Learning with uncertain data: challenges and opportunities

Sébastien Destercke
CNRS senior researcher

Heudiasyc, CNRS Compiegne, France

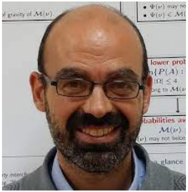
COPA 2022

Talk benefited from collaborations and discussions with

Eyke Hüllermeier

Mylène Masson

Benjamin Quost



Enrique Miranda

Ignacio Montes

Didier Dubois

Thierry Denœux

Ines Couso

Menu

1. Appetizer: on the nature and origins of uncertain data
   - On the nature of data uncertainty
   - On the modelling of data uncertainty: basic models
   - On the modelling of data uncertainty: more complex models
2. Main course: challenges of learning with data uncertainty
   - Challenges of learning under uncertain data
   - Uncertain data and conformal prediction
3. Dessert: when data uncertainty can be useful

Some examples

"Sport car" $\rightarrow$

$Y = \{Lion, Jaguar, Cat, \ldots\}$ $\quad \{4, 9\}$ $\quad \{Porsche, Ferrari, \ldots\}$

$\uparrow$ Ambiguity

$\uparrow$ Ambiguity

$\uparrow$ Coarse data

Menu

1. Appetizer: on the nature and origins of uncertain data
   - On the nature of data uncertainty
   - On the modelling of data uncertainty: basic models
   - On the modelling of data uncertainty: more complex models
2. Main course: challenges of learning with data uncertainty
   - Challenges of learning under uncertain data
   - Uncertain data and conformal prediction
3. Dessert: when data uncertainty can be useful

Kinds of uncertainties

- Epistemic vs Aleatoric
  - Epistemic: due to lack of knowledge
  - Aleatoric: due to inherent randomness
- Statistical vs non-statistical
  - Statistical: concerns a population (over time/space)
  - Non-statistical: concerns an individual
- Reducible vs non-reducible
  - Reducible: further information allow to reduce uncertainty
  - Irreducible: no more information will come

Kinds of uncertainties

Data uncertainty is mostly

- **Epistemic** vs Aleatoric
- Statistical vs **non-statistical**
- **Reducible vs non-reducible**

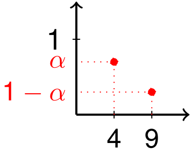
- Epistemic: a datum value is not random
- Non-statistical: uncertainty concerns one individual observation
- Reducible or irreducible: whether or not one has access to better measurement/more expertise

Menu

Probabilistic modelling (a.k.a. soft labels) [18]

Information: rather a 4 than a 9

Uncertainty model: $p(4) = 0.75, p(9) = 0.25$



Why (not) probabilities?

Some pros:

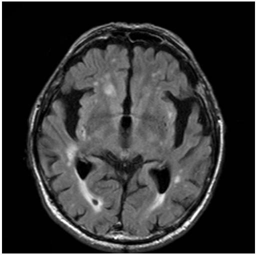
- By far the most used uncertainty model $\rightarrow$ lots of people and tools
- Naturally fits with classical loss function (cross entropy the first)

Some cons:

- Not clear **at all** that data uncertainty has a probabilistic nature
- Important issues when modelling incompleteness/imprecision
- Limit expressiveness/possibilities compared to other theories

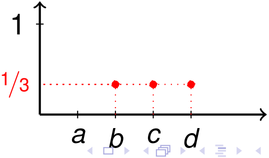
Issue in representing incompleteness

Requesting a doctor to classify disease severity degree



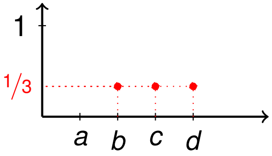
- No disease: *a*
- 3 degrees of severity
- labels $\mathcal{Y} = \{a, b, c, d\}$

Doctor opinion: disease present, severity difficult to assess

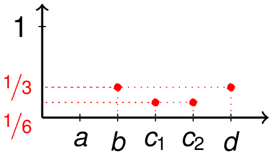


Issue in representing incompleteness

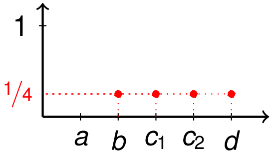
For various reasons, degree c is divided into c_1, c_2. What should



become? Two possibilities are



Coherent with initial model

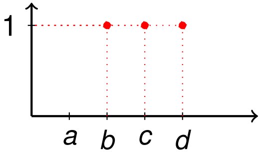


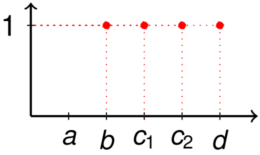
Uniform on $\{b, c_1, c_2, d\}$

No way to be ignorant on $\{b, c, d\}$ and $\{b, c_1, c_2, d\}$ simultaneously!

