Classification robuste et apprentissage faiblement supervisé

Apprendre à partir de données avec des labels incertains

Charles Bouveyron

SAMOS-MATISSE, CES, UMR CNRS 8174
Université Paris 1 Panthéon-Sorbonne
Paris, France

Travail commun avec Stéphane Girard
INRIA Rhône-Alpes, France

Plan de l'exposé

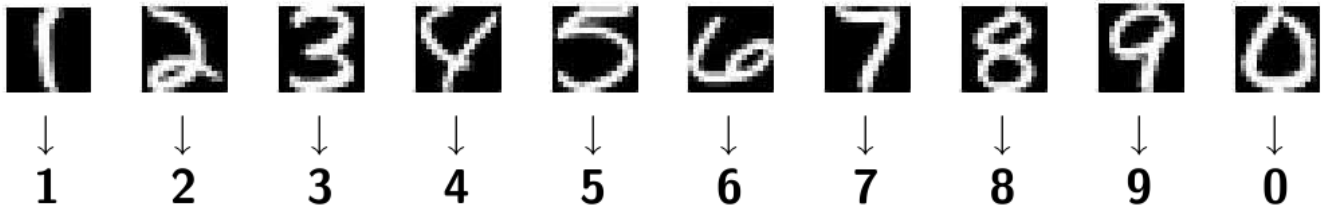
Plan de l'exposé

Introduction

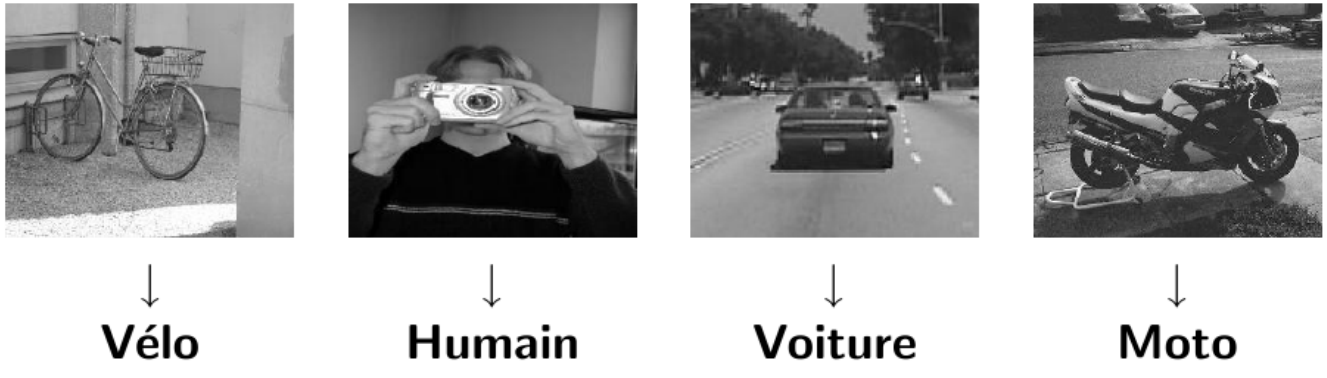
La classification est devenue un problème récurrent :

- qui intervient dans toutes les applications nécessitant une prise de décision,
- et l'approche probabiliste permet de quantifier le risque d'erreur de classement.

Exemple 1 : reconnaissance optique de caractères



Exemple 2 : reconnaissance d'objets à partir d'images



Introduction

En classification supervisée :

- la supervision humaine est requise pour associer des labels à un jeu de données dit d'apprentissage,
- ces données sont ensuite utilisées pour construire un classifieur supervisé.

Cependant, dans de nombreuses applications :

- la supervision humaine peut s'avérer imprécise ou difficile (données complexes, fatigue de l'expert, ...),
- et le coût de la supervision peut limiter le nombre d'observation étiquetées.

Par conséquent :

- quelques erreurs humaines dans la supervision peuvent avoir une grande importance dans la classification finale,
- ce, en particulier, si la taille du jeu d'apprentissage est limité.

Exemples introductifs

Détection automatique de cellules pathologiques (cancer) :

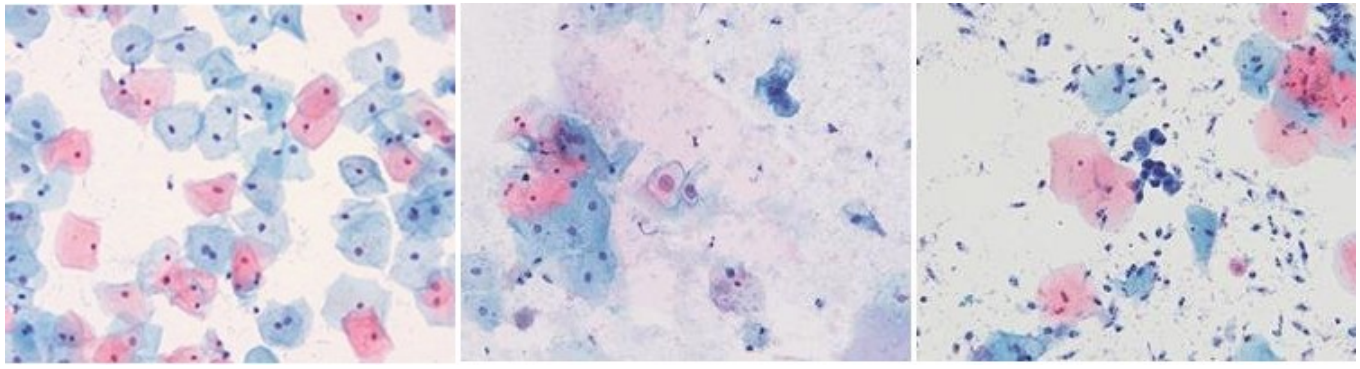


Fig. – Cytologie en couche mince

- proportion (très) faible de données de la classe «cancer» par rapport à la classe «sain»,
- détection difficile des cellules en mutation (sain $\longrightarrow$ cancer),
- tâche de supervision très fatiguante.

