

Distributionally robust, cautious inferences in supervised classification using imprecise probabilities

CARRANZA ALARCÓN Yonatan-Carlos

Ph.D. Candidate in Computer Science

DESTERCKE Sébastien

Ph.D Director

UMR CNRS 7253 Heudiasyc, Sorbonne universités, Université de technologie de Compiègne CS 60319 - 60203 Compiègne cedex, France



December 8, 2020

Overview

- Problem statements
- Imprecise Gaussian Discriminant Analysis
- Cautious inferences in multi-label problems
- Conclusions and perspectives

Overview

- Problem statements
 - Introduction and Motivation
 - Imprecise Probabilities
- Imprecise Gaussian Discriminant Analysis
- Cautious inferences in multi-label problems
- Conclusions and perspectives

Supervised classification approach

Given the training data $\mathcal{D} = \{x_i, y_i\}_{i=0}^N \subseteq \mathbb{R}^p \times \{m_a, \dots, m_e\}$:

Step 1 Learning a classification rule: $\varphi: \mathcal{X} \rightarrow \mathcal{K}$.

Step 2 Making decision on a new instance $\hat{\varphi}(\mathbf{x}), \mathbf{x} \in \mathcal{I}$

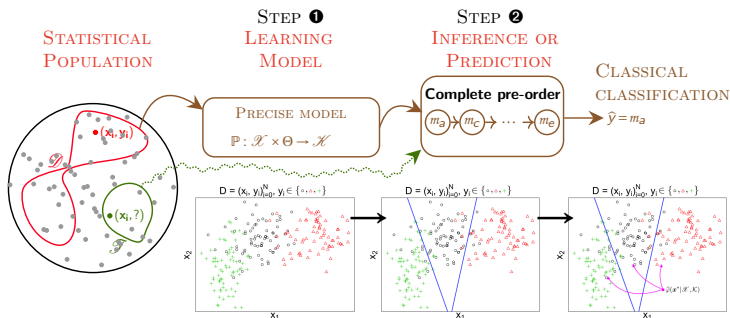


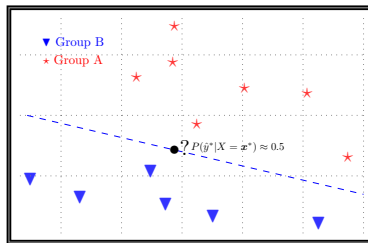
Figure: Supervised classification learning in a precise approach.

What is an important problem in (precise) classification?

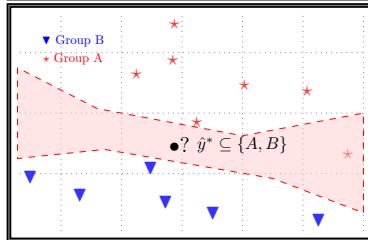
Motivation

What is an important 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.**



- ✓ **Set of potential decisions.**

How can we make cautious decisions?

- Different existing ways of proceeding:
 - Partial reject [[Herbei et al. 2006](#)]
 - Conformal predictions [[Vovk et al. 2018](#)]
 - Partially ordered decisions [[Troffaes 2007](#)]
 - ...
- We adopt an imprecise probabilistic viewpoint.
 - A partial order on a set of decisions

Imprecise Supervised classification approach

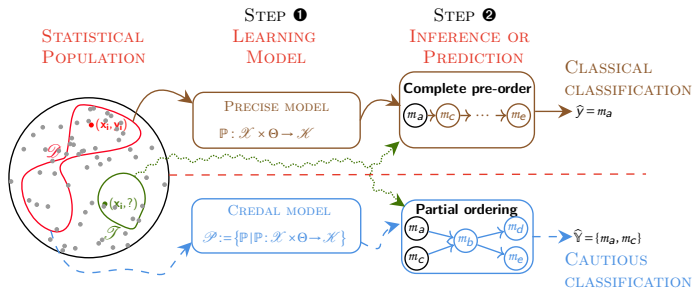


Figure: Statistical learning in imprecise and precise approach.

● Contributions

- Multiclass classification:
 1. An imprecise classifier extending Gaussian discriminant analysis.
- Multi-label classification
 1. More efficient, dedicated algorithm for the Hamming Loss.
 2. First attempt to generalize the classifier chains to IP setting.

Overview

- Problem statements
 - Introduction and Motivation
 - Imprecise Probabilities
- Imprecise Gaussian Discriminant Analysis
- Cautious inferences in multi-label problems
- Conclusions and perspectives

Imprecise probabilities in a nutshell

- ☞ Our uncertainty is described by a convex set \mathcal{P} of probabilities

$$\mathcal{P} := \{\mathbb{P} \mid \mathbb{P} : \mathcal{X} \times \Theta \rightarrow \mathcal{K}\} \quad (\text{Credal set (CS)})$$

- ☞ **How can \mathcal{P} be obtained?**

- Frequentist confidence regions,
- Probability box (P-box),
- ☞ Generalized Bayesian approach (i.e. set of prior distributions)
-

Decision Making under uncertainty

① Decision making using a single precise distribution.