Another model: sets (a.k.a. partial labels) [5]

Available information: set $E = \{b, c, d\}$ (or $E = \{b, c_1, c_2, d\}$)





Derived uncertainty measures

Two binary measures ($\underline{P}, \overline{P} \in \{0, 1\}$) for three possible situations:

- $\underline{P}$ indicates necessarily true, $\overline{P}$ indicates possibly true
- $E \subseteq A$: $y \in A$ certainly true $\rightarrow \underline{P} = \overline{P} = 1$. Ex: $A = \{a, b, c, d\}$
- $E \cap A, E \cap A^c \neq \emptyset$: $y \in A$ possibly true $\rightarrow \underline{P} = 0, \overline{P} = 1$. Ex: $A = \{b, c\}$
- $E \cap A = \emptyset$: $y \in A$ certainly false $\rightarrow \underline{P} = \overline{P} = 0$. Ex: $A = \{a\}$

Why (not) sets

Some pros:

- Very simple uncertainty model
- Naturally models epistemic uncertainty/incompleteness

Some cons:

- Loss adaptation requires some thinking (more on that later)
- Limited expressiveness (yes/no model)

Limited expressiveness

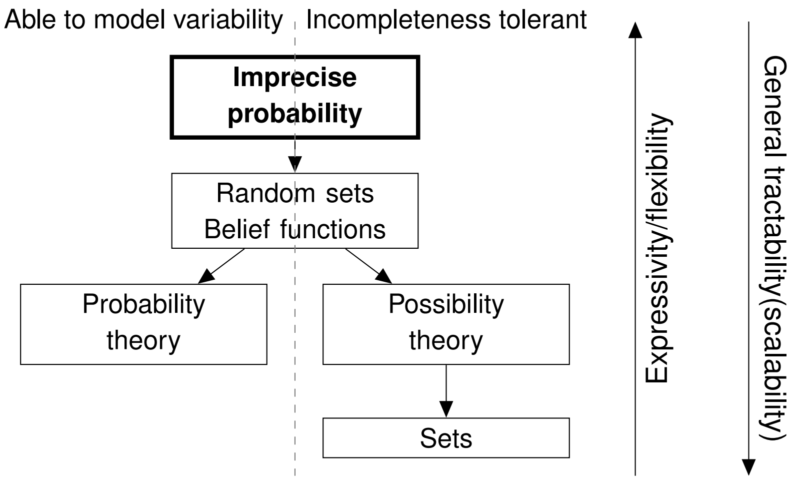
Remember this?

Information: rather a 4 than a 9

No way to model it with sets, probabilistic model reasonably satisfactory (but still requires an arbitrary choice of $p(4) \geq p(9)$)

What else can we do? Generalize them to richer frameworks.

A not completely accurate but useful picture [10]



Menu

1. Appetizer: on the nature and origins of uncertain data
   - On the nature of data uncertainty
   - On the modelling of data uncertainty: basic models
   - On the modelling of data uncertainty: more complex models
2. Main course: challenges of learning with data uncertainty
   - Challenges of learning under uncertain data
   - Uncertain data and conformal prediction
3. Dessert: when data uncertainty can be useful

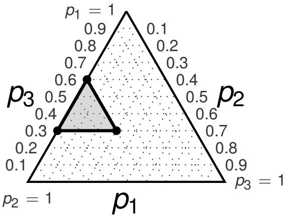
Convex probability sets [10]

Basic tool

A subset $\mathcal{P} \subseteq \Delta_{\mathcal{Y}}$, often convex, of probabilities

- $\overline{P}(A) = \sup_{P \in \mathcal{P}} P(A)$
- $\underline{P}(A) = \inf_{P \in \mathcal{P}} P(A) = 1 - \overline{P}(A^c)$ (duality[a])

[a] Having one of the two on all events is enough



- Probabilities p: one-element set
- Sets: $E \rightarrow$ all distributions with support included in E

