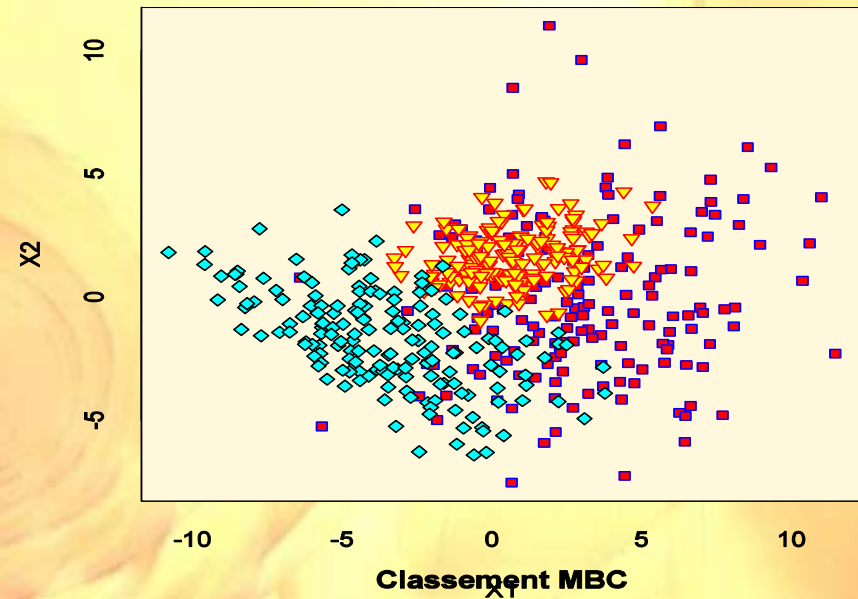
Taux de « bien classés »

	% de données utilisées pour l'apprentissage			
Méthode utilisée	5	10	15	20
MBC	92.38	99.36	99.82	99.99
KPP	76.63	78.14	80.07	81.18

Echantillon d'apprentissage et paramètres estimés



Classement original des données de test



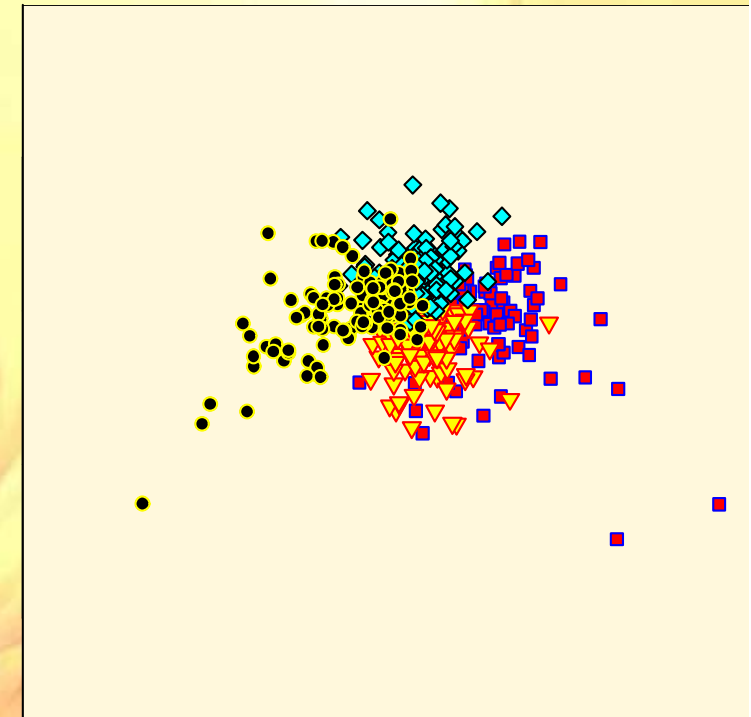
MDS sur données Images

Base de données « Images »

Base de données de 473 images


Dissimilarités calculées à partir d'histogrammes d'orientation en niveaux de gris.

4 classes différentes
Peu d'erreur de mesure
Hypothèse gaussienne peu respecté.




Taux de « bien classés »


	% de données utilisées pour l'apprentissage			
Méthode utilisée	5	10	15	20
MBC	68.25	73.3	75.33	75.40
KPP	69.04	73.71	75.69	76.43




Performances aussi bonnes que les K-PP sur des cas concrets
Meilleures lorsque le modèle est pertinent



Modélisation permettant de prendre en compte l'erreur de mesure.
⇒ information utile lorsqu'elle est présente au sein des données.



Modèle susceptible d'être raffiné
(une erreur de mesure par groupe)



Performances à étudier sur d'autres données
(dissimilitudes, mesures avec erreurs ...)



Bonne soirée ;)