Definition 1 (Complete pre-order)

Given a loss function ℓ and a probability \mathbb{P} , m_a is preferred to m_b if

$$m_a \succ_{\ell}^{\mathbb{P}} m_b \iff \mathbb{E}_{\mathbb{P}} [\ell(m_b, \cdot) - \ell(m_a, \cdot)] > 0$$

② Decision making using a credal set.

Definition 2 (Partial Ordering by Maximality)

Given a loss function ℓ and a set \mathbb{P} , m_a is preferred to m_b if

$$m_a \succ_{\ell}^{\mathcal{P}} m_b \iff \inf_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}} [\ell(m_b, \cdot) - \ell(m_a, \cdot)] > 0.$$

The non-dominated elements of $\succ_{\ell}^{\mathcal{P}}$

$$\mathbb{Y}_{\ell, \mathcal{P}}^M = \left\{ m_a \in \mathcal{K} \mid \nexists m_b : m_b \succ_{\ell}^{\mathcal{P}} m_a \right\}$$



Figure: Complete pre-order

Inference complexity: $\mathcal{O}(|\mathcal{K}|)$

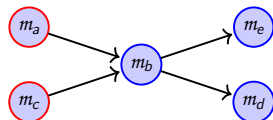


Figure: Partial ordering.

Inference complexity: $\mathcal{O}(|\mathcal{K}|^2)$

Overview

- Problem statements
- Imprecise Gaussian Discriminant Analysis
- Cautious inferences in multi-label problems
- Conclusions and perspectives

(Imprecise) Gaussian discriminant classification

☞ Most existing studies:

- ✓ focus on classifiers applicable only to discrete features \mathcal{X} (e.g. Naive Credal Classifier (NCC) [Zaffalon 2002], Credal C4.5 [Mantas et al. 2014], Credal sum-product networks [Mauá et al. 2017])
- ✓ consider the case of zero-one loss matrix.

☞ Our contribution: extending the Gaussian discriminant analysis

- ✗ not been explored yet,
- ✗ gives an imprecise classifier on continuous features,
- ✗ works under generic loss matrix ℓ .

Overview

- Problem statements
- Imprecise Gaussian Discriminant Analysis
 - Imprecise Gaussian discriminant classification
 - Synthetic data exploring non i.i.d. case
 - Conclusions and Perspective
- Cautious inferences in multi-label problems
- Conclusions and perspectives

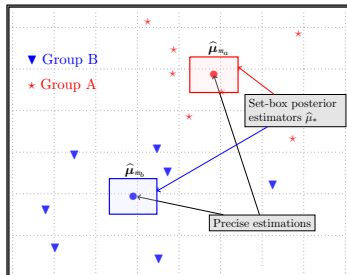
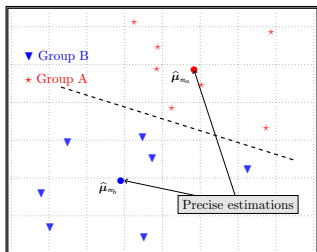
IGDA - Step ① Learning step

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

Assumptions: Precise estimations of $\hat{\Sigma}_{m_k}$ and $\hat{P}(Y = m_k)$.

Proposition: Using the set of prior distributions \mathcal{P}_{μ_k} ([Benavoli et al. 2014, eq 17]).

$\Rightarrow \mu_{m_k}$ belong to an hypercube $\mathbb{G}_{m_k} = \{\mu_{m_k} \mid \mu_{i,m_k} \in [\underline{\mu}_i, \bar{\mu}_i], \forall i \in \{1, \dots, p\}\}$



IGDA - Step ② Predicting/Decision step

- Under the maximality and $\ell_{0/1}$, m_a is preferred to m_b if:

$$\inf_{P \in \mathcal{P}_{X|m_a}} P(\mathbf{x}|Y = m_a) - \sup_{P \in \mathcal{P}_{X|m_b}} P(\mathbf{x}|Y = m_b) > 0$$

- We can reduce it to solving two optimization problems:

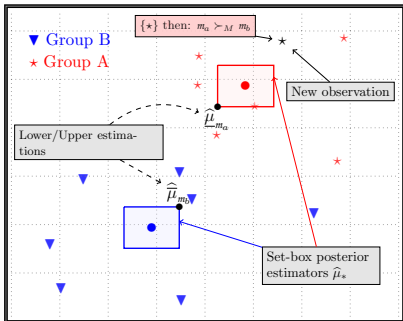
$$\bar{\mu}_{m_b} = \arg \max_{\mu_{m_b} \in \mathbb{G}_{m_b}} -\frac{1}{2}(\mathbf{x} - \mu_{m_b})^T \hat{\Sigma}_{m_b}^{-1}(\mathbf{x} - \mu_{m_b})$$

