Analyse Discriminante Imprécise basée sur l'inférence Bayésienne robuste

Yonatan-Carlos Carranza-Alarcon[1] Sébastien Destercke[1]
[1] UMR CNRS 7253 Heudiasyc, Sorbonne Université
{yonatan-carlos.carranza-alarcon, sebastien.destercke}@hds.utc.fr

Résumé :
L'objectif de cet article est de proposer une nouvelle approche de classification prudente basée sur l'inférence Bayésienne robuste et l'analyse discriminante linéaire. Cette modélisation est conçue pour prendre en compte, dans ses inférences a posteriori, le manque d'information lié aux données. Le principe de cette approche est d'utiliser un ensemble de distributions a priori pour modéliser l'ignorance initiale, plutôt qu'une seule distribution (souvent dite "non-informative") qui peut fortement influencer les résultats en cas de faible quantité de données. Des premières expériences montrent que l'ajout d'imprécision permet d'être prudent en cas de doute sans pour autant diminuer la qualité du modèle, tout en gardant un temps de calcul raisonnable.

Mots-clés :
Analyse Discriminante, Bayésien robuste, Classification

Abstract:
The aim of this paper is to propose a new classification approach based on robust Bayesian analysis and Linear discriminant analysis. This modeling is adapted to account for the lack of prior information (due to lack of data, for example) in its posterior inferences. The principle of this approach is to use a set of prior distributions to model the initial near-ignorance, rather than a single distribution (often called "non-informative") that can strongly influence the results in case of small datasets. First experiments show that the insertion of imprecision allows it to be cautious in cases of doubt without reducing the quality of the model, while keeping a reasonable computing time.

Keywords:
Discriminant Analysis, Robust Bayesian, Classification.

1 Introduction

La classification supervisée classique est un sujet d'étude très actif et assez approfondi offrant des résultats très satisfaisants. Ceci dit, cette dernière réalise toujours une prédiction *"précise"* (ou exacte), quelle que soit la situation d'ignorance ou de certitude du classifieur. On entend par *"ignorance"* le manque de connaissance du classifieur pour distinguer clairement la différence entre deux classes. Or, il existe à présent des classifieurs nommés *"prudents"* afin de mieux gérer cette ignorance.

La classification supervisée fondée sur modèles prudents est devenue un champ de recherche assez actif ces dernières années (e.g. [10] et [11]). En modélisant explicitement le manque de connaissances, ces modèles permettent d'avoir des classifications prudentes, par exemple quand les données sont bruitées ou en petit nombre. Dans cet article, nous proposons une nouvelle méthode de classification prudente basée sur l'analyse discriminante classique dans un contexte Bayésien ; celle-ci se concentre sur l'estimation imprécise de la moyenne de la distribution Gaussienne.

Nous utiliserons le cadre des probabilités imprécises [1] consistant à mesurer l'incertitude attachée aux données en estimant un intervalle $[\underline{P}, \overline{P}]$ de probabilités *"plausibles"* d'un événement, contrairement au cadre classique des probabilités précises. Cette estimation prudente essaie d'appréhender la connaissance manquante, afin de faire des prédictions en conséquence.

Le reste de cet article est organisé comme suit : la section 2 décrit le cadre général de l'inférence et de la classification précise, la section 3 aborde l'inférence imprécise et la nouvelle méthode de classification prudente de l'analyse discriminante, la section 4 présentera des expériences. Dans cet article, nous utiliserons les notations mathématiques du tableau 1.

2 Classification

Étant donné un jeu de données d'apprentissage $\mathcal{D} = \{(\boldsymbol{x}_i, y_i) | i = 1, ..., N\}$ dont chaque ob-

Notation	Valeur	Description
$(\cdot)^T$	-	L'application transposée
$(\boldsymbol{x}_{i,k}, y_{i,k})$	$\{(\boldsymbol{x}_{1,k}, y_{1,k}), \ldots, (\boldsymbol{x}_{n_k,k}, y_{n_k,k})\}$	Des observations de la catégorie k
N	-	Nombre total d'individus
n_k	-	Nombre d'individus de la catégorie k
$\overline{\boldsymbol{x}}_k$	$\frac{1}{n_k}\sum_{i=1}^{n_k} \boldsymbol{x}_{i,k}$	Moyenne empirique de la catégorie k
S_k	$\frac{1}{N-n_k}\sum_{i=1}^{n_k}(\boldsymbol{x}_{i,k} - \overline{\boldsymbol{x}}_k)(\boldsymbol{x}_{i,k} - \overline{\boldsymbol{x}}_k)^T$	Matrice de variance-covariance empirique de la catégorie k
S	$\frac{1}{(N-K)}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(\boldsymbol{x}_{i,k} - \overline{\boldsymbol{x}}_k)(\boldsymbol{x}_{i,k} - \overline{\boldsymbol{x}}_k)^T$	Matrice de variance-covariance empirique globale

