# Imprecise Linear Discriminant Analysis based on robust Bayesian inference

## 27èmes rencontres francophones sur la logique floue et ses applications


**CARRANZA-ALARCON Yonatan-Carlos**

Ph.D. Candidate in Computer Science

**DESTERCKE Sébastien**

Ph.D Director

utc Université de Technologie Compiègne

heudiasyc

cnrs dépasser les frontières


08 Nov 2018 to 09 Nov 2018

# Overview

## Imprecise Linear Discriminant Analysis Classifica tion

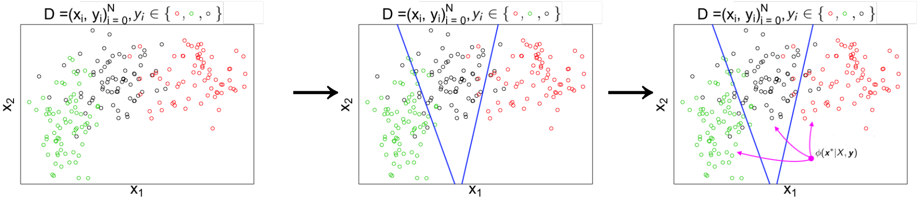
- Classification
  - Motivation
  - Decision Making
  - Discriminant Analysis
- Imprecise Classification
  - Imprecise Linear discriminant analysis
  - Cautious Decision
- Evaluation
  - Cautious accuracy measure
  - Experiments
- Conclusions

# Overview

- Classification
  - Motivation
  - Decision Making
  - Discriminant Analysis
- Imprecise Classification
  - Imprecise Linear discriminant analysis
  - Cautious Decision
- Evaluation
  - Cautious accuracy measure
  - Experiments
- Conclusions

# Classification - Outline (Example)

- Data training $D = \{x_i, y_i\}_{i=0}^{N}$ such that :
  - (Input) $x_i \in \mathscr{X}$ are regressors or features (often $x_i \in \mathbb{R}^p$).
  - (Output) $y_i \in \mathscr{K}$ is a response category variable, with $\mathscr{K} = \{ m_1, ..., m_K \}$




## Objective

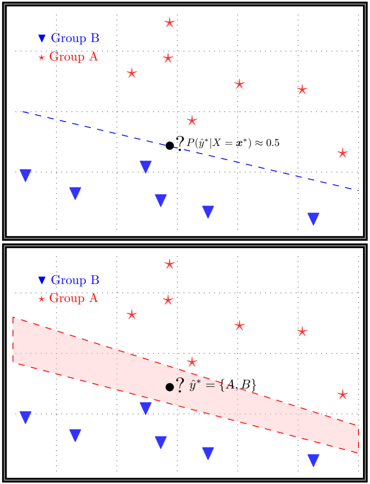
Given training data $D = \{x_i, y_i\}_{i=0}^{N}$, we need to learn a classification rule : $\phi : \mathscr{X} \rightarrow \mathscr{Y}$ in order to predict a new observation $\phi(\boldsymbol{x}^*)$

# Motivation

## What is the bigger problem in (precise) Classification ?

- Precise models can produce many mistakes for hard to predict unlabeled instances.

- One way to recognize such instances and avoid making such mistakes too often → **Making a cautious decision.**



# Decision Making in Statistic

1. In Statistic : An optimal model for classification under $1/0$ loss function is Bayes Classifier :

$$\phi(\boldsymbol{x}^*) := \arg\max_{m_k \in \mathcal{K}} P(y = m_k | X = \boldsymbol{x}^*) \quad (1)$$

2. Preference ordering :

## Definition (Preference ordering [3, pp. 47])

Let $P$ a conditional probability distribution, $m_a$ is preferred to $m_b$ :

$$m_a \succ m_b \iff P(y = m_a | X = \boldsymbol{x}^*) > P(y = m_b | X = \boldsymbol{x}^*)$$

We then take the ***maximal element*** of the complete order $\succ$ :

$$m_{i_K} \succ m_{i_{K-1}} \succ .... \succ m_{i_1} \iff P(y = m_{i_K} | \boldsymbol{x}^*) > .... > P(y = m_{i_1} | \boldsymbol{x}^*)$$