Exemples introductifs

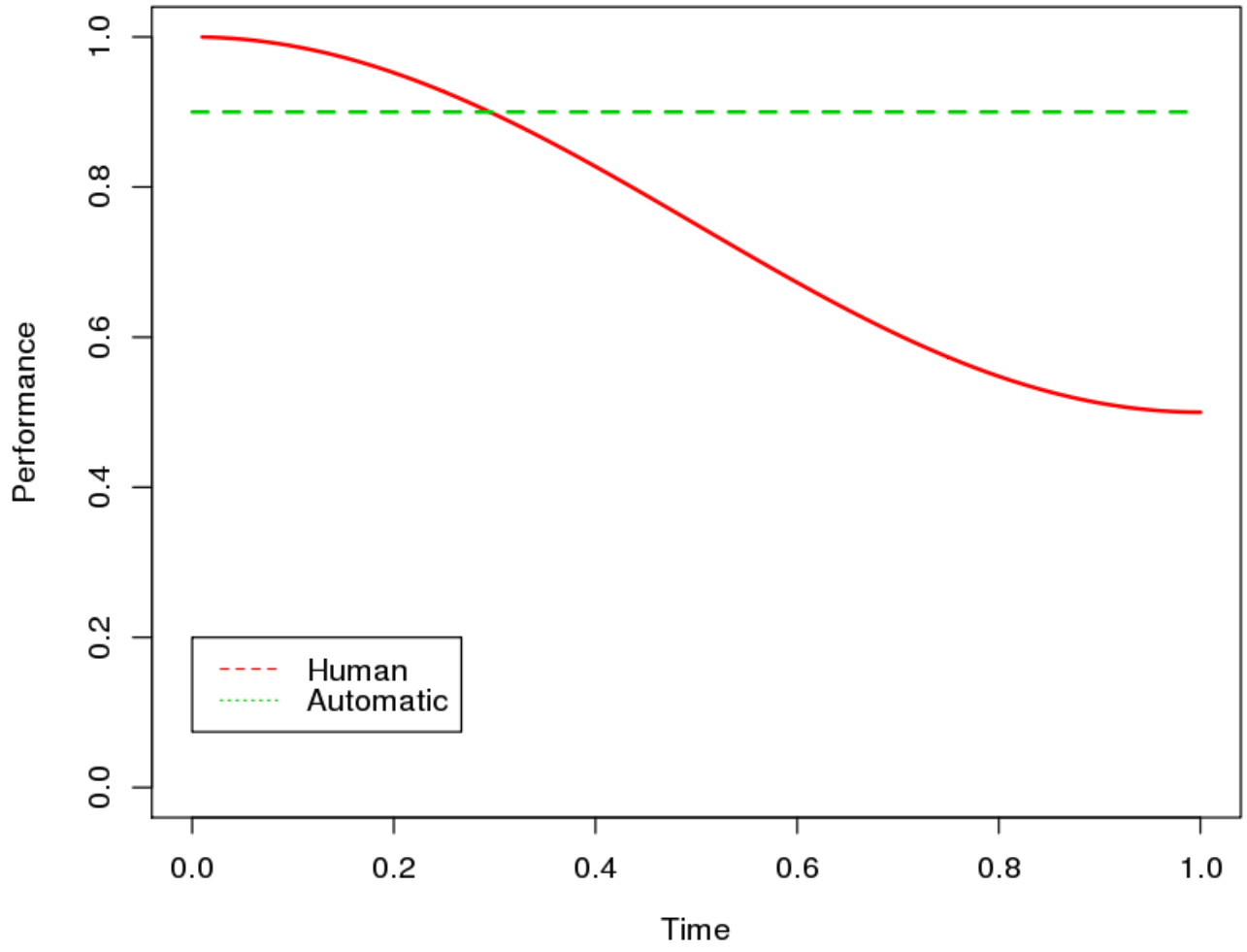


Fig. – Comparaison des performances humaine et de l'ordinateur

Exemples introductifs

Localisation d'objets dans des images :

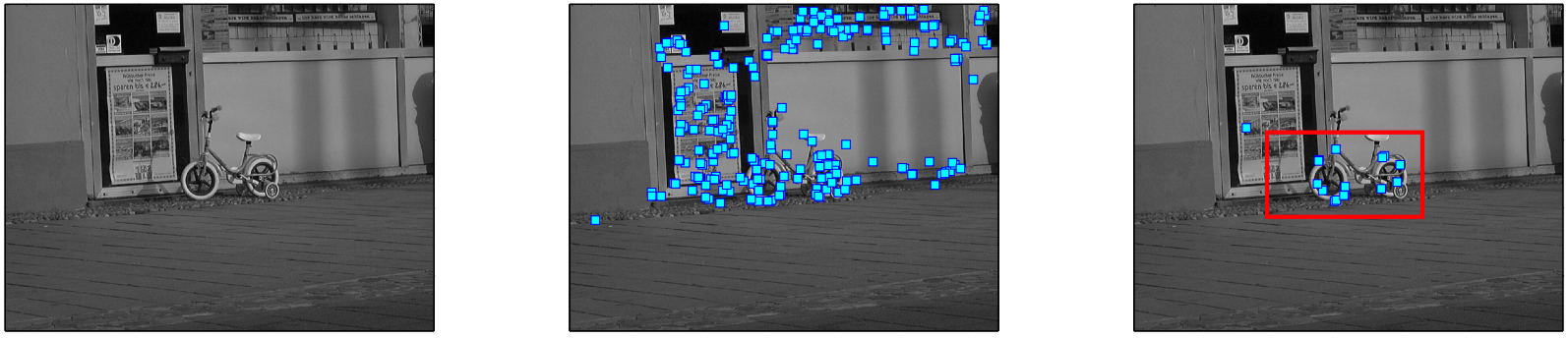


Fig. – Localisation de l'objet «vélo»

- domaine en forte croissante : reconnaissance de visages, détection de piétons, ...
- qui nécessite la supervision manuelle d'un grand nombre d'images,

Les problèmes :

- nombre potentiellement infini de catégories d'objets,
- segmentation des objets fastidieuse et imprécise.