Tableau 1 – Notations mathématiques.

servation est composée par un couple $(\boldsymbol{x}_i, y_i)$ tel que $\boldsymbol{x}_i \in \mathbb{R}^d$ est un individu avec d régresseurs (ou variables explicatives) et $y_i \in \mathscr{Y}$ est la catégorie associée, avec $\mathscr{Y} = \{m_1, ..., m_K\}$ l'ensemble des catégories.

L'objectif de la classification est donc d'attribuer à un nouvel individu la catégorie la plus plausible parmi les K catégories existantes, autrement dit, celle qui minimise l'erreur moyenne de prédiction [9, pp. 21], ou maximise la probabilité d'une catégorie soumise à la fonction de coût classique zéro-un :

$$\hat{y}^* = \underset{m_k \in \mathscr{Y}}{\arg\max}\, P(Y = m_k | X = \boldsymbol{x}^*). \tag{1}$$

Une autre façon de voir cette prise de décision est en établissant une relation de préférence binaire $\succ$ sur les catégories afin de trouver la catégorie la plus plausible :

$$m_a \succ m_b \iff \frac{\mathscr{D}_a(\boldsymbol{x}^*)}{\mathscr{D}_b(\boldsymbol{x}^*)} > 1 \tag{2}$$

avec $\mathscr{D}_k(\boldsymbol{x}^*) = P(Y = m_k | X = \boldsymbol{x}^*)$ (connu aussi sous le nom de ***"score discriminant"***), et où $\mathscr{D}_b(\boldsymbol{x}^*)$ est positif (tjs. le cas ici). Cette équation (2) exprime le fait que m_a est préférable à m_b quand la probabilité de m_a est supérieure à m_b. Ainsi donc, nous pouvons établir un ordre total sur les catégories $m_{i_K} \succ ... \succ m_{i_1}$ et prédire la catégorie m_{i_K} correspondant à l'élément maximal de $\succ$. Ce concept sera étendu dans le cadre des probabilités imprécises dans la section 3.

2.1 Analyse Discriminante

L'analyse discriminante se propose d'estimer la probabilité de la catégorie $Y = m_k$ sachant $\boldsymbol{x}^*$ en passant par le théorème de Bayes, c'est-à-dire en la décomposant de la manière suivante :

$$\mathscr{D}_k(\boldsymbol{x}^*) = \frac{P(X = \boldsymbol{x}^* | Y = m_k) P(Y = m_k)}{\sum_{m_l \in \mathscr{Y}} P(X = x | Y = m_l) P(Y = m_l)} \tag{3}$$

$\mathbb{P}_{X|Y=m_k}$ désigne la distribution conditionnelle d'un groupe d'individus appartenant à la catégorie y_k et $\mathbb{P}_{Y=m_k}$ représente la probabilité marginale de la catégorie m_k.

L'analyse discriminante s'appuie ensuite sur une hypothèse de normalité, qui suppose que chaque distribution conditionnelle $\mathbb{P}_{X|Y=m_k}$ pour chaque catégorie $m_k \in \mathscr{Y}$ suit une loi Gaussienne de moyenne μ_k et de matrice de variance-covariance Σ_k, autrement dit, le modèle paramétrique est représenté par :

$$\mathscr{G}_k := \mathbb{P}_{X|Y=m_k} \sim \mathcal{N}(\mu_k, \Sigma_k). \tag{4}$$

Le modèle de l'analyse discriminante linéaire (ADL) suppose que les matrices de variance-covariance Σ_k sont identiques d'un groupe d'individus à l'autre (i.e. l'hypothèse d'homoscédasticité, $\Sigma_k = \Sigma, \forall m_k \in \mathscr{Y}$) et la distribution marginale $\pi_k := \mathbb{P}_{Y=m_k}$ vérifie $\sum_k \pi_k = 1$. L'analyse discriminante quadratique (ADQ) relâche cette hypothèse d'homoscédasticité.

Dans la plupart des cas pratiques et d'un point de vue fréquentiste, l'estimation *précise* des pa-

ramètres inconnus de la loi gaussienne $\mathscr{G}_k$ à partir de nos données d'apprentissage $\mathcal{D}$ est fixé comme suit [9] : $\hat{\pi}_k = n_k/N$, $\hat{\mu}_k = \overline{\boldsymbol{x}}_k$ et $\hat{\Sigma}_k = S_k$.