# Decision Making in Statistic

$$m_{i_K} \succ m_{i_{K-1}} \succ .... \succ m_{i_1} \iff$$

$$P(y = m_{i_K} | \boldsymbol{x}^*)$$

$$\vee$$

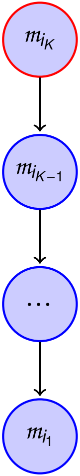
$$P(y = m_{i_K} | \boldsymbol{x}^*)$$

$$\vee$$

$$\cdots$$

$$\vee$$

$$P(y = m_{i_1} | \boldsymbol{x}^*)$$



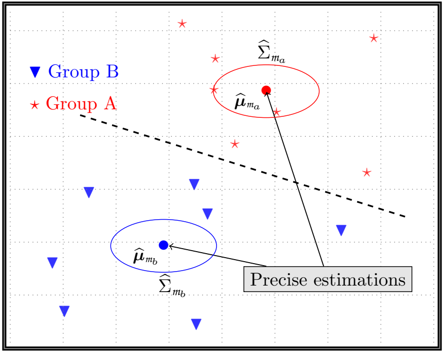

- How can we estimate the cond. probability distribution $P_{Y|X}$ ?

# (Precise) Linear Discriminant Analysis

Applying Baye's rules to $P(Y = m_a | X = \boldsymbol{x}^*)$ :

$$P(y = m_k | X = \boldsymbol{x}^*) = \frac{P(X = \boldsymbol{x}^* | y = m_k) P(y = m_k)}{\sum_{m_l \in \mathcal{K}} P(X = \boldsymbol{x}^* | y = m_l) P(y = m_l)}$$

Normality $P_{X|Y=m_k} \sim \mathcal{N}(\mu_k, \Sigma_k)$ and precise marginal $\pi_k := P_{Y=m_k}$.



# Overview

- Classification
  - Motivation
  - Decision Making
  - Discriminant Analysis
- Imprecise Classification
  - Imprecise Linear discriminant analysis
  - Cautious Decision
- Evaluation
  - Cautious accuracy measure
  - Experiments
- Conclusions

# Imprecise Linear Discriminant Analysis

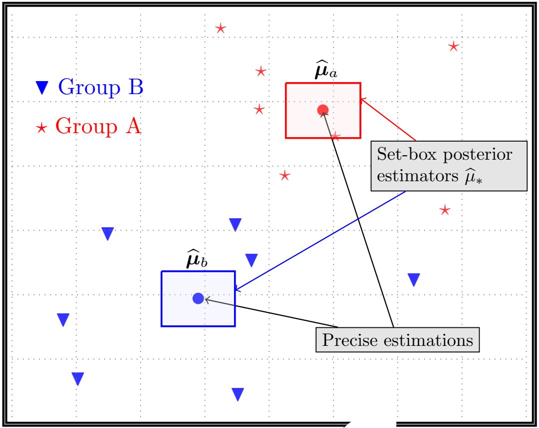
**Objective :** *Making imprecise the parameter mean $\mu_k$ of each Gaussian distribution family $\mathscr{G}_k := P_{X|Y=m_k} \sim \mathscr{N}(\mu_k, \widehat{\Sigma})$*

**Proposition :** Using a set of posterior distribution $\mathscr{P}$ ([4, eq 17]).

$$\mathscr{P}_{\mu_k} = \left\{ \begin{array}{l} \mu_k | \ell \propto \mathscr{N}\left(\dfrac{\ell + n_k \overline{\boldsymbol{x}}_{n_k}}{n_k}, \dfrac{1}{n_k}\widehat{\Sigma}\right) \\ \overline{\boldsymbol{x}}_k = \dfrac{1}{n_k}\sum_{i=1}^{n_k} x_{i,k}, \ell \in \mathbb{L} \end{array} \right\}$$

where $\mathbb{L}$ is a hypercube.

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



# Decision Making in Imprecise Probabilities

## Definition (Partial Ordering by Maximality [1])

Let $\mathscr{P}$ a set of probabilities then $m_a$ is preferred to $m_b$ if the cost of exchanging $m_a$ with $m_b$ if and only if :

$$m_a \succ_M m_b \iff \inf_{P \in \mathscr{P}} \frac{P(y = m_a | X = \boldsymbol{x}^*)}{P(y = m_b | X = \boldsymbol{x}^*)} > 1 \quad (2)$$

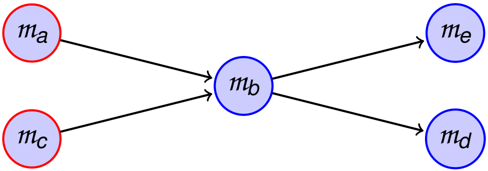
- This definition give us a partial order $\succ_M$, such as the maximal element of partial order is the cautious decision :

$$Y_M = \{m_a \in \mathscr{K} \mid \nexists m_b : m_a \succ_M m_b\}$$

# Decision Making in Imprecise Probabilities

For instance, if we have $\mathscr{K} = \{m_a, m_b, m_c, m_d, m_e\}$, we could have a set of comparisons from last definition :

$$\{m_a \succ_M m_b, m_c \succ_M m_b, m_a \succ\prec_M m_c, m_b \succ_M m_d, m_b \succ_M m_d, m_d \succ\prec_M m_e\}$$