☺ **Polynomial complexity**

$$\underline{\mu}_{-m_a} = \arg \min_{\mu_{m_a} \in \mathbb{G}_{m_a}} -\frac{1}{2}(\mathbf{x} - \mu_{m_a})^T \hat{\Sigma}_{m_a}^{-1}(\mathbf{x} - \mu_{m_a})$$

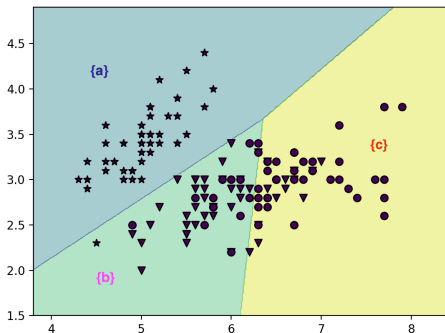
☺ **NP-hard problem**

⇒ solved through Branch & Bound method.

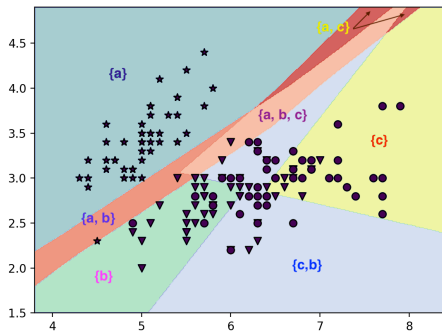


Cautious decision zone of ILDA

$a = \star$, $b = \blacktriangledown$, $c = \bullet$



(a) Precise Classifier



(b) Cautious Classifier

Precise **versus** Cautious

$\{a\}, \{b\}, \{c\}$	$\{a\}, \{b\}, \{c\}$ $\{a, b\}, \{a, c\}, \{b, c\}$ $\{a, b, c\}$
-----------------------	--

Additional theoretical results [Pattern Recognition-2020]

- ✌ We propose different imprecise classifiers depending on assumptions about $\hat{\Sigma}_{m_k}$.

Assumptions	Imprecise GDA	Complexity
Heteroscedasticity: $\hat{\Sigma}_{m_k} = \hat{\Sigma}_k$	IQDA	$\geq \mathcal{O}(p^2)$
Homoscedasticity: $\hat{\Sigma}_{m_k} = \hat{\Sigma}$	ILDA	$\geq \mathcal{O}(p^2)$
Feature independence: $\hat{\Sigma}_{m_k} = \hat{\sigma}_k^T \mathbb{I}$	INDA	$\mathcal{O}(p)$
Unit-variance feature indep.: $\hat{\Sigma}_{m_k} = \mathbb{I}$	IEDA	$\mathcal{O}(p)$

- ✌ We explore the imprecise prior marginal and generic loss matrix cases.
- ✓ Solvable using linear programming or extreme points of \mathcal{P}_Y ,
 - ✓ **The complexity is not increased by much !.**

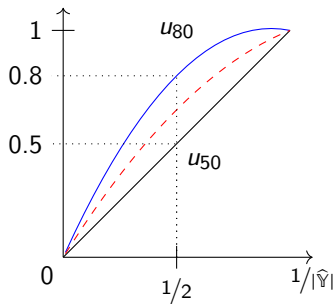
Datasets and experimental setting

- ☞ Data sets issued from UCI repository [Frank et al. 2010].
- ☞ 10×10-fold cross-validation procedure.
- ☞ Utility-discounted accuracy measure proposed in [Zaffalon et al. 2012].

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

with $u(\hat{Y}, y) = 1$ if $|\hat{Y}| = 1$ and $\hat{Y} = y$

- Discounted accuracy: $\alpha = 1 \Rightarrow u(\hat{Y}, y) = \frac{1}{|\hat{Y}|}$
 - no reward to cautiousness
 - (cautiousness \equiv randomness)
- u_{65} : $\alpha = 1.6$, moderate reward to cautiousness
- u_{80} : $\alpha = 2.2$, big reward to cautiousness



↓
2 classes predicted,
good one in it

Experimental results

	LDA	ILDA		QDA	IQDA	
#	acc.	u_{80}	u_{65}	acc	u_{80}	u_{65}
iris	97.96	98.38	97.16	97.29	98.08	97.13
wine	98.85	98.99	98.95	99.03	99.39	99.09
forest	94.61	94.56	94.05	89.43	91.77	88.90
seeds	96.35	96.59	96.51	94.64	95.20	94.72
dermatology	96.58	97.06	96.94	82.47	84.24	84.05
vehicle	77.96	81.98	79.59	85.07	87.96	86.13
vowel	60.10	67.45	62.41	87.83	89.96	88.40
wine-quality	59.25	65.83	60.31	55.62	65.85	60.36
wall	67.96	71.34	66.65	65.87	71.79	69.75
avg.	83.68	86.05	84.03	80.34	87.16	85.33

