Distributionally Robust, Skeptical Binary Inferences in Multi-label Problems

YC. Carranza-Alarcón, Sébastien Destercke

{yonatan-carlos.carranza-alarcon, sebastien.destercke}@hds.utc.fr

(1) Problem statement

Setting: Let us consider that we have:

① a training data set $\mathscr{D}$,
② an uncertainty model $\mathscr{P}$ fitted to $\mathscr{D}$,
③ the Hamming loss ℓ_H and
④ the maximality criterion $\succ^{\mathscr{P}}_{\ell_H}$.

$\Longrightarrow$

Goals:
① Making skeptical decisions $\mathbb{Y} \subseteq \mathscr{Y}$.
② Reducing the time complexity of the inference-step, i.e. $\mathscr{O}(2^{2m})$.

X_1	X_2	X_3	X_4	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6
0.34	0	34	174	1	0	1	0	1	1
0.54	1	4	434	1	1	1	0	0	1
1.44	0	14	574	0	0	0	0	1	1
3.44	1	45	584	1	1	1	0	0	0
4.94	1	6	884	0	1	1	0	1	0

Table 1: Example of a data set $\mathscr{D}$

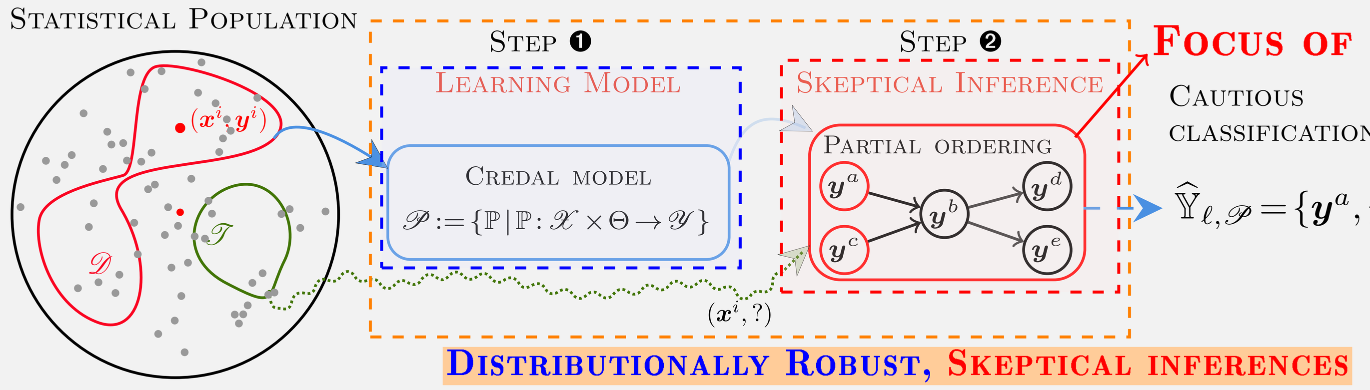
How can we obtain Skeptical Binary Inferences?:

Imprecise Supervised Classification Approach

Given an uncertainty model defined as a set of distributions $\mathscr{P}$ and fitted to the training data set:

$$\mathscr{D} = \{\boldsymbol{x}^i, \boldsymbol{y}^i\}_{i=0}^N \subseteq \mathbb{R}^p \times \mathscr{Y},\ \mathscr{Y} = \{0,1\}^m.$$

We want make a skeptical inference for a new observation $(\boldsymbol{x}^i, ?)$. (see fig. $\rightarrow$).



(2) Skeptic Inference for Hamming loss

Definition (Maximality) [4, 3]

Maximality consists in returning the maximal, non-dominated elements of the partial order $\succ^{\mathscr{P}}_{\ell}$ such that $\boldsymbol{y} \succ^{\mathscr{P}}_{\ell} \boldsymbol{y}'$ if

$$\underline{\mathbb{E}}\left[\ell(\boldsymbol{y}',\cdot) - \ell(\boldsymbol{y},\cdot)\right] := \inf_{P \in \mathscr{P}} \mathbb{E}_P\left[\ell(\boldsymbol{y}',\cdot) - \ell(\boldsymbol{y},\cdot)\right] > 0, \tag{1}$$

that is if exchanging $\boldsymbol{y}'$ for $\boldsymbol{y}$ is guaranteed to give a positive expected loss. The maximality rule returns the prediction set

$$\hat{\mathbb{Y}}^M_{\ell,\mathscr{P}} = \{\boldsymbol{y} \in \mathcal{Y} \mid \nexists \boldsymbol{y}' \in \mathcal{Y} \text{ s.t. } \boldsymbol{y}' \succ^{\mathscr{P}}_{\ell} \boldsymbol{y}\}. \tag{2}$$