Where cautious decision is a set-value categories : $Y_M = \{m_a, m_c\}$

# Decision Making in ILDA

- We take to equation (2) of the criterion of maximality and applying Bayes' rule :

$$\inf_{P_{X|y} \in \mathscr{P}_{X|y}} \frac{P(\boldsymbol{x}^*|y = m_a)P(y = m_a)}{P(\boldsymbol{x}^*|y = m_b)P(y = m_b)} > 1$$

- Assuming an precise estimation for marginal : $\pi_k := \widehat{P}_{Y=m_k}$ of LDA.

$$\frac{\widehat{P}(y = m_a)}{\widehat{P}(y = m_b)} \inf_{P_{X|y} \in \mathscr{P}_{X|y}} \frac{P(\boldsymbol{x}^*|y = m_a)}{P(\boldsymbol{x}^*|y = m_b)} > 1$$

- Given that normality assumption $P_{X|Y=m_k} \sim \mathcal{N}(\mu_k, \widehat{\Sigma})$ of LDA, each component of numerator/denominator are independent, then :

$$\frac{\widehat{P}(y = m_a)}{\widehat{P}(y = m_b)} \frac{\inf_{P_{X|y=m_a} \in \mathscr{P}_{X|y=m_a}} P(\boldsymbol{x}^*|y = m_a)}{\sup_{P_{X|y=m_b} \in \mathscr{P}_{X|y=m_b}} P(\boldsymbol{x}^*|y = m_b)} > 1$$

# Decision Making in ILDA

- We take to equation (2) of the criterion of maximality and applying Bayes' rule :

$$\inf_{P_{X|y} \in \mathscr{P}_{X|y}} \frac{P(\boldsymbol{x}^*|y = m_a)P(y = m_a)}{P(\boldsymbol{x}^*|y = m_b)P(y = m_b)} > 1$$

- Assuming an precise estimation for marginal : $\pi_k := \widehat{P}_{Y=m_k}$ of LDA.

$$\frac{\widehat{P}(y = m_a)}{\widehat{P}(y = m_b)} \inf_{P_{X|y} \in \mathscr{P}_{X|y}} \frac{P(\boldsymbol{x}^*|y = m_a)}{P(\boldsymbol{x}^*|y = m_b)} > 1$$

- Given that normality assumption $P_{X|Y=m_k} \sim \mathscr{N}(\mu_k, \widehat{\Sigma})$ of LDA, each component of numerator/denominator are independent, then :

$$\frac{\widehat{P}(y = m_a)}{\widehat{P}(y = m_b)} \frac{\inf_{P_{X|y=m_a} \in \mathscr{P}_{X|y=m_a}} P(\boldsymbol{x}^*|y = m_a)}{\sup_{P_{X|y=m_b} \in \mathscr{P}_{X|y=m_b}} P(\boldsymbol{x}^*|y = m_b)} > 1$$

# Decision Making in ILDA

- We take to equation (2) of the criterion of maximality and applying Bayes' rule :

$$\inf_{P_{X|y} \in \mathscr{P}_{X|y}} \frac{P(\boldsymbol{x}^*|y = m_a)P(y = m_a)}{P(\boldsymbol{x}^*|y = m_b)P(y = m_b)} > 1$$

- Assuming an precise estimation for marginal : $\pi_k := \widehat{P}_{Y=m_k}$ of LDA.

$$\frac{\widehat{P}(y = m_a)}{\widehat{P}(y = m_b)} \inf_{P_{X|y} \in \mathscr{P}_{X|y}} \frac{P(\boldsymbol{x}^*|y = m_a)}{P(\boldsymbol{x}^*|y = m_b)} > 1$$

- Given that normality assumption $P_{X|Y=m_k} \sim \mathcal{N}(\mu_k, \widehat{\Sigma})$ of LDA, each component of numerator/denominator are independent, then :

$$\frac{\widehat{P}(y = m_a)}{\widehat{P}(y = m_b)} \quad \frac{\inf_{P_{X|y=m_a} \in \mathscr{P}_{X|y=m_a}} P(\boldsymbol{x}^*|y = m_a)}{\sup_{P_{X|y=m_b} \in \mathscr{P}_{X|y=m_b}} P(\boldsymbol{x}^*|y = m_b)} > 1$$