Exemples introductifs

Reconnaissance / localisation d'objets dans des images :

Fig. – Supervision faible grâce à Google Image

Plan de l'exposé

Le problème de la classification

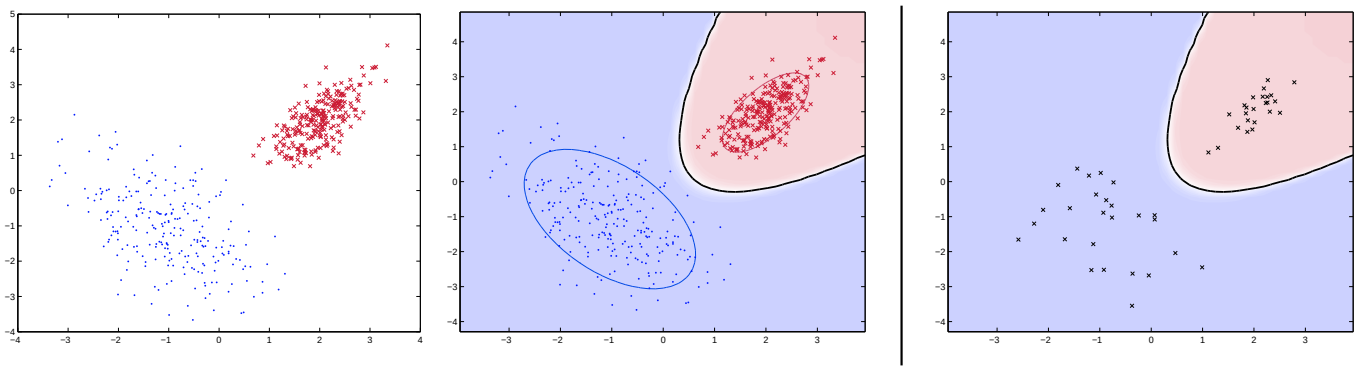
Le problème de la classification est :

- à partir d'un jeu d'apprentissage $\mathcal{A}$:

$$\mathcal{A} = \{(x_1, z_1), ..., (x_n, z_n)\} \in \mathbb{R}^p \times \{1, ..., K\}$$

- construire une règle de décision δ :

$$\begin{array}{rcl} \delta : \mathbb{R}^p & \rightarrow & \{1, ..., K\}. \\ x & \rightarrow & z. \end{array}$$



Apprentissage ***Classification***

La règle optimale δ^* est celle qui affecte x à la classe la plus probable a posteriori (règle du MAP) :

$$\delta^*(x) = \underset{i=1,...,K}{\operatorname{argmax}} P(Z = i | X = x).$$

La classification probabiliste

Le modèle de mélange :

- on suppose classiquement que les observations $x_1, ..., x_n$ sont des réalisations indépendantes,
- d'un vecteur aléatoire $X \in \mathbb{R}^p$ dont la densité s'écrit :

$$f(x) = \sum_{k=1}^{K} \pi_k f(x, \theta_k).$$

La règle du MAP s'écrit dans ce cas :

$$\begin{aligned}
\delta^*(x) &= \underset{k=1,...,K}{\operatorname{argmax}} P(Z = k | X = x) \\
&= \underset{k=1,...,K}{\operatorname{argmax}} P(Z = k) P(X = x | Z = k) \\
&= \underset{k=1,...,K}{\operatorname{argmax}} \pi_k f(x, \theta_k).
\end{aligned}$$

Remarque *: la construction de la règle de décision consiste à (i) estimer les paramètres θ_k et (ii) calculer la valeur de $H_k(x)$.*

Le problème du bruit de labels

En apprentissage statistique :

- il est très commun de supposer que les données sont bruitées,
- le problème du bruit sur les variables explicatives a été étudié largement dans la littérature,
- alors que le bruit sur la variable à prédire a reçue nettement moins d'attention (excepté en régression).

En classification supervisée :

- le bruit sur les labels est un problème important puisque toutes les méthodes supervisées font une confiance totale aux labels,
- et leur règle de décision sont par conséquent très sensibles à ce type de bruit :
  - les approches discriminatives de part la construction de la frontière,
  - les approches génératives de part l'estimation des paramètres.

Solutions existantes

Méthodes de «nettoyage des données» :

- ce sont les approches les plus anciennes : Gates (1972), Dasarathy (1980), ...
- l'idée est d'enlever les observations mal étiquetées en se basant sur des recherches d'*outliers*,
- cela induit cependant un biais dans la procédure d'apprentissage.

Estimation robuste des paramètres :

- dans le contexte des méthodes génératives, quelques travaux ont portés sur l'estimation robuste des paramètres du modèle,
- mais la réduction du taux de mauvaise classification s'est avérée faible.

Solutions existantes

Prise en compte de l'incertitude sur les labels :

- Côme *et al.* ont proposé récemment une approche basée sur le modèle de mélange pour classer des données avec labels imprécis,
- mais la méthode requiert de connaître l'incertitude de chacun des labels.

Modélisation du bruit de labels :

- Lawrence et Sholköpf ont proposé récemment un algorithme construisant un classifieur à noyau et incluant une modélisation explicite du bruit de labels,
- Li *et al.* ont étendu ce travail en permettant la modélisation de chaque classe par un mélange de gaussiennes,
- cependant, ces approches considèrent uniquement le cas de la classification binaire.

Plan de l'exposé

L'idée du modèle

L'idée de notre approche est :

- de comparer l'information supervisée portée par les données d'apprentissage,
- avec une modélisation non supervisée par modèle de mélange des données .

Avec une telle approche :

- la comparaison de l'information supervisée avec le modèle non supervisé devrait permettre de détecter les labels inconsistants,
- et il sera ainsi possible de construire un classifieur supervisé robuste en donnant une faible importance aux observations d'apprentissage ayant des labels inconsistants.

Un modèle de mélange pour la classification robuste

Nous considérons un modèle de mélange avec :

- une structure non supervisée de K groupes représentée par la variable aléatoire discrète S,
- et une structure supervisée de k classes representée par la variable aléatoire discrète C.