Table: AVERAGE UTILITY-DISCOUNTED ACCURACIES (%)

✌ Including an imprecise component in the Gaussian discriminant analysis produces reasonably robust cautious predictions

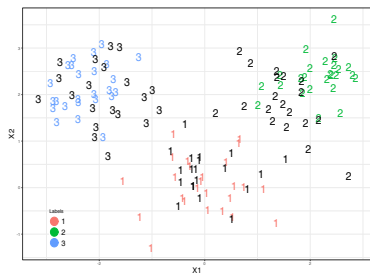
Overview

- Problem statements
- Imprecise Gaussian Discriminant Analysis
 - Imprecise Gaussian discriminant classification
 - Synthetic data exploring non i.i.d. case
 - Conclusions and Perspective
- Cautious inferences in multi-label problems
- Conclusions and perspectives

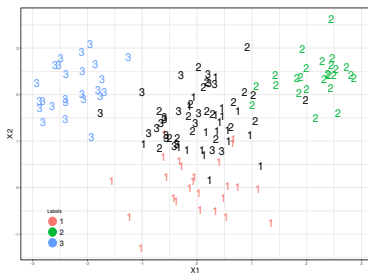
Setting on synthetic data sets non i.i.d.

☞ (Shifting mean) Noise-corrupted test instances of synthetic data sets.

$$\mathcal{T}_1(\epsilon) = \left\{ \mathcal{T}^{m_k} \sim \mathcal{N}(\tilde{\mu}_{m_k}, \Sigma_{m_k}), \tilde{\mu}_{m_k} = (1 - \epsilon)\mu_{m_k} + \epsilon\mu_G, \mu_G = 1/K \sum_{k=1}^K \mu_{m_k} \right\}$$



\mathcal{T}_1 with $\epsilon=0.18$

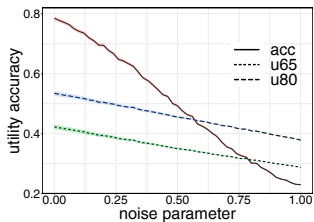


\mathcal{T}_1 with $\epsilon=0.88$

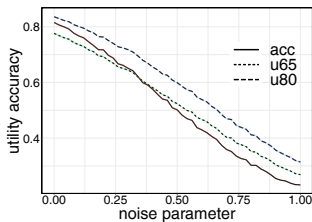
- ☞ Test instances are moved away from its ground-truth sub-population.
- ☞ At higher values of $\epsilon \in [0, 1]$, test instances highly overlap.

Experiments on synthetic data sets

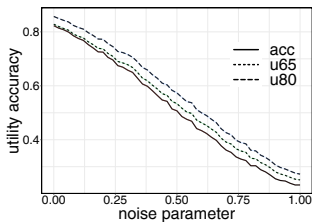
👉 Results of (I)QDA model on corrupt test dataset $\mathcal{T}_1(\epsilon)$ using training data sets with different number of samples:



10 samples



25 samples



50 samples

- 👉 As number of samples increases, performance of the precise and imprecise classifiers converge to similar values.
- 👉 For a small number of training data, the imprecise approaches are quite robust to change in the distributions.

Overview

- Problem statements
- Imprecise Gaussian Discriminant Analysis
 - Imprecise Gaussian discriminant classification
 - Synthetic data exploring non i.i.d. case
 - Conclusions and Perspective
- Cautious inferences in multi-label problems
- Conclusions and perspectives

Conclusions and further issues

✌ Works done

- ✓ A new continuous imprecise classifier extending the classical Gaussian discriminant analysis.
- ✓ A first empirical study concerning the case of non-identically distributed data.
- ✓ An optimized algorithm for a cautious prediction using the maximality criterion.

📖 Some questions to explore

- ? Making imprecise the covariance matrix Σ_{m_k} .
- ? Making imprecise the components eigen-values and -vectors of Σ_{m_k} .
- ? Dealing with a high number of classes and features.

Overview

- Problem statements
- Imprecise Gaussian Discriminant Analysis
- Cautious inferences in multi-label problems
- Conclusions and perspectives

Multi-label classification problem

👉 **The goal of multi-label problem:**

Given a training data: $\mathcal{D} = \{\mathbf{x}^i, \mathbf{y}^i\}_{i=0}^N \subseteq \mathbb{R}^p \times \mathcal{Y}$

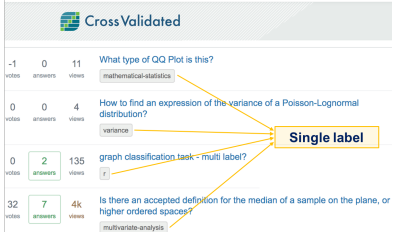
where: $\mathcal{Y} = \{0, 1\}^m, |\mathcal{Y}| = 2^m$

Learning a multi-label classification rule: $\varphi : \mathbb{R}^p \rightarrow \mathcal{Y}$