Revisiting our example

Information: rather a 4 than a 9

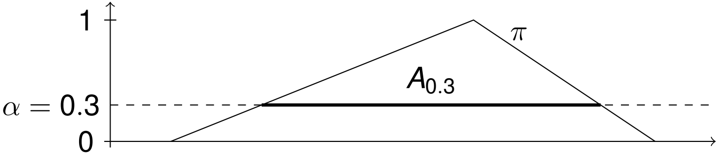
Modelling by $\mathcal{P} = \{P : P(4) \geq P(9)\}$

$$\underline{P}(9) = 0 \leq P(9) \leq \overline{P}(9) = 0.5,$$
$$\underline{P}(4) = 0.5 \leq P(4) \leq \overline{P}(4) = 1$$

Set $\mathcal{P}$ in practice: possibility distributions [1, 13]

A mapping π inducing $\mathcal{P}_\pi$ and

$$\overline{P}(A) = \sup_{x \in A} \pi(x); \qquad \underline{P}(A) = 1 - \overline{P}(A^c)$$



$$\underline{P}_\pi(A_{0.3}) = 0.7$$

Why is it interesting?

Confidence interval interpretation:

$$P \in \mathcal{P}_\pi \text{ iff } \forall \alpha \in [0,1] P(\{y : \pi(y) \geq \alpha\}) \geq 1 - \alpha$$

Linking possibilities and conformal approaches

One can use the equivalence [4, 15]:

$$\text{conformal p-value}$$

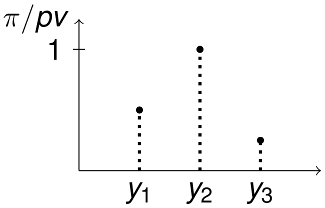
$$pv(y)$$

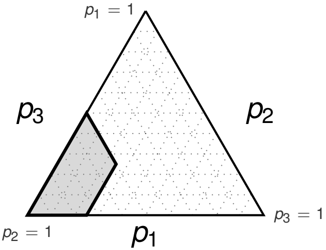
$$\equiv$$

$$\text{possibility degree}$$

$$\pi(y).$$

Conformal predictors can be seen as inducing a possibility distributions





Another interesting practical $\mathcal{P}$ [7]

Imprecise Prob. Assignments

For any $y \in \mathcal{Y}$, interval

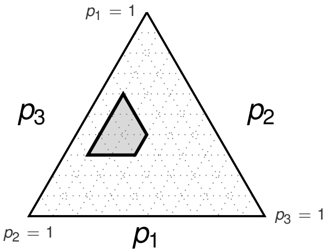
$$[\underline{p}(y), \overline{p}(y)]$$

inducing a set $\mathcal{P}_{[\underline{p},\overline{p}]}$.

Why look at it?

- Direct connection with Venn predictors
- Interesting mathematical properties

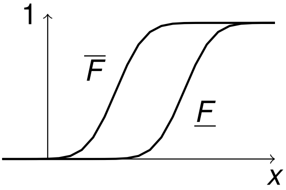
- $p(y_3) \in [0.1, 0.3]$
- $p(y_1) \in [0.3, 0.6]$
- $p(y_2) \in [0.3, 0.6]$



A last kind of model [9]

Two cumulative distributions $[\underline{F}, \overline{F}]$ bounding an unknown one

$$\mathcal{P}_{[\underline{F},\overline{F}]} = \{P : \forall x, \underline{F} \leq F_P(x) \leq \overline{F}\}$$



Why may it be interesting?

- Akin to predictive distributions $\rightarrow$ useful to extend them if needed?
- Nice mathematical properties to handle them

Wrap-up on apetizers

Data uncertainty is essentially non-statistical (in my view)

Probabilistic modelling not the only option, and even questionable to some extent.

Imprecise probability as a richer framework

Credal/imprecise probabilistic approaches embed sets and probabilities in a very expressive setting.

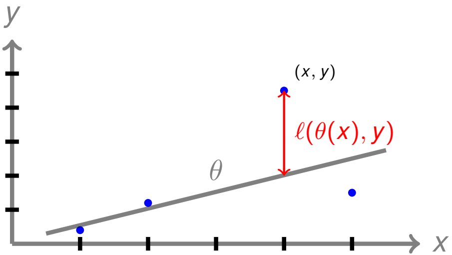
Strong links with conformal approaches

Conformal and Venn predictors naturally map to such representations. Some works even argue that calibration cannot be achieved without using such representations [4, 11].