Definition (Hamming Loss)

The Hamming loss amounts to compute the Hamming distance between the ground truth $\boldsymbol{y}$ and a prediction $\hat{\boldsymbol{y}}$, that is

$$\ell_H(\hat{\boldsymbol{y}}, \boldsymbol{y}) = \sum_{i=1}^m 1_{(\hat{y}_i \neq y_i)} = |\mathcal{I}_{\hat{\boldsymbol{y}} \neq \boldsymbol{y}}| \tag{3}$$

Lemma 1

In the case of Hamming loss and given $\boldsymbol{y}^1, \boldsymbol{y}^2$, we have that: $\underline{\mathbb{E}}\left[\ell_H(\boldsymbol{y}^2,\cdot) - \ell_H(\boldsymbol{y}^1,\cdot)\right] = \sum_{i=1}^m P(Y_i = y_i^1) - P(Y_i = y_i^2)$

(3) General case

Proposition 3 (Ceteris paribus comparison)

For a given set of indices $\mathscr{I} \subseteq [\![m]\!]$, let us consider an assignment $\boldsymbol{a}_{\mathscr{I}}$ and its complement $\overline{\boldsymbol{a}}_{\mathscr{I}}$. Then, for any two vectors $\boldsymbol{y}^1, \boldsymbol{y}^2$ such that $\boldsymbol{y}^1_{\mathscr{I}} = \boldsymbol{a}_{\mathscr{I}}$, $\boldsymbol{y}^2_{\mathscr{I}} = \overline{\boldsymbol{a}}_{\mathscr{I}}$ and $\boldsymbol{y}^1_{-\mathscr{I}} = \boldsymbol{y}^2_{-\mathscr{I}}$, we have

$$\boldsymbol{y}^1 \succ_M \boldsymbol{y}^2 \iff \inf_{P \in \mathscr{P}} \sum_{i \in \mathscr{I}} P(Y_i = a_i) > \frac{|\mathscr{I}|}{2} \tag{4}$$

Example (Imprecise tree model $\mathscr{P}$)

Let us consider the imprecise tree model in the bottom right. Applying Prop. 3, we have

$$\underline{\mathbb{E}}[\ell_H((1,*),\cdot)] = 0.444 > 0.5 \implies (0,*) \not\succ_M (1,*),$$
$$\underline{\mathbb{E}}[\ell_H((0,*),\cdot)] = 0.456 > 0.5 \implies (1,*) \not\succ_M (0,*),$$
$$\underline{\mathbb{E}}[\ell_H((*,1),\cdot)] = 0.498 > 0.5 \implies (*,0) \not\succ_M (*,1),$$
$$\underline{\mathbb{E}}[\ell_H((*,0),\cdot)] = 0.354 > 0.5 \implies (*,1) \not\succ_M (*,0),$$
$$\underline{\mathbb{E}}[\ell_H((1,1),\cdot)] = 0.942 > 1.0 \implies (0,0) \not\succ_M (1,1),$$
$$\underline{\mathbb{E}}[\ell_H((1,0),\cdot)] = 0.846 > 1.0 \implies (0,1) \not\succ_M (1,0),$$
$$\underline{\mathbb{E}}[\ell_H((0,1),\cdot)] = 1.001 > 1.0 \implies (\mathbf{1},\mathbf{0}) \succ_M (\mathbf{0},\mathbf{1}),$$
$$\underline{\mathbb{E}}[\ell_H((0,0),\cdot)] = 0.810 > 1.0 \implies (1,1) \not\succ_M (0,0).$$