Avec les hypothèses et notations du modèle de mélange :

- les données $(x_1, ..., x_n)$ sont des réalisations indépendantes d'un vecteur aléatoire $X \in \mathbb{R}^p$ dont la densité est :

$$p(x) = \sum_{j=1}^{K} P(S = j) p(x|S = j), \tag{1}$$

- où $P(S = j)$ et $p(x|S = j)$ sont respectivement la probabilité *a priori* et la densité conditionnelle du $j$ème groupe.

Un modèle de mélange pour la classification robuste

Nous introduisons à présent l'information supervisée :

- puisque $\sum_{i=1}^{k} P(C = i|S = j) = 1$ pour tout $j = 1, ..., K$, nous pouvons introduire cette quantité dans (1) pour obtenir :

$$p(x) = \sum_{i=1}^{k} \sum_{j=1}^{K} P(C = i|S = j)P(S = j)p(x|S = j), \qquad (2)$$

- où $P(C = i|S = j)$ peut être interprété comme la probabilité que le $j$ème groupe appartienne à la $i$ème classe.

En utilisant les notations classiques :

- l'équation (2) peut être reformulée :

$$p(x) = \sum_{i=1}^{k} \sum_{j=1}^{K} r_{ij}\pi_j f(x, \theta_j), \qquad (3)$$

- où $r_{ij} = P(C = i|S = j)$, $\pi_j = P(S = j)$ et f est la densité conditionnelle du $j$ème groupe de paramètre θ_j.

Choix du modèle de mélange

Ce choix dépend de la nature des données :

- données quantitatives :
  - modèle gaussien classique,
  - modèle gaussien parcimonieux,
  - modèle gaussien pour les données de grande dimension.
- données qualitatives :
  - modèle multinomial
  - les données ne doivent pas être de trop grande dimension.

Le choix du modèle :

- est laissé à la discrétion du praticien,
- qui pourra s'aider d'outils comme le critère BIC si besoin.

Règle de classification

De façon classique, nous utilisons la règle du MAP :

- qui affecte une nouvelle observation x à la classe la plus probable *a posteriori*,
- ainsi, l'étape de classification consiste principalement à calculer $P(C = i|X = x)$ pour chaque classe $i = 1, ..., k$.

Dans le cas du modèle présenté :

- la probabilité *a posteriori* $P(C = i|X = x)$ vaut :

$$P(C = i|X = x) = \sum_{j=1}^{K} r_{ij} P(S = j|X = x),$$

- nous avons donc pour ce faire à estimer les paramètres r_{ij} et les probabilités *a posteriori* $P(S = j|X = x)$.

Liens avec Mixture Discriminant Analysis

Mixture Discriminant Analysis:

- chaque classe est modélisée par un mélange de K_i gaussiennes,
- MDA suppose que la densité conditionnelle de chaque classe est :

$$p(x|C=i) = \sum_{j=1}^{K} \pi_{ij} \phi(x; \mu_j, \Sigma_j),$$

Par conséquent :

- nous pouvons réécrire la densité $p(x)$:

$$p(x) = \sum_{i=1}^{k} \sum_{j=1}^{K} r_{ij} \pi_j \phi(x; \mu_j, \Sigma_j),$$

- où $r_{ij} = P(C=i|S=j)$ est connu et vaut $r_{ij} = 1$ si le $j$ème groupe appartient à la $i$ème classe et $r_{ij} = 0$ sinon.

Plan de l'exposé

Estimation des paramètres du mélange

Du fait de la nature du modèle présenté :

- la procédure d'estimation est faite de deux étapes,
- correspondant respectivement aux parties non supervisée et supervisée de la comparaison.

Estimation des paramètres du mélange :

- dans cette première étape, les labels ne sont pas utilisés pour former K groupes homogènes,
- l'algorithme EM est utilisé pour estimer les paramètres du mélange par maximisation de la vraisemblance,
- les formules de mise à jour de l'algorithme EM dépendent du modèle choisi (gaussien, gaussien HD, multinomial, ...).

Estimation des paramètres r_{ij}

Estimation des paramètres r_{ij} par MV :

- la log-vraisemblance du modèle proposé est :

$$\ell(R) = \sum_{i=1}^{k} \sum_{x \in \mathcal{C}_i} \log \left(\sum_{j=1}^{K} r_{ij} P(S = j | X = x) \right) + C^{ste}.$$

- ce qui conduit au problème d'optimisation sous contraintes :

$$\begin{cases} \text{maximiser :} & \sum_{i=1}^{k} \sum_{x \in \mathcal{C}_i} \log \left(R_i \Psi(x) \right), \\ \text{sous la contrainte :} & r_{ij} \in [0,1], \forall i = 1, \ldots, k, \forall j = 1, \ldots, K, \\ \text{et la contrainte :} & \sum_{i=1}^{k} r_{ij} = 1, \ \forall j = 1, \ldots, K, \end{cases}$$

où $\Psi(x) = (P(S = 1|X = x), \ldots, P(S = K|X = x))^t$ et R_i est la $i$ème ligne de $R = (r_{ij})$.

Choix du nombre de groupes

Choix de K :

- nous pouvons utiliser les outils classiques du modèle de mélange,
- en particulier le critère BIC est tout à fait adapté à la situation.

Remarque :

- le choix du nombre de composantes par classe dans RMDA est beaucoup plus simple que le choix des K_i dans MDA,
- pour RMDA, il suffit de choisir K et le nombre de composantes par classe s'adaptent automatiquement de part l'estimation des paramètres r_{ij}.