Menu

How good is a model?

- Learning $\theta : \mathcal{X} \to \mathcal{Y}$ from observations (x_i, y_i)
- $\ell(\theta(x) = \hat{y}, y)$: loss of predicting $\hat{y}$ using θ if y is observed.



Loss and selection

- $\ell(\theta(x) = \hat{y}, y)$: loss incurred by predicting $\hat{y}$ if y is observed.
- A model θ will produce predictions $\theta(x)$, and its global loss on observed training data (x_i, y_i) will be evaluated as[1]

$$R_{emp}(\theta) = \sum_{i=1}^{N} \ell(\theta(x_i), y_i)$$

  possibly regularizing to avoid overfitting (not this talk topic)
- The optimal model is

$$\theta^* = \arg\min_{\theta \in \Theta} R_{emp}(\theta),$$

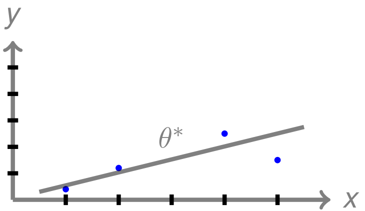
  the one with lowest possible average loss

[1] Used as approximation of $R(\theta) = \int_{\mathcal{X}\times\mathcal{Y}} \ell(\theta(x), y) dP(x, y)$.

Prototypical cases

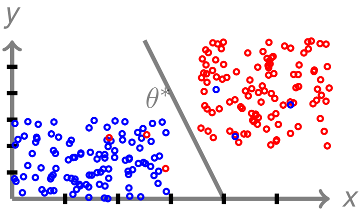
Regression

$$L(y, \hat{y}) = (y - \hat{y})^2$$



Classification (binary log reg)

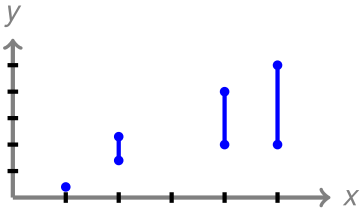
$$L(y,p) = \begin{cases} -\log(p) & \text{if } y = 1 \\ -\log(1-p) & \text{if } y = 0 \end{cases}$$



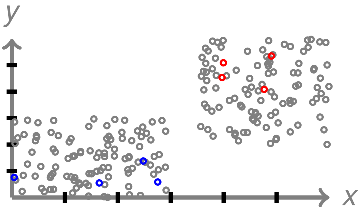
Menu

The imprecise setting illustrated

Regression



Classification (binary log reg)



How to define h_{θ^*}?

Induction with imprecise data

- We observe possibly imprecise input/output (X, Y) containing the truth (one $(x, y) \in (X, Y)$ are true, unobserved values)
- Losses[2] become set-valued [6]:

$$\ell(\theta(X), Y) = \{\ell(\theta(X), Y) | y \in Y, x \in X\}$$

- Previous induction principles are no longer well-defined
- What if we still want to get one model?

[2] And likelihoods/posteriors alike

Various options

- If we know the “imprecisiation” process $P_{obs}((X, Y)|(x, y))$, no theoretical problem $\rightarrow$ “merely” a computational one
- If not, common approaches are to redefine a precise criterion:
  - Optimistic (Maximax/Minimin) approach [14, 5]:

$$\ell_{opt}(\theta(x), Y) = \min\{\ell(\theta(x), Y)|y \in Y\}$$

  - Pessimistic (Maximin/Minimax) approach [12]:

$$\ell_{pes}(\theta(x), Y) = \max\{\ell(\theta(x), Y)|y \in Y\}$$

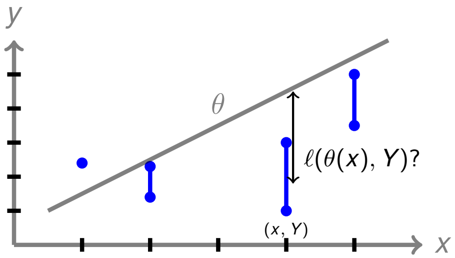
  - "EM-like" or averaging/weighting approaches[3]

$$\ell_w(\theta(x), Y) = \sum_{y \in Y} w_y \ell(\theta(x), y),$$

[3] With likelihood $\sim L_{av}(\theta|(x, Y)) = P((x, Y)|\theta)$ [8]