3 Classification prudente

Le processus d'apprentissage dans des modèles prudents est assez semblable aux processus classiques (e.g. la classification supervisée). Dans cette section, nous présentons d'abord l'inférence (ou prise de décision), puis l'estimation de paramètres inconnus de notre modèle.

Rappelons que le cadre des probabilités imprécises consiste à remplacer des estimations précises par des estimations imprécises $\mathscr{P}$ (i.e. l'ensemble de distributions), le plus souvent sous la forme d'ensembles convexes.

3.1 Décision imprécise

Dans le contexte des probabilités imprécises, nous pouvons trouver différentes méthodes étendant le critère de décision donné par l'équation (2), (voir [14]) .

Ici, pour classifier un nouvel individu $\boldsymbol{x}^*$, nous allons faire usage du *critère de maximalité* [1, §8.6] ayant de fortes justifications théoriques [15, §3.9.5] et pratiques [17, 10, 16, 2]. Celui-ci étend l'équation (2) dès lors que nous sommes intéressés par des prédictions imprécises, mais fiables, sous la forme d'un sous-ensemble de catégories (mutuellement exclusives entre-elles).

Eu égard à ce qui précède, le critère de maximalité peut être défini comme suit :

Définition 3.1 (Maximalité) *Étant donné un ordre partiel sur les catégories $\succ_M$ et un ensemble de probabilités $\mathcal{P}$ donné :*

$$m_a \succ_M m_b \iff \inf_{P \in \mathcal{P}} \frac{\mathscr{D}_a(\boldsymbol{x}^*)}{\mathscr{D}_b(\boldsymbol{x}^*)} > 1 \qquad (5)$$

avec $\mathscr{D}_k(\boldsymbol{x}^*) = P(Y = m_k | X = \boldsymbol{x}^*)$.

L'équation (5) revient à demander que (2) soit vrai pour toutes les probabilités possibles. En pratique, $\succ_M$ peut être un ordre partiel avec plusieurs éléments maximaux, auquel cas la prédiction devient imprécise. Néanmoins en cas d'incertitude nulle (e.g. $N \to \infty$), ces modèles prudents peuvent aussi prédire une seule classe. L'ensemble de décisions prudentes en utilisant $\succ_M$ est donc :

$$Y_M = \left\{ m_f \in \mathscr{Y} \;\middle|\; \nexists m_g : m_g \succ_M m_f \right\} \qquad (6)$$

À présent, en reprenant les équations (5) et (3), nous obtenons $m_a \succ_M m_b$ si et seulement si :

$$\inf_{\substack{P_{X|m_a} \in \mathscr{G}_a P_{X|m_b} \in \mathscr{G}_b \\ P_{m_a}, P_{m_b} \in \mathbb{P}_Y}} \frac{P(\boldsymbol{x}^* | y = m_a) P(y = m_a)}{P(\boldsymbol{x}^* | y = m_b) P(y = m_b)} > 1 \qquad (7)$$

Étant donné que les distributions $P(\boldsymbol{x}^* | y = m_a) \in \mathscr{G}_a$ et $P(\boldsymbol{x}^* | y = m_b) \in \mathscr{G}_b$ sont indépendantes, et en supposant que la loi marginale $\mathbb{P}_Y$ soit précise, nous pouvons donc appliquer l'infimum au numérateur et le supremum au dénominateur de l'équation (7). Notre problème se réduit donc à résoudre les deux équations suivantes :

$$\overline{P}(\boldsymbol{x}^* | y = m_b) = \sup_{P \in \mathscr{G}_b} P(\boldsymbol{x}^* | y = m_b) \qquad (8)$$

$$\underline{P}(\boldsymbol{x}^* | y = m_a) = \inf_{P \in \mathscr{G}_a} P(\boldsymbol{x}^* | y = m_a) \qquad (9)$$

Et, en reprenant l'hypothèse de normalité (4) de l'analyse discriminante précise, nous notons :

$$(\overline{\mu}_b, \overline{\Sigma}_b) = \arg\inf_{(\mu,\Sigma)_{\mathscr{G}_b}} \frac{1}{2} (\boldsymbol{x}^* - \mu_b)^T \Sigma_b^{-1} (\boldsymbol{x}^* - \mu_b) \qquad (10)$$

$$(\underline{\mu}_a, \underline{\Sigma}_a) = \arg\sup_{(\mu,\Sigma)_{\mathscr{G}_a}} \frac{1}{2} (\boldsymbol{x}^* - \mu_a)^T \Sigma_a^{-1} (\boldsymbol{x}^* - \mu_a) \qquad (11)$$

L'estimation $(\overline{\mu}_b, \overline{\Sigma}_b)$ (resp. $(\overline{\mu}_a, \overline{\Sigma}_a)$) s'obtient quand l'équation (10) (resp. (11)) atteint la borne inférieure (resp. borne supérieure). Ainsi, le problème principal est à présent réduit à connaître les valeurs des paramètres inconnus $\hat{\mu}_*$ et $\hat{\Sigma}_*$ atteignant ces bornes. Il nous reste à définir comment peuvent être obtenues les familles de gaussiennes, i.e. comment les ensembles de paramètres $\hat{\mu}_*$ et $\hat{\Sigma}_*$ possibles peuvent être estimés.