Plan de l'exposé

Protocole expérimental

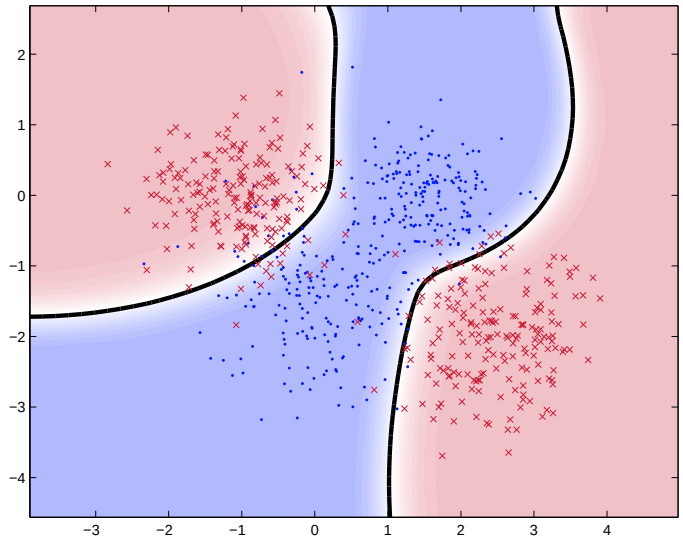
Simulation du bruit de labels :

- nous considérons le cas de l'échange de labels,
- les labels des observations sont échangés selon une loi de Bernoulli de paramètre $\eta \in [0, 1]$,
- η représente donc le taux de contamination.

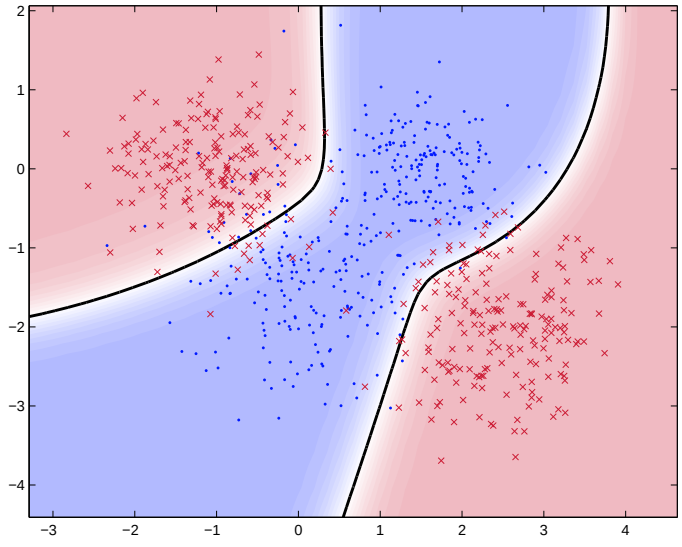
Méthodes comparées :

- Linear Discriminant Analysis (LDA),
- Mixture Discriminant Analysis (MDA),
- Robust Linear Discriminant Analysis (RLDA),
- Robust Mixture Discriminant Analysis (RMDA),

Classification binaire (données simulées)



MDA avec $\eta = 0$

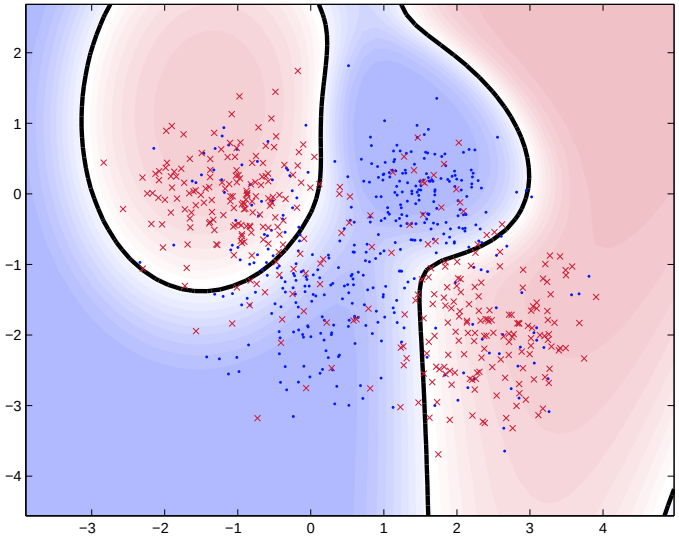


RMDA avec $\eta = 0$

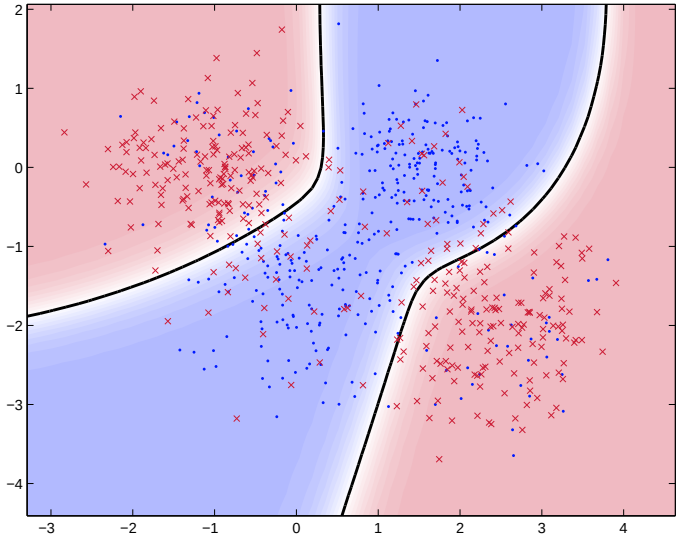
Données simulées :

- deux classes formées par un mélange de 2 gaussiennes,
- dans un espace de dimension $p = 2$.

Classification binaire (données simulées)