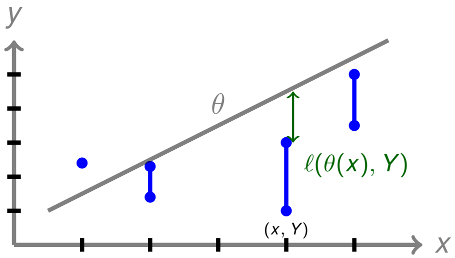
Focusing on regression

Regression



Focusing on regression

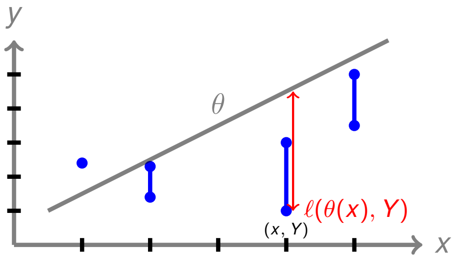
Regression



- Minimum: ℓ_{opt}

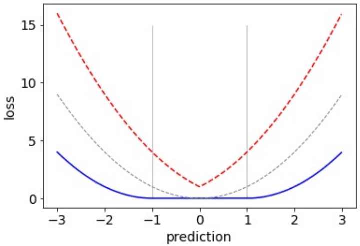
Focusing on regression

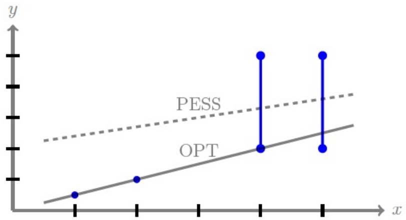
Regression



- Minimum: ℓ_{opt}
- Maximum: ℓ_{pes}

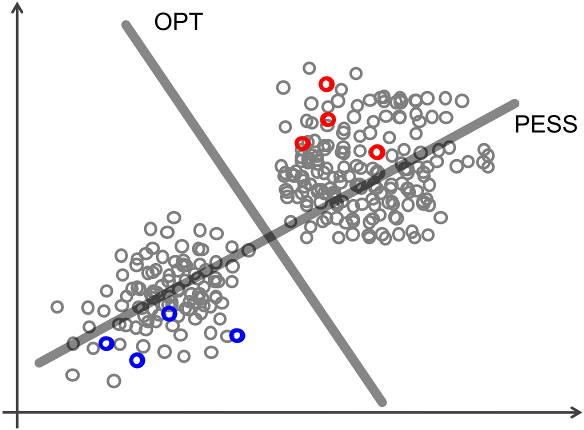
Not a trivial choice: regression example





- Pessimistic tries to be good for every replacement
- Optimistic tries to be the best for one replacement

A logistic regression example

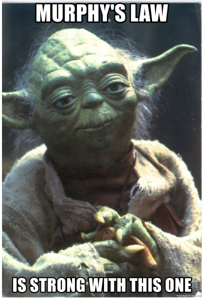


Which one should I be?

Optimist . . .

or. . .

Pessimist?



→ pretty much depends on the context!

Some elements of answer

When to be optimist?

- Reasonably sure model space Θ can capture a good predictor and is not too flexible (overfitting!)
- "imprecisiation" process random/not designed to make you fail
- can capture the best model

Optimism $\simeq$ semi-sup. learning if imprecision=missingness.

When to be pessimist?

- want to obtain guarantees in all possible scenarios ($\simeq$ distributional robustness)
- facing an "adversarial" process
- partial data=set of situations for which you want to perform reasonably well (ontic interpretation)

But what if we do not have intervals?

If our data is of the more general form $(x_i, \mathcal{P}_i)$, same kind of problems arise.

- Considering a loss $\ell : \Delta_{\mathcal{Y}} \times \Delta_{\mathcal{Y}} \to \mathbb{R}$ (e.g., cross-entropy)
- Considering that the model θ outputs a probability
- We can then define the analogous of the previous losses:

$$\ell_{opt}(\mathcal{P}, \theta(x)) = \min\{\ell(p, \theta(x)) | p \in \mathcal{P}\},$$

$$\ell_{pes}(\mathcal{P}, \theta(x)) = \max\{\ell(p, \theta(x)) | p \in \mathcal{P}\}.$$

Menu

1. Appetizer: on the nature and origins of uncertain data
   - On the nature of data uncertainty
   - On the modelling of data uncertainty: basic models
   - On the modelling of data uncertainty: more complex models
2. Main course: challenges of learning with data uncertainty
   - Challenges of learning under uncertain data
   - Uncertain data and conformal prediction
3. Dessert: when data uncertainty can be useful

Uncertain data and conformal prediction?