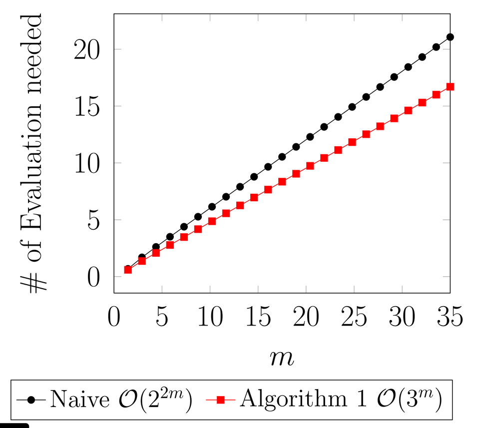
We get $3^2 - 1 = 8$ comparisons and skeptical inference is the set

$$\hat{\mathbb{Y}}^M_{\ell_H,\mathscr{P}} = \{(1,0),(0,0),(1,1)\}$$

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

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

1 comparaison instead of 2^n.



Remark

Let us define a partial-binary vector space as

$$\mathfrak{Y} = \{0,1,*\}^m$$

and not every solution $\hat{\mathbb{Y}}^M_{\ell_H,\mathscr{P}}$ can not be represented as a partial-binary vector $\mathfrak{Y}$.

How can we obtain $\hat{\mathbb{Y}}^M_{\ell_H,\mathscr{P}^*} \in \mathfrak{Y}$?

(4) Binary relevance and partial vectors

Under the assumption of label independence, i.e. $Y_1 \perp Y_2 \cdots \perp Y_m$:

$$\mathscr{P}_{BR} := \left\{ \prod_{\{i|y_i=1\}} p_i \prod_{\{i|y_i=0\}} (1-p_i) \;\middle|\; p_i \in [\underline{p}_i, \overline{p}_i],\ \ p_i := P(Y_i = y_i|X = x) \right\}.$$

Proposition 8 (Domain restriction on $\mathscr{P}$)

Given a probability set $\mathscr{P}_{BR}$ and the Hamming loss ℓ_H, the set $\hat{\mathbb{Y}}^M_{\ell_H,\mathscr{P}_{BR}} \in \mathfrak{Y}$.

✓ $\hat{\mathbb{Y}}^M_{\ell_H,\mathscr{P}_{BR}}$ can be represented as partial vector $\mathfrak{Y}$.
✓ $\hat{\mathbb{Y}}^M_{\ell_H,\mathscr{P}_{BR}}$ is equal to known outer-approximation [1].
✓ **The time complexity of skeptical inference becomes linear on m, i.e. $\mathscr{O}(m)$!**

(5) Experiments and results

Setting: The imprecise tree model (see App. 1) is here used to represent our credal set $\mathscr{P}$ (but our results hold for any credal).

Exact vs approximate skeptic inference

Goal: Evaluate how accurate the outer-approximation [1] $\hat{\boldsymbol{y}}^*_{\ell_H,\mathscr{P}}$ is in comparison to our exact estimation of the set $\hat{\mathbb{Y}}^M_{\ell_H,\mathscr{P}}$.

Setting:
☞ We simulate imprecise binary trees $\mathscr{P}$ using an imprecise parameter ϵ.
☞ Metric evaluation (how large is $\hat{\boldsymbol{y}}^*$?): $d^\epsilon(\hat{\boldsymbol{y}}^*, \hat{\mathbb{Y}}) = |\hat{\boldsymbol{y}}^*_{\ell_H,\mathscr{P}}| - |\hat{\mathbb{Y}}^M_{\ell_H,\mathscr{P}}|$..

Results: The quality $\hat{\boldsymbol{y}}^*_{\ell_H,\mathscr{P}}$ decreases as the number of labels increases and seems to be the worst for moderate imprecision.

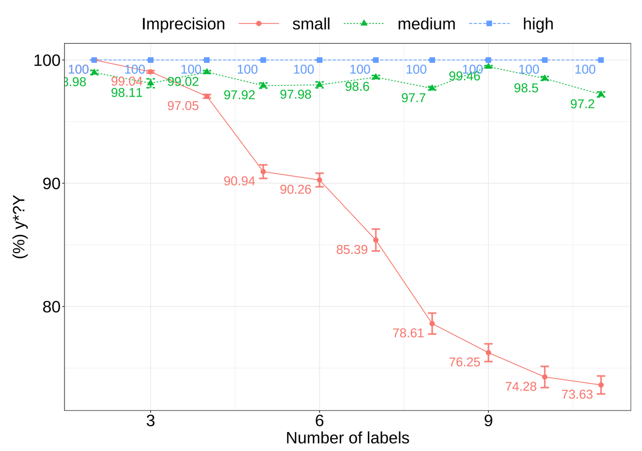


Figure 1: % of instances where $\hat{\boldsymbol{y}}^* = \hat{\mathbb{Y}}^M_{\ell_H,\mathscr{P}}$.