Dans le contexte des probabilités imprécises, ces paramètres peuvent être estimés en utilisant ; (1) des méthodes d'inférence fréquentiste

imprécise [6] (robustesse fréquentiste), ou sinon, (2) des méthodes d'inférence Bayésienne généralisée (robustesse Bayésienne).

Dans la suite, cette étude se focalise sur l'usage de la seconde méthode d'inférence. Ainsi, l'inférence Bayésienne demande en principe une connaissance a priori sur les paramètres inconnus, celle-ci étant dans la plupart des cas absente ou difficile à obtenir pendant l'étude. En l'absence de cette dernière, il est naturel de choisir un ensemble d'a priori *« vide d'information »* aussi grande que possible, i.e. représentant la méconnaissance ou l'ignorance des paramètres inconnus (aussi connue sous le terme de l'*ignorance préalable* ou en anglais *prior near-ignorance*)[15, §4.6.9].

Selon la littérature sur l'inférence bayésienne, l'ignorance préalable peut être traitée en prenant une loi a priori non-informative pour chaque paramètre, cependant cette dernière ne modèlise pas l'absence de connaissance, mais seulement l'invariance par translation [15], et elles sont aussi souvent impropres. Une autre alternative plus prudente proposée notamment par [3] dans le contexte des probabilités imprécises, est de considérer un ensemble de distributions a priori pour chaque paramètre inconnu au lieu d'une seule distribution.

Dans la suite, nous nous concentrons sur l'estimation de μ_k et Σ_k, en rendant seulement les moyennes μ_k imprécises. Dans le but d'être concis, nous supprimerons l'indice k de μ et Σ, en gardant toujours à l'esprit que ces estimations sont liées à une catégorie k.

3.2 Analyse discriminante linéaire imprécise

Dans cette sous-section, nous reprenons l'hypothèse d'homoscédasticité (i.e. $\Sigma_k = \Sigma, \forall m_k \in \mathscr{Y}$), et nous supposons une estimation précise pour la matrice variance-covariance, i.e. $\hat{\Sigma} = S$.

La modélisation Bayésienne s'appuie principalement sur deux composantes ; la vraisemblance et la loi a priori, pour ensuite construire des procédures d'inférence a posteriori sur le paramètre inconnu, ici μ. La première est le produit de probabilités conditionnelles $\prod_i \mathbb{P}_{X_i|Y_i,\mu}$ et le second est la loi a priori $\mathbb{P}_\mu$.

Afin de simplifier la procédure de modélisation Bayésienne, nous ferons l'usage de l'outil des lois conjuguées des familles exponentielles régulières (i.e. $\mathscr{FE}\mathscr{xp}$) définies dans [13, §3.3.4] et [4, §5.2], pour laquelle la distribution a priori et a posteriori appartient à une même famille de lois.

Nous tenons donc la vraisemblance de lois Gaussiennes qui fait partie de la $\mathscr{FE}\mathscr{xp}$, et en nous appuyant sur [3, §th. 4.6], nous proposons de considérer l'ensemble de lois a priori non-informatives [1], aussi appartenant à la $\mathscr{FE}\mathscr{xp}$, suivant [2] :

$$\mathcal{M}_0^\mu = \left\{ \mu \in \mathbb{R}^d \,\middle|\, \begin{array}{l} p(\mu|\ell) \propto \exp(\ell^T \mu), \\ \ell = [\ell_1, ..., \ell_d]^T \in \mathbb{L} \end{array} \right\} \quad (12)$$

où ℓ est un hyperparamètre appartenant à l'espace convexe suivant :

$$\mathbb{L} = \left\{ \ell \in \mathbb{R}^d : \begin{array}{r} \ell_i \in [-c_i, c_i], c_i > 0, \\ i = \{1, \ldots, d\} \end{array} \right\} \quad (13)$$

En utilisant cet ensemble de lois a priori, nous pouvons aisément trouver la famille de lois a posteriori en appliquant le théorème de Bayes (ou en appliquant directement l'équation (17) de l'article [3]).