Assume some data are of the form $(x_i, \mathcal{P})$

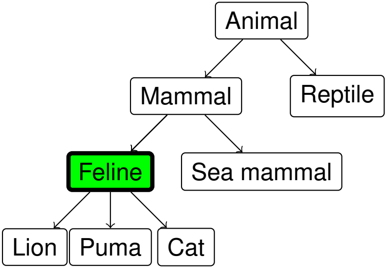
- How should we adapt conformal prediction?
- In the case of set-valued data (x_i, E_i), we can for instance
  - Compute set-valued conformity scores $[\underline{\alpha}_i, \overline{\alpha}_i]$
  - Retain one possibility, e.g. $\underline{\alpha}_i$ to be conservative
  - Proceed as usual
- Other theoretically sound options?

$\rightarrow$ some very recent work[4] on the topic does exist [3], notably looking at large prediction spaces such as rankings.

[4] Thanks to Andrea Panizza for pointing out the reference.

Why considering such issues?

- Weak labels in standard applications
- Structured information (hierarchical classification, ranking, . . .)



- Cascading approaches using predictions of previous steps (stacking, . . .)

Wrap-up on main course

Many ways to embed data uncertainty in learning

- Considering sets or imprecision provides an interesting perspective (pessimist vs optimist)
- Not all options equivalent in terms of computations and hypothesis, with a potential huge impact on practical results.

Uncertain data and conformal approaches

- Some initial work for set-valued/weak labels [3]
- Soft/probabilistic labels?
- Going beyond set-valued and probabilistic labels?

Menu

Possible utilities of uncertain data

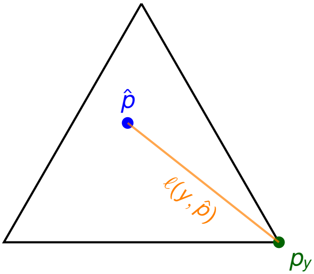
By being more cautious about the label certainty, uncertain data can:

- **Provide more regularised and better calibrated predictors, at least empirically**
- Help in self- or co-supervised learning, by being more cautious about automatically labelled examples

Cross-entropy: standard labels

$$\ell(y, \hat{p}) = -\log(\hat{p}(y)) = -\sum_y p_y(y) \log(\hat{p}(y))$$

with $p_y(y) = 1$

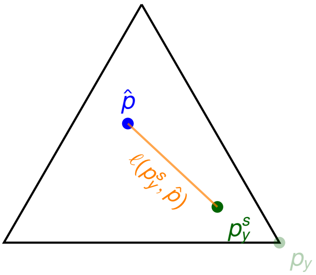


- Model is encouraged to strongly correct the prediction towards p_y
- p_y not equal to the distribution $p(|x)$

Cross-entropy: soft labels

$$\ell(p_y^s, \hat{p}) = -\sum_y p_y^s(y) \log(\hat{p}(y))$$

with $p_y^s = \alpha p_y + (1-\alpha)$*uniform*

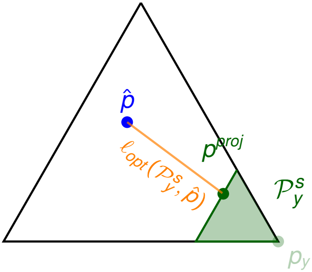


- Model still correct itself, but less strongly (regularise)
- p_y^s may be closer to $p(|x)$

Cross-entropy: credal labels

$$\ell_{opt}(\mathcal{P}_y^s, \hat{p}) = - \min_{p \in \mathcal{P}_y^s} \sum_y p(y) \log(\hat{p}(y)) = \begin{cases} 0 & \text{if } \hat{p} \in \mathcal{P}_y^s \\ L(p^{proj}, \hat{p}) \end{cases}$$

with $\mathcal{P}_y^s$ some model seen during appetizers.