Skeptic inference with Binary relevance

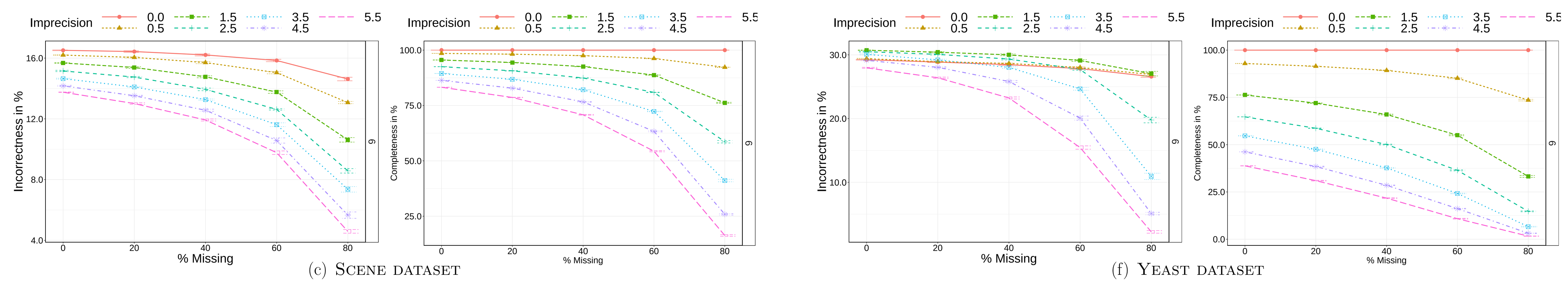
Goal: We investigate what happens when some labels are missing.

Results: The precise model ($s = 0.$) is not really affected by randomly missing labels, whereas our proposal becomes more cautious as the number of missing labels increases.

Setting: ☞ Incorrectness and Completness (Q denotes the set of predicted label s.t. $\hat{y}_i = 1$ or $\hat{y}_i = 0$).

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

☞ Missing labels pick at random a percentage of $\{20, 40, 60, 80\}$.



(c) Scene dataset (f) Yeast dataset

(6) Conclusion and Perspectives

☞ Works done since submission of ISIPTA paper:

- We provide efficient procedures to solve the maximality criterion under **Hamming loss** and **generic probability sets**.
- 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 provide new implications (an implication $A \rightarrow B$ means that $A \subset B$) for **the different decision criteria**, namely **M**aximality, **E**-admissibility, **Γ**-minimax, **Γ**-minimin and **I**nterval **D**ominance, when we use **the restricted probability set $\mathscr{P}_{BR}$** and the **Hamming loss ℓ_H** (see Fig.2)

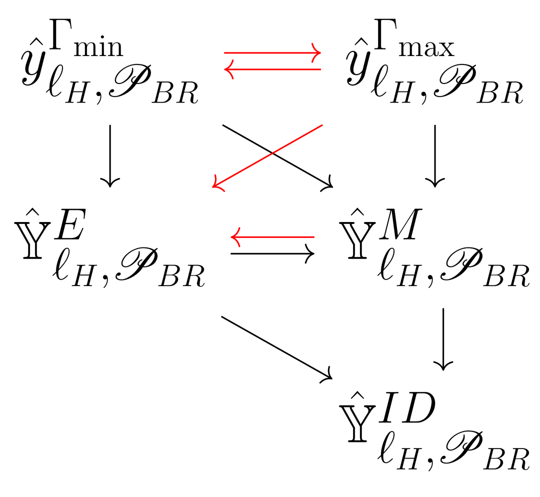


Figure 2: Decision relation under a $\mathscr{P}_{BR}$ and a ℓ_H. In **red arrow**, the new implications.