👉 **Example:**

Classical classification

$\mathcal{H} = \{\text{mathematical-statistics, variance, poisson-distribution, lognormal, qq-plot, ...}\}$

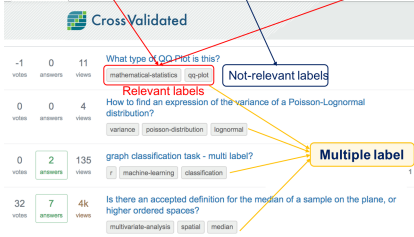


Single label



Multi-label classification

$\mathcal{H} = \{\text{mathematical-statistics, variance, poisson-distribution, lognormal, qq-plot, ...}\}$



Relevant labels

Not-relevant labels

Multiple label

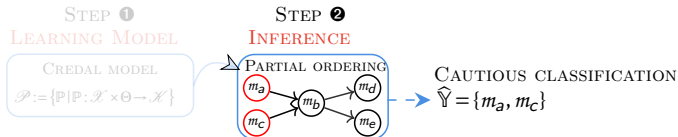
Existing results for precise and cautious inferences

- ✌ Inference in precise case difficult, but there are
 - ✓ Efficient algorithms for specific losses [...; [Dembczyński et al. 2012](#); [Waegeman et al. 2014](#)]
 - ✓ Several simplified learning model: Binary relevance, Classifier chains [[Read et al. 2019](#)], ...

- ✂ This issue is poorly explored in IP [[Destercke 2015](#); [Antonucci et al. 2017](#)], and even less in other cautious settings [[Nguyen et al. 2019](#); [Pillai et al. 2013](#)].

- ✌ Our contribution consists in providing:
 - 👉 More efficient, dedicated algorithm for the Hamming Loss.
 - 👉 First attempt to generalize the classifier chains to IP setting.

Cautious inferences in form of set-valued solutions



Problem setting and challenges:

✓ Step 1: The uncertainty model \mathcal{P} is known.

✗ Step 2: Under the maximality criterion and a generic loss matrix
 \implies the set-valued solutions require at worst $2^m(2^m - 1)$ computations.

✗ **Example:** $|\mathcal{Y}| = 10$, it needs to $2^{10}(2^{10} - 1) = 1047552$ computations.

$$(0, 0, \dots, 0) \succ_{\ell}^{\mathcal{P}} (0, \dots, 1, 1) ?$$

$$(0, 0, \dots, 0) \succ_{\ell}^{\mathcal{P}} (0, \dots, 1, 0) ?$$

\vdots
 \vdots
 \vdots
 \vdots



A set-valued solution

$$\hat{\mathcal{Y}}_{\ell, \mathcal{P}}^M = \begin{pmatrix} 1 & 1 & 0 & \dots & 0 \\ 0 & 1 & 1 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix}$$

👉 Can we obtain cautious predictions efficiently?

Overview

- Problem statements
- Imprecise Gaussian Discriminant Analysis
- Cautious inferences in multi-label problems
 - General case for the Hamming loss
 - Experimental results
 - Conclusion and additional results
- Conclusions and perspectives

General case for the Hamming case

Proposition 3 (Ceteris paribus comparison)

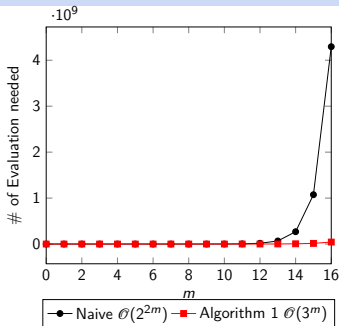
For a given set of indices $\mathcal{J} \subseteq \llbracket m \rrbracket$, let us consider an assignment $\mathbf{a}_{\mathcal{J}}$ and its complement $\bar{\mathbf{a}}_{\mathcal{J}}$. Then, for any two vectors $\mathbf{y}^1, \mathbf{y}^2$ such that $\mathbf{y}_{\mathcal{J}}^1 = \mathbf{a}_{\mathcal{J}}$, $\mathbf{y}_{\mathcal{J}}^2 = \bar{\mathbf{a}}_{\mathcal{J}}$ and $\mathbf{y}_{-\mathcal{J}}^1 = \mathbf{y}_{-\mathcal{J}}^2$, we have

$$\mathbf{y}^1 \succ_M \mathbf{y}^2 \iff \inf_{P \in \mathcal{P}} \sum_{i \in \mathcal{J}} P(Y_i = a_i) > \frac{|\mathcal{J}|}{2} \quad (1)$$

Prop. 3 amounts to focus on partial binary vector, e.g. $|\mathcal{Y}| = n + 3$, $\mathbf{a} = (0, 0, 0, *, \dots, *)$

$$(0, 0, 0, \underbrace{*, \dots, *}_{n \text{ labels}}) \succ_{\ell_H}^{\mathcal{P}} (1, 1, 1, \underbrace{*, \dots, *}_{n \text{ labels}})$$

1 comparison instead of 2^n .