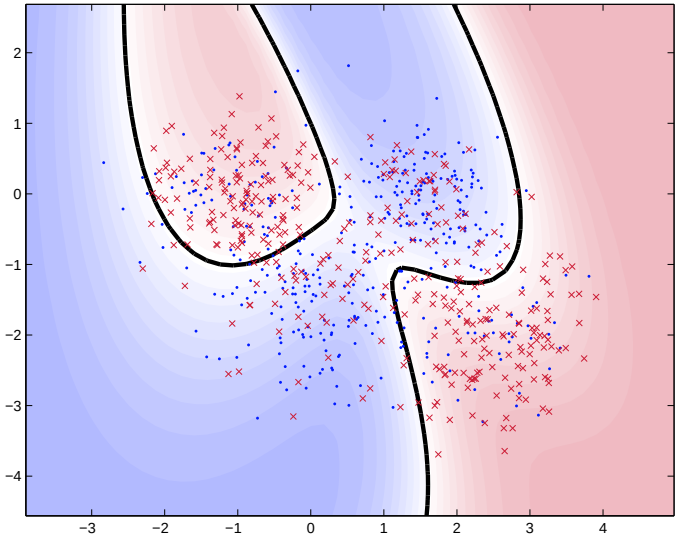


MDA avec $\eta = 0.15$

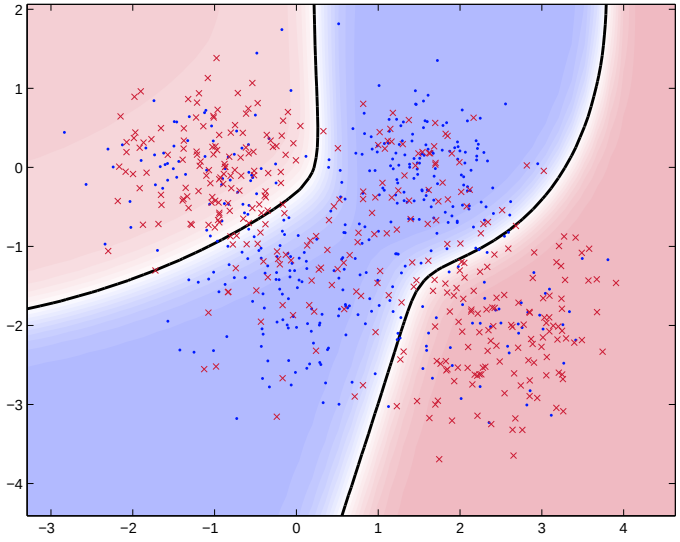


RMDA avec $\eta = 0.15$

Classification binaire (données simulées)

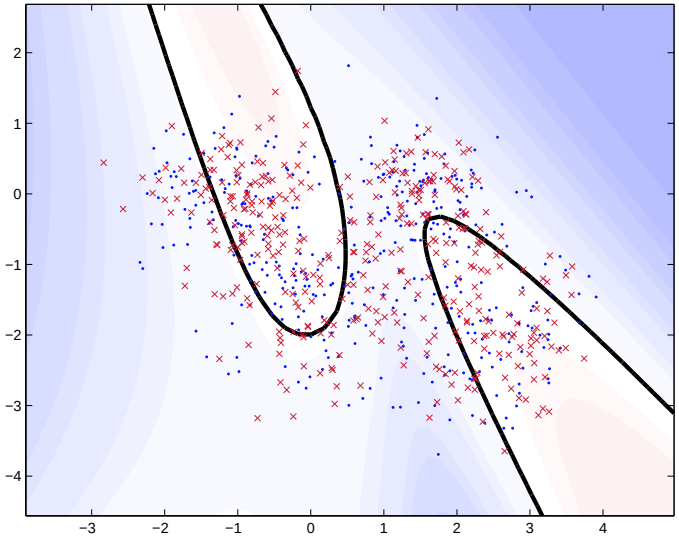


MDA avec $\eta = 0.30$

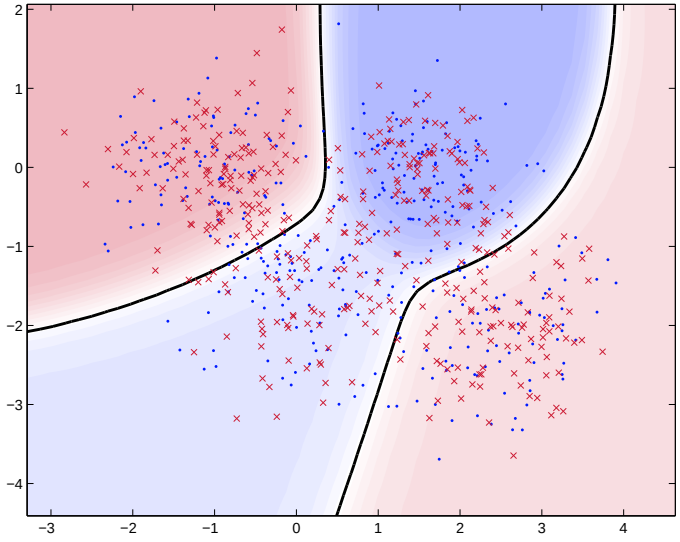


RMDA avec $\eta = 0.30$

Classification binaire (données simulées)