# Decision Making in ILDA

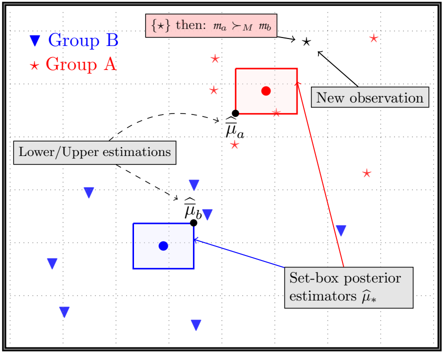
- We then have two optimization : box-constrained quadratic problem (BQP) and non-convex BQP (NBQP)

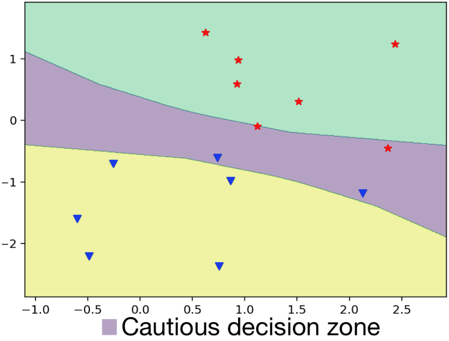
$$\inf_{P \in \mathscr{P}_{\mu_a}} P(\boldsymbol{x}^* | y = m_a) \iff \underline{\mu}_a = \arg\min_{\mu_a \in \mathscr{P}_{\mu_a}} -\frac{1}{2}(\boldsymbol{x}^* - \mu_a)^T \widehat{\Sigma}^{-1} (\boldsymbol{x}^* - \mu_a) \text{ (NBQP)}$$

$$\sup_{P \in \mathscr{P}_{\mu_b}} P(\boldsymbol{x}^* | y = m_b) \iff \overline{\mu}_b = \arg\max_{\mu_b \in \mathscr{P}_{\mu_a}} -\frac{1}{2}(\boldsymbol{x}^* - \mu_b)^T \widehat{\Sigma}^{-1} (\boldsymbol{x}^* - \mu_b) \text{ (BQP)}$$

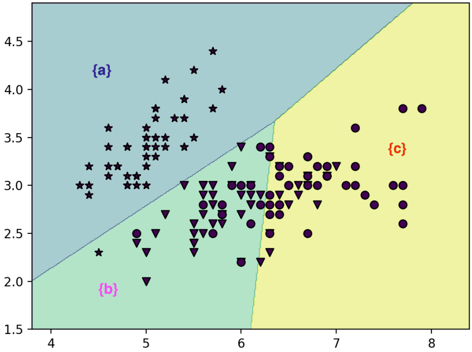
- First problem non-convex BQP → solved through Branch and Bound method.

# Cautious decision zone (example with 2 class)

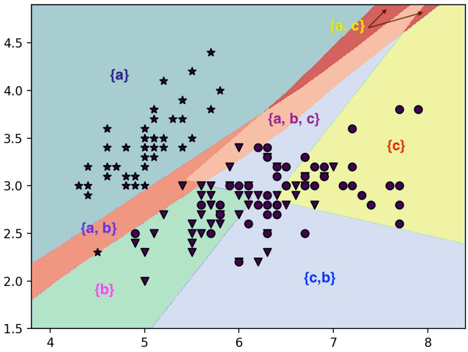


# Another Example with 3 class $\{a, b, c\}$




(a) Precise Classifier




(b) Cautious Classifier

# Overview

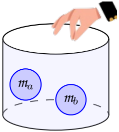
- Classification
  - Motivation
  - Decision Making
  - Discriminant Analysis
- Imprecise Classification
  - Imprecise Linear discriminant analysis
  - Cautious Decision
- Evaluation
  - Cautious accuracy measure
  - Experiments
- Conclusions

# Utility-discount accuracy measure

**Example :** Given a ground-truth category $\{m_a\}$ and a binary cautious prediction $\{m_a, m_b\}$, how should we reward cautious prediction ?

**Case 1 :** If we reward it 0.5 :

$$\text{Reward} = \frac{1}{|\{m_a, m_b\}|}$$




$\Longrightarrow$ Equivalent to randomly pick out $m_b$ or $m_a$.

→ too low reward confusing cautiousness and randomness.

**Case 2 :** If we reward it with 1, then classifier always returning all classes is the best → too high reward to cautiousness, no penalty for non-informativeness.