$$\mathcal{M}_n = \left\{ \mu|\overline{\boldsymbol{x}}_n, \ell \propto \mathcal{N}\left(\frac{\ell + n\overline{\boldsymbol{x}}_n}{n}, \frac{1}{n}\hat{\Sigma} \right), \atop \ell \in \mathbb{L} \right\} \quad (14)$$

où n et $\overline{\boldsymbol{x}}_n$ représentent la nombre d'individus et la moyenne empirique d'une catégorie. Et en appliquant le corollaire 4.7 d'après [3], les estimations MAP pour le paramètre $\hat{\mu}$ se trouvent dans l'espace convexe suivant :

$$\mathbb{G} = \left\{ \hat{\mu} \in \mathbb{R}^d \,\middle|\, \begin{array}{c} \hat{\mu}_i \in \left[\dfrac{-c_i + n\overline{\boldsymbol{x}}_{i,n}}{n}, \dfrac{c_i + n\overline{\boldsymbol{x}}_{i,n}}{n} \right], \\ \forall i = \{1, ..., d\} \end{array} \right\} \quad (15)$$

1. Celui-ci répond aux exigences de l'ignorance préalable (i.e. l'invariant à translation, l'ignorance a priori, l'apprentissage et la convergence).
2. Le lecteur courageux étant renvoyé à l'article [3] pour les développements théoriques.

Contrairement à l'analyse discriminante linéaire classique où l'estimateur μ a une valeur précise (ou exacte), ici nous avons un ensemble convexe d'estimateurs plausibles.

Remarque 3.1 *La propriété de convergence de la famille des lois a priori proposée dans (12) nous assure que peu importe la valeur initiale de notre espace convexe $\mathbb{L}$, lorsque le nombre d'observations tend vers l'infini, i.e. $n \to \infty$, leur influence sur l'inférence des lois a posteriori disparaîtra, i.e $\hat{\mu} = \frac{\ell + n\overline{\boldsymbol{x}}_n}{n} \xrightarrow[n\to\infty]{} \overline{\boldsymbol{x}}_n$, et deviendra l'estimateur asymptotique de la loi gaussienne précise.*

Exemple 3.1 *L'intérêt de modéliser une moyenne imprécise est de pouvoir créer une zone d'imprécision de prise de décision dans laquelle on pourra prédire un ensemble de catégories. Par exemple dans la figure 1, nous avons simulé deux groupes d'individus x_{a*} et x_{b*} (i.e. cas binaire), chacun ayant deux régresseurs non-corrélés et de moyennes différentes.*

$$\begin{pmatrix} x_{a1} \\ x_{a2} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0.25 \\ 0.5 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$$
$$\begin{pmatrix} x_{b1} \\ x_{b2} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0.5 \\ -1.0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$$
$$\mathbb{L} = \left\{\ell \in \mathbb{R}^2 : \ell_i \in [-c_i, c_i], c_i = 2\right\}$$

Dans cet exemple, nous avons représenté dans la figure 1 ; les groupes x_{a*} et x_{b*} avec le symbole $\star$ et $\blacktriangledown$ respectivement, l'espace convexe initial $\mathbb{L}$ (en pointillé) et l'espace convexe $\mathbb{G}$ (en solide) après l'information apportée par nos données.

*Nous avons aussi représenté la moyenne précise comme un point solide au centre de l'ellipse de chaque catégorie, et un point noir (•) représentant un nouvel individu ainsi que les emplacements des estimations inférieures et supérieures de la **moyenne imprécise** de chaque catégorie.*

Dans la figure 2, nous constatons la création d'une zone d'indécision issue de la moyenne imprécise, celle-ci a deux frontières de décision linéaire par morceaux[3].

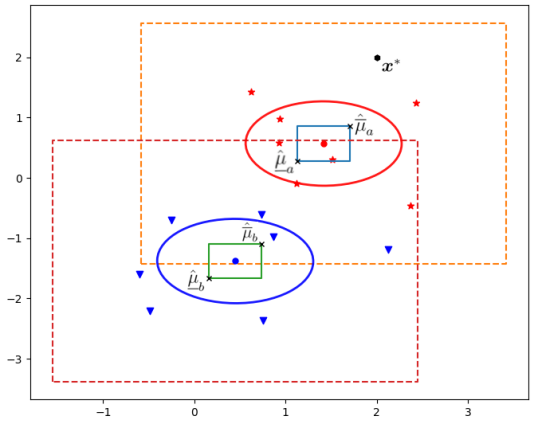

Figure 1 – Estimation supérieure et inférieure de la moyenne imprécise.