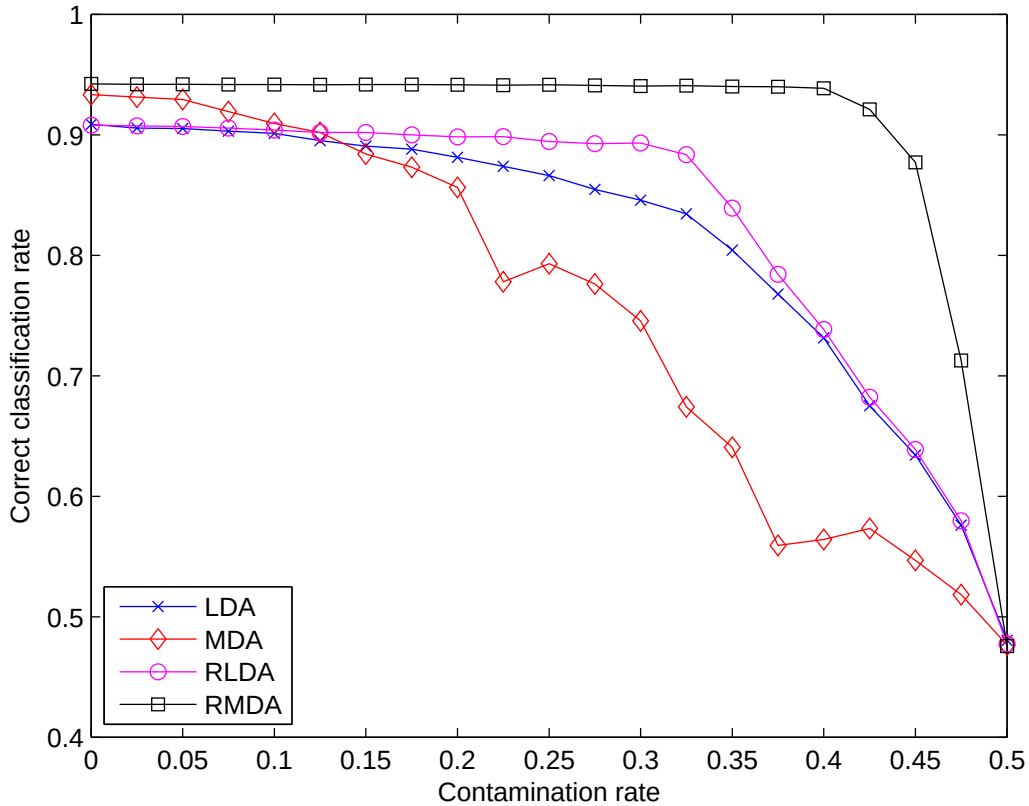


MDA avec $\eta = 0.45$



RMDA avec $\eta = 0.45$

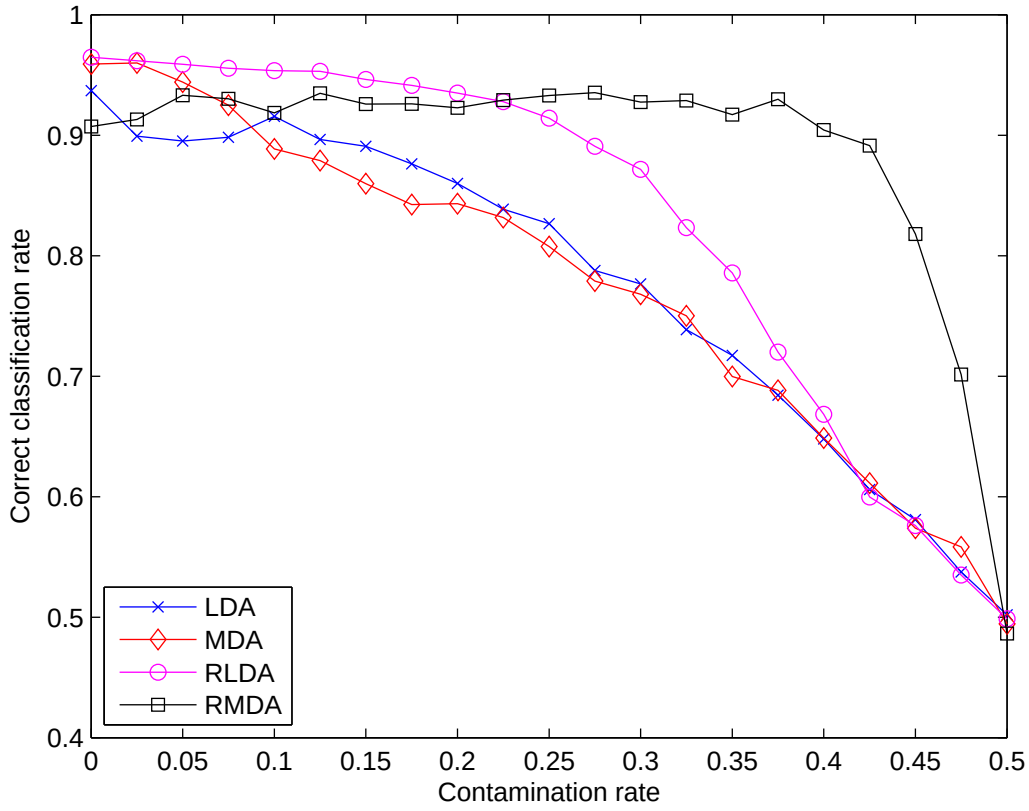
Classification binaire (données simulées)



Données simulées :

- deux classes formées par un mélange de 2 gaussiennes,
- dans un espace de dimension $p = 50$,
- 750 obs. pour l'apprentissage, le bruit de label $\eta = 0, ..., 0.5$,
- l'expérience a été répétée 25 fois pour moyenner les résultats.

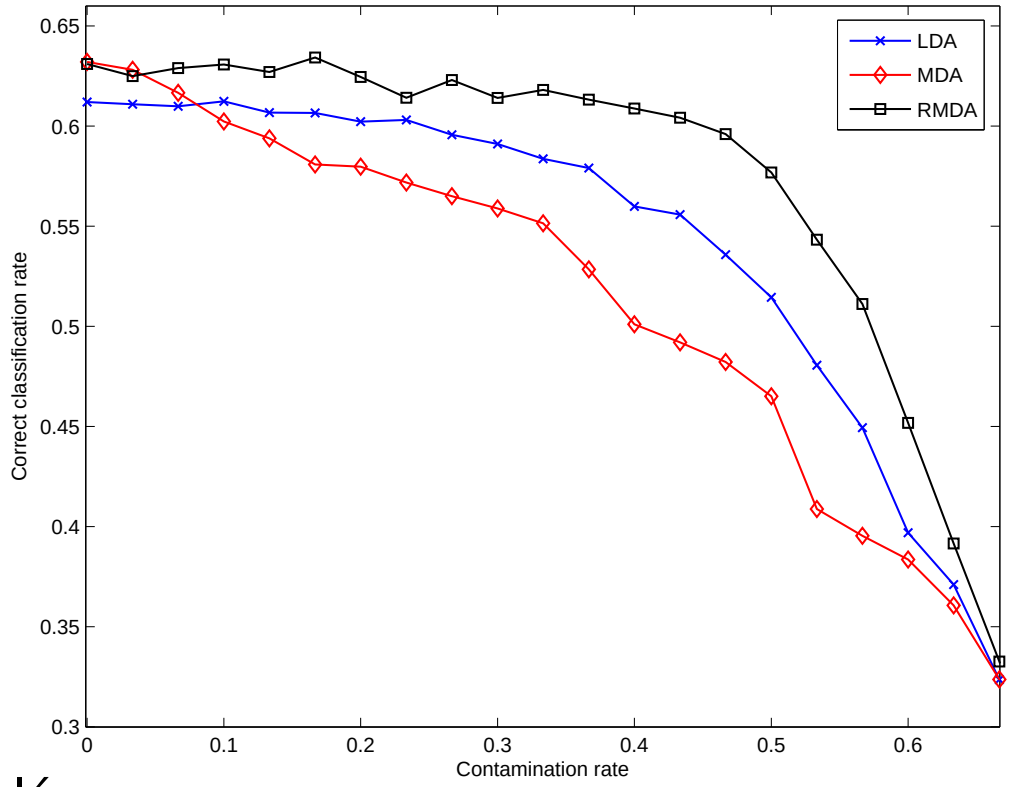
Classification binaire (données réelles)



Données réelles :

- données de reconnaissance optique de caractères (USPS),
- 2 classes (caractères '2' et '4') en dimension 256,
- 7250 obs. pour l'apprentissage et 25 rép. de l'expérience.