☞ What remains to finish (or in progress):

✘ Compare our proposal against those **rejecting** and **abstaining** approaches.
✘ Solve the maximality criterion using other loss functions, e.g.; **Ranking loss and F-measure.**

Appendix 1 - Example - Imprecise probabilistic tree and lower expected loss

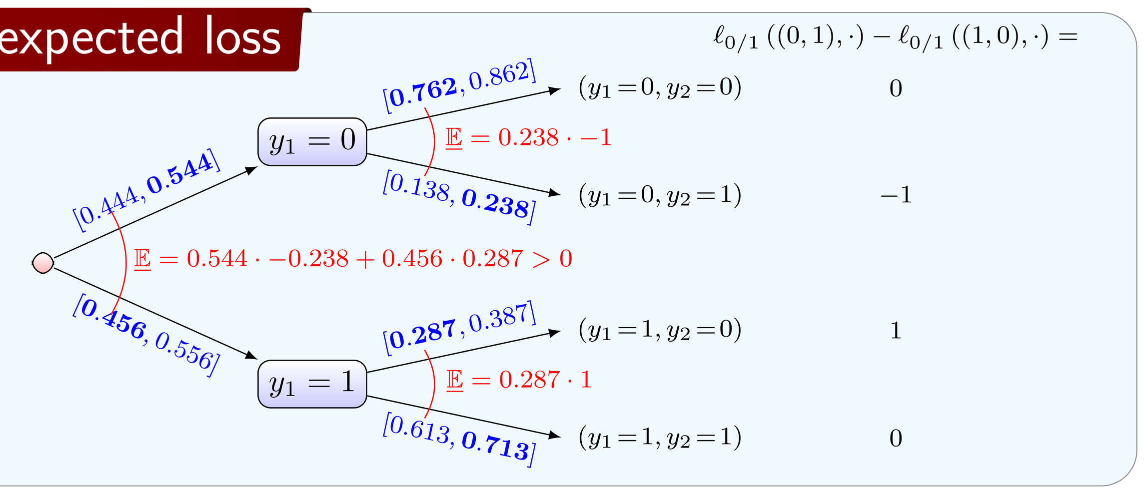
Inference in binary trees [2]: For computing the infimum expectation given an assignment $\boldsymbol{a}_{\mathcal{I}}$, we use the law of iterated lower expectation [2]:

$$\underline{\mathbb{E}}_{\mathbf{Y}}\left[\ell_H(\cdot,\overline{a}_{\mathcal{I}})\right] = \underline{\mathbb{E}}_{Y_1}\left[\underline{\mathbb{E}}_{Y_2}\left[\dots\underline{\mathbb{E}}_{Y_m}\left[\ell_H(\cdot,\overline{a}_{\mathcal{I}})\middle|Y_{\mathcal{I}_{[\![m-1]\!]}}\right]\dots\right]\right].$$

Proposition 4

For a given set $\mathcal{I}$ of indices, an assignment $\boldsymbol{a}_{\mathcal{I}}$ and its complement $\overline{\boldsymbol{a}}_{\mathcal{I}}$. We have

$$\inf_{P \in \mathscr{P}} \sum_{i \in \mathcal{I}} P(Y_i = a_i) = \underline{\mathbb{E}}[\ell^*_H(\overline{\boldsymbol{a}}_{\mathcal{I}},\cdot)] \tag{5}$$



Acknowledgments

This work was carried out in the framework of the Labex MS2T, funded by the French Government, through the National Agency for Research (Reference ANR-11-IDEX-0004-02).

References

[1] Sébastien Destercke. Multilabel predictions with sets of probabilities: the hamming and ranking loss cases. *Pattern Recognition*, 48(11):3757–3765, 2015.
[2] Filip Hermans, Gert Quaeghebeur, et al. Imprecise markov chains and their limit behavior. *Probability in the Engineering and Informational Sciences*, 23(4):597–635, 2009.
[3] Matthias CM Troffaes. Decision making under uncertainty using imprecise probabilities. *International Journal of Approximate Reasoning*, 45(1):17–29, 2007.
[4] P. Walley. *Statistical reasoning with imprecise Probabilities.* Chapman and Hall, 1991.