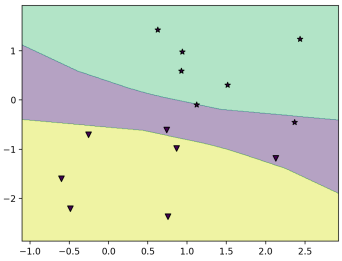

Figure 2 – Frontière d'indécision de prise de décision (couleur violet).

Ainsi donc, en reprenant les équations (10) et (11), nous pouvons voir que la fonction à optimiser sous contraintes est une fonction convexe de $\hat{\mu}$ car ; (1) $\hat{\Sigma}$ est semi-définie positive (et donc son inverse aussi) et (2) les contraintes $\mathbb{G}$ appartiennent aussi à un espace convexe. Nous posons donc le problème d'optimisation suivi de l'équation (10) afin de calculer la borne supérieure de la moyenne imprécise d'un nouvel individu appartenant *plausiblement* à la catégorie a.

$$\hat{\bar{\mu}}_a = \arg\min \frac{1}{2}\hat{\mu}_a^T \hat{\Sigma}^{-1} \hat{\mu}_a + q^T \hat{\mu}_a$$
$$s.t. \ \frac{-c_j + n\overline{x}_{j,n}}{n} \leq \hat{\mu}_{j,a} \leq \frac{c_i + n\overline{x}_{j,n}}{n} \quad \text{(BQP)}$$
$$\forall j = \{1, ..., d\}$$
$$q^T = -\boldsymbol{x}^{*T}\hat{\Sigma}^{-1}$$

Cette formulation est bien connue comme un problème d'optimisation quadratique sous

3. Du fait que les équations (10)- (11) ont différentes solutions en différents endroits de l'espace des attributs.

contrainte de boîte [8] (ou, *Box-constraint Quadratic Program* (BQP)).

Obtenir une solution optimale globale en temps polynomial de ce dernier problème d'optimisation est à nos jours facile en utilisant n'importe quel langage de programmation. Dans notre cas, nous avons utilisé la libraire python nommé *CvxOpt* [7].

Néanmoins, dans le cas du calcul de la borne inférieure de la probabilité $\underline{P}(\boldsymbol{x}^*|y = m_a)$ soumis aux mêmes contraintes, nous nous retrouvons avec la maximisation du problème convexe (BQP) avec contraintes. Ce problème est non-convexe (NBQP) si la matrice $\hat{\Sigma}^{-1}$ est semi-définie positive ou indéfinie.

$$\underline{\hat{\mu}}_a = \arg\max_{\hat{\mu}_a \in \mathbb{G}} \frac{1}{2} \mu_a^T \hat{\Sigma}^{-1} \hat{\mu}_a + q^T \hat{\mu}_a \qquad \text{(NBQP)}$$

NBQP est NP-Hard [12]. Cependant, Burer and Vandenbussche ont développé un puissant algorithme de séparation et évaluation [4] (B&B) [5] (i.e. *Branch-and-Bound*) qui emploie une ramification finie basée sur les conditions de Karush-Kuhn-Tucker [5] du premier ordre, et en appliquant dans chaque noeud de l'arbre B&B un problème relaxé semi-défini polyédrique. Ce dernier algorithme trouve en moyenne la solution optimale en temps *"polynomial"*, malgré un pire cas en temps exponentiel. Dans la section (4), nous présenterons les temps d'exécution pour différents jeux de données.

4 Expérimentations

Cette section aborde les premiers résultats expérimentaux dans le but d'évaluer la performance de l'approche d'analyse discriminante linéaire imprécise (ADLI) (ou classifieur prudent).

4.1 Choix du paramètre c_i

Le choix de la valeur c_i détermine l'impact sur les inférences a posteriori, i.e. s'il est aussi grand que possible, les inférences a posteriori seront donc plus prudentes et cohérentes avec le manque d'information a priori. En l'absence d'informations *a priori*, nous considérons une boîte symétrique autour de $\mathbb{L}$ [3], i.e. $c_i = c, \forall i = \{1, ..., d\}, c > 0$.