Existing approximate results for Hamming loss

- The partial vector $\hat{\mathbf{y}}^* = (\hat{y}_1^*, \dots, \hat{y}_m^*) \in \mathcal{Y}^* = \{0, 1, *\}$

$$\hat{y}_j^* = \begin{cases} 1 & \text{if } \underline{P}_x(Y_j = 1) > 0.5 \\ 0 & \text{if } \overline{P}_x(Y_j = 1) < 0.5 \\ * & \text{if } 0.5 \in [\underline{P}_x(Y_j = 1), \overline{P}_x(Y_j = 1)] \end{cases}$$

is an outer-approximation of $\hat{\mathbf{Y}}_{\ell_H, \mathcal{P}}^M$ [Destercke 2015]

- Only requires to know imprecise marginal bounds \mathcal{P}_{Y_i} on each label.
- Note that not all partial multi-label predictions can be exactly represented as a partial vector

$$\begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix} \Rightarrow \text{cannot be represented in } \mathcal{Y}^* \quad \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = (*, *, 0) \in \mathcal{Y}^*$$

Exact $\hat{Y}_{\ell, \mathcal{P}}^M$ vs. \hat{y}^* outer-approximation inferences

Imprecision — small — medium — high

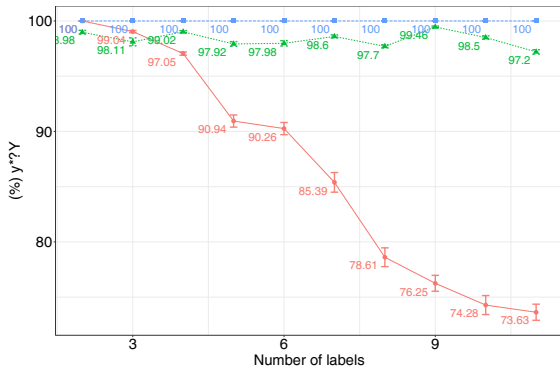


Figure: % of instances where $\hat{y}^* = \hat{Y}_{\ell, \mathcal{P}}^M$.

- ✓ The quality of \hat{y}^* decreases as the number of labels increases.
- ✓ The quality of \hat{y}^* seems to be the worst for moderate imprecision.

Exact $\hat{Y}_{\ell_H, \mathcal{P}}^M$ vs. \hat{y}^* outer-approximation inferences

Imprecision — small — medium — high

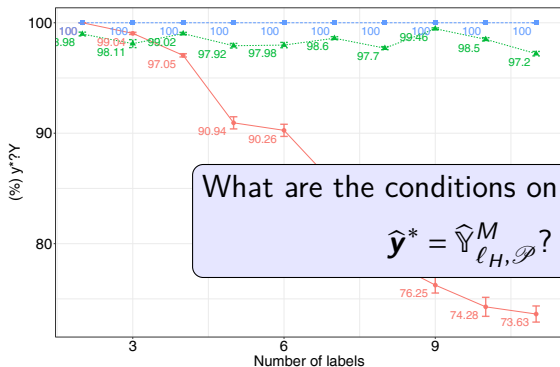


Figure: % of instances where

What are the conditions on \mathcal{P} ensuring

$$\hat{y}^* = \hat{Y}_{\ell_H, \mathcal{P}}^M?$$

- ✓ The quality of \hat{y}^* decreases as the number of labels increases.
- ✓ The quality of \hat{y}^* seems to be the worst for moderate imprecision.

Binary relevance and partial vectors

Under the assumption of label independence:

$$\mathcal{P}_{BR} := \left\{ \prod_{\{i|y_i=1\}} p_i \prod_{\{i|y_i=0\}} (1-p_i) \mid p_i \in [\underline{p}_i, \bar{p}_i] \right\}.$$

Proposition 4 (Domain restriction on \mathcal{P})

Given a probability set \mathcal{P}_{BR} and the Hamming loss, $\hat{Y}_{\ell_H, \mathcal{P}_{BR}}^M \in \mathcal{Y}^*$.

- ✓ $\hat{Y}_{\ell_H, \mathcal{P}_{BR}}^M$ can be represented as partial vector \mathcal{Y}^* .
- ✓ $\hat{Y}_{\ell_H, \mathcal{P}_{BR}}^M$ is equal to outer-approximation \hat{y}^* [Destercke 2015].
- ✓ The time complexity becomes linear on m , i.e. $\mathcal{O}(m)$!

Overview

- Problem statements
- Imprecise Gaussian Discriminant Analysis
- Cautious inferences in multi-label problems
 - General case for the Hamming loss
 - Experimental results
 - Conclusion and additional results
- Conclusions and perspectives

Dataset and experimental setting

Material/Imprecise Classifier/Metrics

- The data set issued from MULAN repository.