$\Longrightarrow$ *Actual reward should be between those two cases, and depend on how much we are cautiousness seeking.*

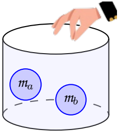
# Utility-discount accuracy measure

**Example :** Given a ground-truth category $\{m_a\}$ and a binary cautious prediction $\{m_a, m_b\}$, how should we reward cautious prediction ?

**Case 1 :** If we reward it 0.5 :

$$\text{Reward} = \frac{1}{|\{m_a, m_b\}|}$$

$\Longrightarrow$ Equivalent to randomly pick out $m_b$ or $m_a$.




→ too low reward confusing cautiousness and randomness.

**Case 2 :** If we reward it with 1, then classifier always returning all classes is the best → too high reward to cautiousness, no penalty for non-informativeness.

$\Longrightarrow$ *Actual reward should be between those two cases, and depend on how much we are cautiousness seeking.*

# Utility-discount accuracy measure

*Zaffalon et al* in [2] proposes an utility-discounted accuracy measure :

$$u(y, \widehat{Y}_M) = \begin{cases} 0 & \text{if } y \notin \widehat{Y}_M \\ \frac{\alpha}{|\widehat{Y}_M|} - \frac{1-\alpha}{|\widehat{Y}_M|^2} & \text{else} \end{cases}$$

Where $u(\cdot, \cdot)$ rewards to be more informative in cautious decision : $u_{65}$ with a gain of 0.65 and $u_{80}$ with a gain of 0.8.

# Experiments

Cross-Validation with 60% training data and 40% validation data, it is repeated with 10 resampling.

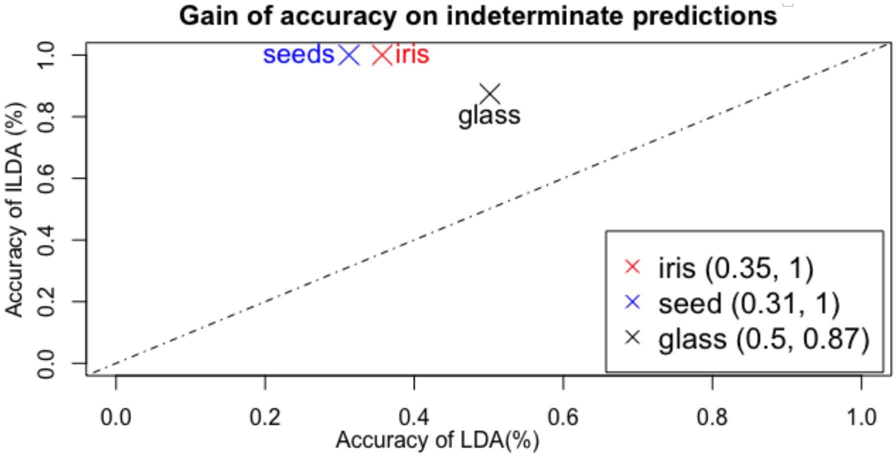
| # | Name | # Obs. | # Regr. | # Classes |
|---|---|---|---|---|
| a | iris | 150 | 4 | 3 |
| b | seeds | 210 | 7 | 3 |
| c | glass | 214 | 9 | 6 |

TABLE – Data sets used in the experiments

| # | DLA | IDLA | | Inference time |
|---|---|---|---|---|
| | | $u_{65}$ | $u_{80}$ | |
| 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 |

TABLE – Average utility-discounted accuracies

# Experiments



# Overview

- Classification
  - Motivation
  - Decision Making
  - Discriminant Analysis
- Imprecise Classification
  - Imprecise Linear discriminant analysis
  - Cautious Decision
- Evaluation
  - Cautious accuracy measure
  - Experiments
- Conclusions

# Conclusions and Perspectives

## Imprecise Linear Discriminant Classification Analysis

- Increasing in imprecision on the estimators has allowed us to be more cautious in doubt and to improve the prediction of classification.
- More experiments with all imprecise components.
- Creation of new imprecise statistic models for a sensibility analysis and a more (cautious) robust prediction.

# References

Matthias CM Troffaes. “Decision making under uncertainty using imprecise probabilities”. In : *International Journal of Approximate Reasoning* 45.1 (2007), p. 17-29.

Marco Zaffalon, Giorgio Corani et Denis Mauá. “Evaluating credal classifiers by utility-discounted predictive accuracy”. In : *International Journal of Approximate Reasoning* 53.8 (2012), p. 1282-1301.

James O Berger. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

Alessio Benavoli et Marco Zaffalon. “Prior near ignorance for inferences in the k-parameter exponential family”. In : *Statistics* 49.5 (2014), p. 1104-1140.


Thank
You!