Ici, pour fixer c, nous nous baserons sur le taux de convergence de l'imprécision a posteriori [15] (i.e. la différence entre l'espérance supérieure et inférieure) :

$$\overline{E}[\mu|\overline{\boldsymbol{x}}_n, c] - \underline{E}[\mu|\overline{\boldsymbol{x}}_n, c] \xrightarrow[n\to\infty]{} 0 \qquad (16)$$

Avec de petites valeurs de ℓ, on atteint une convergence plus rapide de l'équation (16) vers une valeur précise, la largeur de l'intervalle $[\underline{E}[\mu|\overline{\boldsymbol{x}}_n, c], \overline{E}[\mu|\overline{\boldsymbol{x}}_n, c]]$ étant $\frac{2c}{n}$, avec une valeur conseillée par [3, §4.3, §8] de $c \leq 0.75$. Donc pour obtenir un bon compromis entre justesse et précision des inférences lequels il sera montré de façon empirique dans la figure 4, nous restreignons c à l'intervalle $[0.01, 2]$. Cet intervalle sera donc discrétisé comme suit : $[0.01, 0.02, ..., 0.09, 0.1, 0.2, ..., 2]$, et la valeur optimale décidée par une validation croisée 10-folds sur le jeu de données d'apprentissage.

4.2 Configuration

Ici, nous présenterons les premiers résultats expérimentaux effectués sur 3 jeux de données extraits de dépôt UCI (c.f. Tableau 2). Nous comparons la performance de l'approche ADLI [6] contre l'ADL précis [7] en utilisant la *"validation croisée hold-out"* (i.e. celle-ci divise l'échantillon d'apprentissage en deux : un ensemble d'entraînement contenant 60 % des données et un ensemble de validation), que nous répétons

4. Nommé QuadProgBB et placé sur `https://github.com/sburer/QuadProgBB`

5. Appelé aussi KKT, celui-ci permet de résoudre des problèmes d'optimisation sous contraintes non linéaires d'inégalités.

6. L'implémentation utilisée se trouve au dépôt *Github* : `https://github.com/sdestercke/classifip`

7. L'implémentation LDA utilisée se trouve dans : `http://scikit-learn.org/`

10 fois avec ré-échantillonnage.

Tableau 2 – Jeux de données d'expérimentation

#	Nom	# observations	# régresseurs	# catégories
a	iris	150	4	3
b	seeds	210	7	3
c	glass	214	9	6

Dans le but d'évaluer la performance de l'ensemble de prédictions plausibles $\hat{Y}$ effectuée par notre classifieur ADLI issue de l'utilisation de la maximalité définie auparavant, contre la seule prédiction plausible $\hat{y}$ du classifieur ADL précise, nous avons besoin d'un évaluateur (i.e. une fonction). Nous reprenons la métrique proposée et justifiée théoriquement par [18], qui permet de récompenser l'imprécision de manière plus ou moins forte. Cette métrique s'écrit :

$$u(y, Y) = \begin{cases} 0 & \text{si } y \notin Y, \\ \frac{\alpha}{|Y|} - \frac{\alpha-1}{|Y|} & \text{autrement.} \end{cases} \quad (17)$$

[18] montre qu'une valeur $\alpha = 1$ revient à ne pas récompenser le fait d'être prudent (l'imprécision étant confondue avec l'aléatoire), et que $\alpha > 1$ récompense l'imprécision. Nous utilisons les valeurs usuelles u_{65} avec $\alpha = 1.6$ et u_{80} avec $\alpha = 2.2$.

4.3 Evaluations

Les résultats moyens obtenus en fonction de u_{65} et u_{80} et le temps moyen d'exécution pour prédire la catégorie d'un nouvel individu sont affichés dans le tableau 3. En premier lieu, nous constatons une légère augmentation de précision si la moyenne devient imprécise, et en deuxième lieu, les temps d'exécutions sont raisonnables compte-tenu des problèmes à résoudre (e.g. un problème non-convexe).

Dans la figure (3), et en utilisant deux régresseurs du jeu de données IRIS, nous voulons souligner comme l'imprécision joue un rôle important en créant une frontière de décision (ne pas confondre avec aire de rejet) plus épaisse, laquelle héberge l'ensemble des catégories plausibles.

Tableau 3 – Précision moyenne des fonctions d'utilité u_{65} et u_{80}.

#	ADL	ADLI		Temps de prédiction d'un individu avec ADLI
		u_{65}	u_{80}	
a	0.961	0.969	0.975	0.56 sec.
b	0.959	0.959	0.962	1.50 sec.
c	0.594	0.589	0.642	8.66 sec

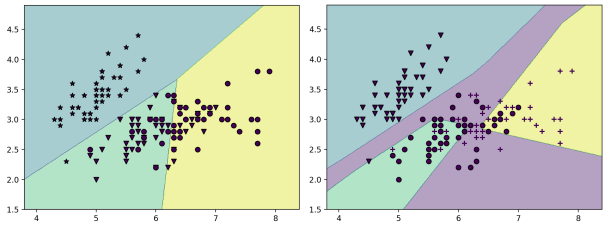

Figure 3 – Frontière de décision précise (à gauche) et d'indécision (à droite) du jeu de données IRIS.