Data set	#Features	#Labels	#Instances	#Cardinality	#Density
yeast	103	14	2417	4.23	0.30

- Naive credal classifier (NCC) [Zaffalon 2002]
- Metric evaluations: (Q denotes the set of predicted label s.t. $\hat{y}_i = 1$ or $\hat{y}_i = 0$)

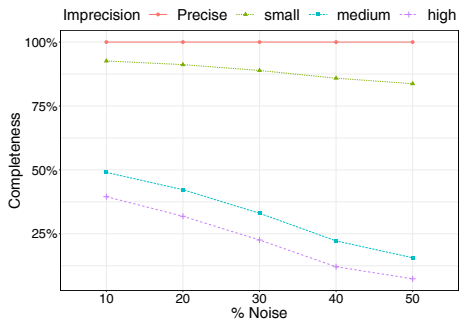
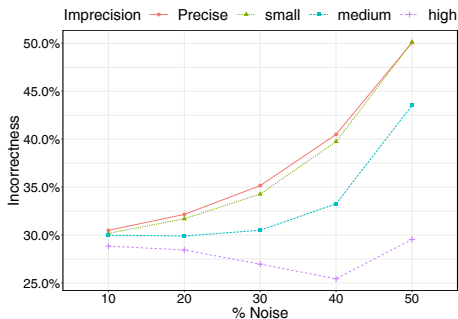
$$IC(\hat{Y}, \mathbf{y}) = \frac{1}{|Q|} \sum_{\hat{y}_i \in Q} 1_{(\hat{y}_i \neq y_i)} \quad \text{and} \quad CP(\hat{Y}, \mathbf{y}) = \frac{|Q|}{m}$$

Reversing Noise (“adversarial” perturbations)

We reverse the current value of a selected label j and an instance i , i.e.

$Y_{j,i} = 1 \rightarrow Y_{j,i} = 0$ or $Y_{j,i} = 0 \rightarrow Y_{j,i} = 1$. For example:

Features					Noise-Reversing		
X_1	X_2	X_3	X_4	X_5	Y_1	Y_2	Y_3
107.1	25	Blue	60	1	1	0→1	0
-50	10	Red	40	0	1	0	1→0
200.6	30	Blue	58	1	0→1	0	0
...



Evolution of the incorrectness (left) and the completeness (right) in average (%) for each level of imprecision (a curve for each one), with respect to the % of noise.

- 👉 Cautious inferences provide some level of protection by abstaining on those hard-to-predict instances where adversarial noise was introduced.
- 👉 Including some imprecision limits the increase in incorrectness, but it decreases the completeness.

Overview

- Problem statements
- Imprecise Gaussian Discriminant Analysis
- Cautious inferences in multi-label problems
 - General case for the Hamming loss
 - Experimental results
 - Conclusion and additional results
- Conclusions and perspectives

Conclusion and additional results for the Hamming Loss

- Given a probability set \mathcal{P}_{BR} and the Hamming loss ℓ_H , we proved some additional relations

$$\Gamma\text{-minimax} \iff \Gamma\text{-minimin}$$

$$\Gamma\text{-minimax} \implies \mathbf{E}\text{-admissibility}$$

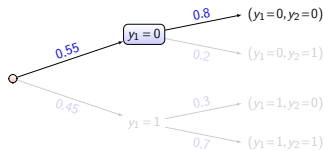
- When considering sets of distributions and cautious inferences, **it is not sufficient to consider marginal probabilities to get exact set-valued predictions**, as opposed to the case of precise distributions.
- We now have a better knowledge of computational issues for the Hamming loss.

Imprecise Binary relevance allows for efficiency, but it does not integrate the dependence between labels. So, **how can we tackle this issue?**

Imprecise classifier-chains (ICC) approach

What?

Multi-label chaining using a **set of probability models** instead of a **single probability model**.

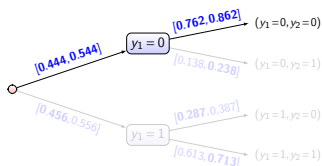


Chaining with precise probabilistic models

$$\mathbb{P}$$

$$P(Y_j | Y_1, \dots, Y_{j-1}, X)$$

widely studied!



Chaining with imprecise probabilistic models

$$\mathcal{P}$$

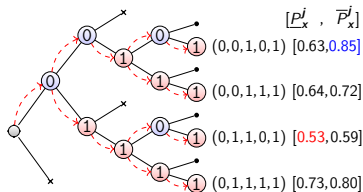
$$[\underline{P}(Y_j | Y_1, \dots, Y_{j-1}, X), \overline{P}(Y_j | Y_1, \dots, Y_{j-1}, X)]$$

how can we do it?

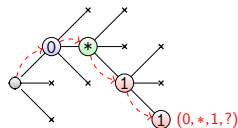
Imprecise classifier-chains (ICC) approach

How can we do it?

- 👉 We propose 2 strategies when having probability bounds in the chain.
- 👉 They differ by how we treat labels for which we abstain ($0.5 \in [P, \bar{P}]$).