Classification multi-classes (données simulées)



Données simulées :

- 3 classes formées par un mélange de 2 gaussiennes
- dans un espace de dimension 50,
- 750 obs. pour l'apprentissage, le bruit de label $\eta = 0, ..., 2/3$,
- l'expérience a été répétée 25 fois pour moyenner les résultats.

Plan de l'exposé

Classification faiblement supervisée

La classification faiblement supervisée :

- nouveau type de classification supervisée ($\neq$semi-supervisée),
- qui utilise comme supervision des informations disponibles facilement et à faible coût,
- exemples :
  - informations de contexte $\longrightarrow$ images, documents,
  - segmentation rapide $\longrightarrow$ images.

L'idée de l'extension :

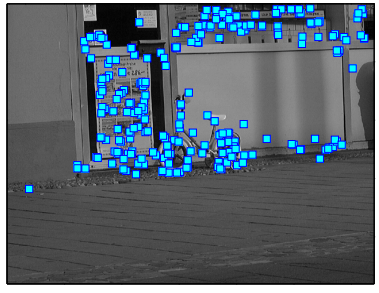
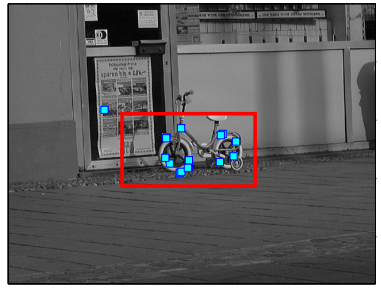
- utiliser des sources d'information pour associer un label à des groupes d'observations,
- en ayant conscience que l'on commet des erreurs de labels,
- et utiliser RMDA pour discriminer les classes présentes dans les données.

Localisation d'objets dans des images

La localisation d'objets dans des images naturelles :

- est un des problèmes les plus difficiles en vision par ordinateur,
- du fait du très grand nombre de catégories d'objets,
- et de la grande variété des objets considérés.

L'approche classique :

- chaque image est représentée par des points d'intérêt,
- qui sont décrits grâce à des descripteurs locaux de grande dimension ($p = 128$),
- puis ces descripteurs sont affectés à la classe «objet» ou à la classe «fond».

Localisation faiblement supervisée

Localisation supervisée :

- segmentation manuelle des objets dans les images d'apprentissage,
- descripteurs locaux x_i :
  - x_i sur l'objet : $z_i = 1$,
  - x_i sur le fond : $z_i = 0$,

Localisation faiblement supervisée :

- on utilise un moteur de recherche pour trouver des images contenant l'objet étudié,
- descripteurs locaux x_i :
  - $x_i \in$ à une image contenant l'objet : $z_i = 1$,
  - $x_i \in$ à une image ne contenant pas l'objet : $z_i = 0$,
- les labels 0 sont tous justes alors qu'une partie des labels 1 sont faux.

Protocole expérimental

Les données :

- nous avons choisi d'utiliser le *Pascal dataset* (*www.pascal-network.org*),
- qui a servi à une compétition de localisation d'objets,
- comparaison des meilleures méthodes actuelles.

Le *Pascal dataset* :

- 684 images pour l'apprentissage,
- 2 jeux de test : 689 et 956 images resp. pour *test1* et *test2*,
- quatre catégories d'objets : moto, vélo, voiture et humains.

Évaluation des résultats :

- nous avons utilisé la mesure AP$\in [0, 1]$ proposée lors de la compétition,
- qui mesure l'adéquation entre la localisation proposée et réelle de l'objet.

Protocole expérimental

Résultats

Database	Pascal test 1		Pascal test 2	
Supervision	full	weak	full	weak
RMDA $[a_{ij}b_iQ_id_i]$	0.302	0.273	0.172	0.145
RMDA $[a_{ij}bQ_id_i]$	**0.318**	**0.287**	0.181	0.147
RMDA $[a_ib_iQ_id_i]$	0.313	0.285	**0.183**	0.142
RMDA $[a_ibQ_id_i]$	**0.318**	0.283	0.176	**0.148**
RMDA $[a_ib_iQ_id]$	0.314	**0.287**	0.179	0.130
RMDA spherical	0.271	0.216	0.149	0.106
RMDA diagonal	0.276	0.227	0.161	0.110
RMDA common	0.267	0.246	0.164	0.116
Best method of [9]	0.279	/	0.112	/

Résultats

Fig. Localisation d'une instance d'un objet.

Fig. Localisation de plusieurs instances d'un objet..

Plan de l'exposé

Conclusion

Nous avons donc proposé un classifieur supervisé robuste :

- qui prend en compte l'incertitude sur les labels,
- en comparant l'information portée par les labels à une modélisation non supervisée des données,
- qui s'adapte aux différents types de données par le choix du modèle de mélange.

Extension à la classification faiblement supervisée :

- nous avons montré qu'il est possible d'étendre la classification à partir de labels incertains,
- à une classification supervisée par des informations disponibles facilement,
- en acceptant qu'une partie des labels soient erronés.

Perspectives

D'un point de vue théorique :

- estimation du pourcentage d'incertitude par classe,
- incorporation de l'incertitude sur les labels donnée par l'expert,
- travail sur le cas des données qualitatives.

Du point de vue des applications :

- application à la cytologie et à la spectrométrie de masse,
- application en Biologie au DNA barcoding.