Dans la figure (4), nous montrons l'évolution des évaluations en fonction de l'imprécision des estimateurs. Nous voyons qu'un paramètre c trop élevé est dommageable, et devrait rester limité en valeur.

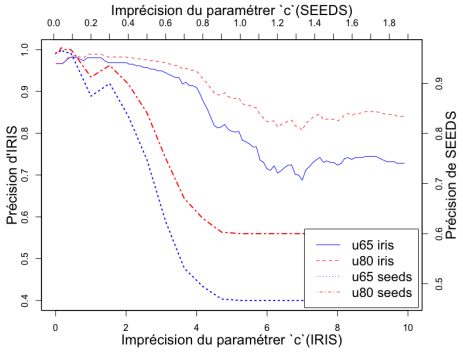

Figure 4 – Rapport de performance de prédiction issue à la moyenne imprécision.

5 Conclusions

Cet article introduit une nouvelle méthode de classification prudente (ou imprécise), qui généralise le modèle d'analyse discriminante linéaire.

Plusieurs pistes de recherche s'offrent à nous : premièrement considérer l'ensemble des modèles Gaussiens possibles (matrices de covariances diagonales, hétéroscédasticité, etc.) et la complexité des problèmes d'inférences associés, deuxièmement considérer le problème où

μ est précis mais Σ est imprécise. Cela pose le problème de définir un ensemble convexe pour les matrices inverses Σ^{-1}.

Références

[1] Thomas Augustin, Frank PA Coolen, Gert de Cooman, and Matthias CM Troffaes. *Introduction to imprecise probabilities*. John Wiley & Sons, 2014.

[2] Alessio Benavoli and Branko Ristic. Classification with imprecise likelihoods : A comparison of tbm, random set and imprecise probability approach. In *Proceedings of the 14th International Conference on Information Fusion*, pages 1–8. IEEE, 2011.

[3] Alessio Benavoli and Marco Zaffalon. Prior near ignorance for inferences in the k-parameter exponential family. *Statistics*, 49(5) :1104–1140, 2014.

[4] José M Bernardo and Adrian FM Smith. *Bayesian Theory*. John Wiley & Sons Ltd., 2000.

[5] Samuel Burer and Dieter Vandenbussche. Globally solving box-constrained nonconvex quadratic programs with semidefinite-based finite branch-and-bound. *Computational Optimization and Applications*, 43 (2) :181–195, 2009.

[6] Marco EGV Cattaneo. *Statistical decisions based directly on the likelihood function*. PhD thesis, ETH Zurich, 2007.

[7] Joachim Dahl and Lieven Vandenberghe. CVXOPT : A python package for convex optimization, 2004. URL `http://cvxopt.org/`.

[8] Pasquale L De Angelis, Panos M Pardalos, and Gerardo Toraldo. Quadratic programming with box constraints. In *Developments in global optimization*, pages 73–93. Springer US, 1997.

[9] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Springer New York Inc., 2001.

[10] Yang Gen, Destercke Sébastien, and Masson Marie-Hélène. Nested dichotomies with probability sets for multi-class classification. In *Proceedings of the Twenty-first European Conference on Artificial Intelligence*, pages 363–368. IOS Press, 2014.

[11] Vu-Linh Nguyen, Sébastien Destercke, and Marie-Hélène Masson. K-nearest neighbour classification for interval-valued data. In *International Conference on Scalable Uncertainty Management*, pages 93–106. Springer, 2017.

[12] Panos M Pardalos and Stephen A Vavasis. Quadratic programming with one negative eigenvalue is np-hard. *Journal of Global Optimization*, 1(1) :15–22, 1991.

[13] Christian Robert. *Le choix bayésien : Principes et pratique*. Springer Paris, 2005.

[14] Matthias CM Troffaes. Decision making under uncertainty using imprecise probabilities. *International Journal of Approximate Reasoning*, 45(1) :17–29, 2007.

[15] P. Walley. *Statistical reasoning with imprecise Probabilities*. Chapman and Hall, 1991.

[16] Marco Zaffalon. Statistical inference of the naive credal classifier. In *International Symposium on Imprecise Probability : Theories and Applications*, pages 384–393, 2001.

[17] Marco Zaffalon. The naive credal classifier. *Journal of statistical planning and inference*, 105(1) :5–21, 2002.

[18] Marco Zaffalon, Giorgio Corani, and Denis Mauá. Evaluating credal classifiers by utility-discounted predictive accuracy. *International Journal of Approximate Reasoning*, 53(8) :1282–1301, 2012.