(a) Imprecise Branching



(b) Marginalization (* = {0, 1})

Exploring the tree

Consider all possible paths in the chaining on which we abstain.

Pruning the tree

Take out or ignore labels on which we abstain.

Conclusion Imprecise classifier-chains (ICC) approach

ICC using Naive credal classifier

- ✌ Inference complexity of the **IMPRECISE BRANCHING** strategy using NCC is between $\mathcal{O}(m^2)$ and $\mathcal{O}(m)$.
- ✌ Inference complexity of the **MARGINALIZATION** strategy using NCC is $\mathcal{O}(m)$.

Experimental results

- ✌ Good balance between abstained labels and performance.
- ✌ Our proposal overcomes those precise ones in noisy setting.

Open issues

- ? How to come up with general but efficient optimisation methods to solve the strategies (IB) and (MAR).
- ? Investigating the performance of our proposed strategies on other imprecise classifier (e.g., continuous classifier).
- ? Fully investigating issue of label ordering.

Overview

- Problem statements
- Imprecise Gaussian Discriminant Analysis
- Cautious inferences in multi-label problems
- Conclusions and perspectives

Conclusions and Perspectives

- ✓ Including an imprecise component in supervised classification problems in a smart way allow for reasonable, limited cautiousness while offering a good protection on noisy, ambiguous, ill-informed instances.
- ✗ Describing our uncertainty by a set of probabilities distributions over combinatorial domains leads to difficult optimisation problems, that largely remain to be solved.



References I



Zaffalon, Marco (2002). "The naive credal classifier". In: *Journal of statistical planning and inference* 105.1, pp. 5–21.



Herbei, Radu and Marten H Wegkamp (2006). "Classification with reject option". In: *Canadian Journal of Statistics* 34.4, pp. 709–721.



Troffaes, Matthias CM (2007). "Decision making under uncertainty using imprecise probabilities". In: *International Journal of Approximate Reasoning* 45.1, pp. 17–29.



De Cooman, Gert and Filip Hermans (2008). "Imprecise probability trees: Bridging two theories of imprecise probability". In: *Artificial Intelligence* 172.11, pp. 1400–1427.



Frank, A. and A. Asuncion (2010). *UCI Machine Learning Repository*. URL: <http://archive.ics.uci.edu/ml>.



Read, Jesse et al. (2011). "Classifier chains for multi-label classification". In: *Machine learning* 85.3, p. 333.



Dembczynski, Krzysztof, Willem Waegeman, and Eyke Hüllermeier (2012). "An Analysis of Chaining in Multi-Label Classification.". In: *ECAI*, pp. 294–299.



Dembczyński, Krzysztof et al. (2012). "On label dependence and loss minimization in multi-label classification". In: *Machine Learning* 88.1-2, pp. 5–45.



Zaffalon, Marco, Giorgio Corani, and Denis Mauá (2012). "Evaluating credal classifiers by utility-discounted predictive accuracy". In: *International Journal of Approximate Reasoning* 53.8, pp. 1282–1301.

References II



Pillai, Ignazio, Giorgio Fumera, and Fabio Roli (2013). "Multi-label classification with a reject option". In: *Pattern Recognition* 46.8, pp. 2256–2266.



Benavoli, Alessio and Marco Zaffalon (2014). "Prior near ignorance for inferences in the k-parameter exponential family". In: *Statistics* 49.5, pp. 1104–1140.



Mantas, Carlos J and Joaquin Abellan (2014). "Credal-C4. 5: Decision tree based on imprecise probabilities to classify noisy data". In: *Expert Systems with Applications* 41.10, pp. 4625–4637.



Waegeman, Willem et al. (2014). "On the bayes-optimality of f-measure maximizers". In: *Journal of Machine Learning Research* 15, pp. 3333–3388.



Destercke, Sébastien (2015). "Multilabel predictions with sets of probabilities: the Hamming and ranking loss cases". In: *Pattern Recognition* 48.11, pp. 3757–3765.



Antonucci, Alessandro and Giorgio Corani (2017). "The multilabel naive credal classifier". In: *International Journal of Approximate Reasoning* 83, pp. 320–336.



Mauá, Denis D et al. (2017). "Credal sum-product networks". In: *Proceedings of the Tenth International Symposium on Imprecise Probability: Theories and Applications*, pp. 205–216.



Vovk, Vladimir and Claus Bendtsen (2018). "Conformal predictive decision making". In: *Conformal and Probabilistic Prediction and Applications*, pp. 52–62.



Nguyen, Vu-Linh and Eyke Hüllermeier (2019). "Reliable Multi-label Classification: Prediction with Partial Abstention". In: *CoRR* abs/1904.09235. arXiv: 1904.09235. URL: <http://arxiv.org/abs/1904.09235>.

References III



Read, Jesse et al. (2019). "Classifier Chains: A Review and Perspectives". In: *arXiv preprint arXiv:1912.13405*.