- If model close enough, no correction, otherwise still regularise
- Non-null chances to include $p(|x)$

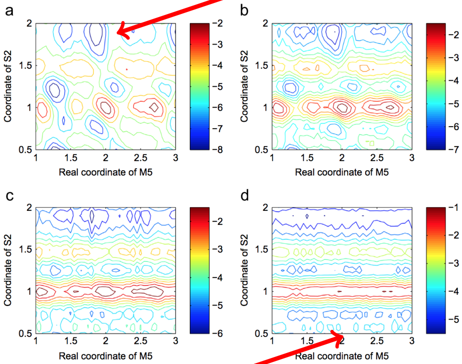
Example of results [17]

Data set	Probabilist		Credal	
	Accuracy	Calib. (ECE)	Accuracy	Calib. (ECE)
MNIST	0.98	0.11	0.98	0.01
Fash.-MNIST	0.91	0.15	0.91	0.06
CIFAR 10	0.93	0.13	0.93	0.03

$\rightarrow$ Roughly the same accuracy, but much better calibration.

Sound source separation [19]

No uncertainty description



Data uncertainty described

Possible utilities of uncertain data

By being more cautious about the label certainty, uncertain data can:

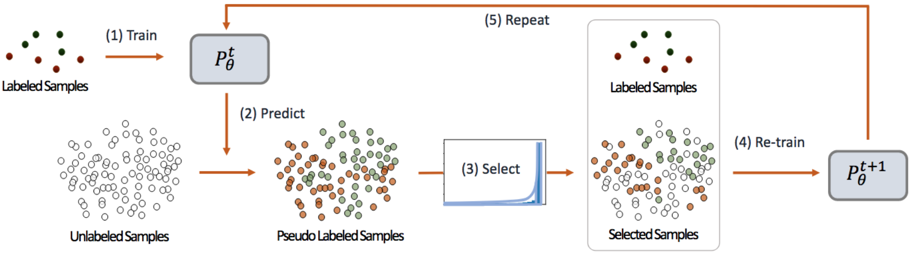
- Provide more regularised and better calibrated predictors, at least empirically
- **Help in self- or co-supervised learning, by being more cautious about automatically labelled examples**

Issue

Labelled points

Unlabelled points

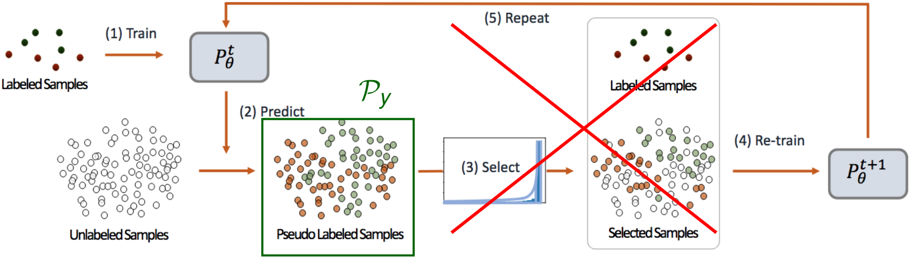
Self-labelling process [2]



Classical approach

- Replace unlabelled examples by hard labels
- Potential bias

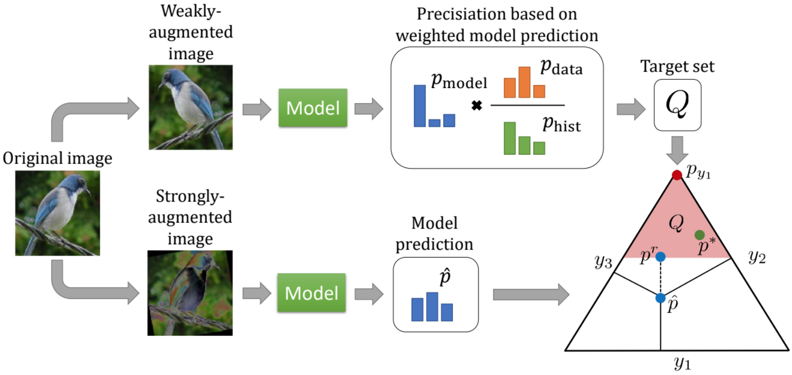
Self-labelling process [2]



Credal approach

- Replace unlabelled examples by uncertain (calibrated) labels
- Avoid potential bias while still improving

Recent use in self-supervised deep learning [16]



Some results

	CIFAR-10		SVHN	
	40 lab.	4000 lab.	40 lab.	1000 lab.
FixMatch ($\tau = 0.0$)	18.50 ±2.92	6.88 ±0.11	13.82 ±13.57	**2.73** ±0.04
FixMatch ($\tau = 0.8$)	11.99 ±2.32	7.08 ±0.13	3.52 ±0.44	2.85 ±0.08
FixMatch ($\tau = 0.95$)	14.73 ±3.29	8.26 ±0.09	5.85 ±5.10	3.03 ±0.07
LSMatch	11.60 ±2.68	7.24 ±0.21	7.04 ±3.29	2.76 ±0.05
CSSL	**10.04** ±3.32	**6.78** ±0.94	**3.50** ±0.49	2.84 ±0.06

Conclusion or final wrap-up

Rich landscape of ways to deal with uncertain data

- Credal sets offer a rich way to model uncertain data
- Some of them are strongly linked to conformal approaches

Richness can mean troubles

How to adapt learning procedure to such data is non-trivial, but raises interesting questions, notably in terms of underlying assumptions.

But richness also means opportunities

Accurate/rich uncertainty description can help in improving learned models $\rightarrow$ many things remain to do in combining conformal approaches not only with prediction, but with learning itself.

References I

[1] C. Baudrit and D. Dubois.
Practical representations of incomplete probabilistic knowledge.
Computational Statistics and Data Analysis, 51(1):86–108, 2006.

[2] Paola Cascante-Bonilla, Fuwen Tan, Yanjun Qi, and Vicente Ordonez.
Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning.
In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6912–6920, 2021.

[3] Maxime Cauchois, Suyash Gupta, Alnur Ali, and John Duchi.
Predictive inference with weak supervision.
arXiv preprint arXiv:2201.08315, 2022.

[4] Leonardo Cella and Ryan Martin.
Validity, consonant plausibility measures, and conformal prediction.
International Journal of Approximate Reasoning, 141:110–130, 2022.

[5] Timothee Cour, Ben Sapp, and Ben Taskar.
Learning from partial labels.
Journal of Machine Learning Research, 12(May):1501–1536, 2011.

[6] Inés Couso and Luciano Sánchez.
Machine learning models, epistemic set-valued data and generalized loss functions: An encompassing approach.
Information Sciences, 358:129–150, 2016.

[7] L.M. de Campos, J.F. Huete, and S. Moral.
Probability intervals: a tool for uncertain reasoning.
I. J. of Uncertainty, Fuzziness and Knowledge-Based Systems, 2:167–196, 1994.

[8] Thierry Denoeux.
Maximum likelihood estimation from uncertain data in the belief function framework.
IEEE Transactions on knowledge and data engineering, 25(1):119–130, 2013.

References II

[9] S. Destercke, D. Dubois, and E. Chojnacki.
Unifying practical uncertainty representations: I generalized p-boxes.
Int. J. of Approximate Reasoning (In press), 2008.

[10] Sébastien Destercke and Didier Dubois.
Special cases.
Introduction to Imprecise Probabilities, (chapter 4):79–91, 2014.

[11] Peter Grünwald.
Safe probability.
Journal of Statistical Planning and Inference, 195:47–63, 2018.

[12] Romain Guillaume, Inés Couso, and Didier Dubois.
Maximum likelihood with coarse data based on robust optimisation.
In *Proceedings of the Tenth International Symposium on Imprecise Probability: Theories and Applications*, pages 169–180, 2017.

[13] Dominik Hose and Michael Hanss.
A universal approach to imprecise probabilities in possibility theory.
International Journal of Approximate Reasoning, 133:133–158, 2021.

[14] Eyke Hüllermeier.
Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization.
International Journal of Approximate Reasoning, 55(7):1519–1534, 2014.

[15] Julian Lienen, Caglar Demir, and Eyke Hüllermeier.
Conformal credal self-supervised learning.
arXiv preprint arXiv:2205.15239, 2022.

References III

[16] Julian Lienen and Eyke Hüllermeier.
Credal self-supervised learning.
Advances in Neural Information Processing Systems, 34, 2021.

[17] Julian Lienen and Eyke Hüllermeier.
From label smoothing to label relaxation.
In *Proceedings of the 35th AAAI Conference on Artificial Intelligence, AAAI, Online*, 2021.

[18] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna.
Rethinking the inception architecture for computer vision.
In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[19] Xun Wang, Benjamin Quost, Jean-Daniel Chazot, and Jérôme Antoni.
Estimation of multiple sound sources with data and model uncertainties using the em and evidential em algorithms.
Mechanical Systems and Signal Processing, 66:159–177, 2016.