Learning with uncertain data: challenges and opportunities

Sébastien Destercke

Heudiasyc, CNRS Compiegne, France

SystemX seminar - chaire SAFE AI

Plan

Some examples

"Sport car" $\rightarrow$

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

$\uparrow$ Ambiguity

$\uparrow$ Ambiguity

$\uparrow$ Coarse data

Outline

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: we only look at one item
- Reducible or irreducible: whether or not one has access to better measurement/more expertise

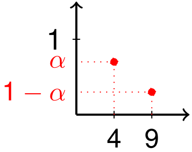
Outline

1. Appetizer: on the nature and origins of uncertain data
   - On the nature of data uncertainty
   - On the modelling of data uncertainty
   - On the origins of data uncertainty
2. Main course: learning with data uncertainty, challenges and opportunities
   - Challenges of learning under uncertain data
   - Leveraging uncertain data opportunities
3. Dessert: conclusions and beyond

Probabilistic modelling (a.k.a. soft labels)

Information: rather a 4 than a 9

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



Why (not) probabilities?

Some pros:

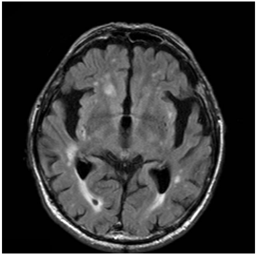
- By far the most used uncertainty model $\rightarrow$ lots of people and tools
- Naturally fits with classical loss function (cross entropy the first)

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

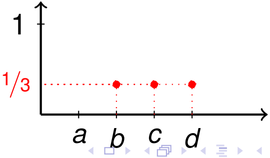
Issue in representing incompleteness

Requesting a doctor to classify Alzeimher severity degree



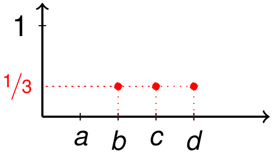
- No disease: a
- 3 degrees of severity
- labels $\mathcal{Y} = \{a, b, c, d\}$

Doctor opinion: disease present, severity difficult to assess

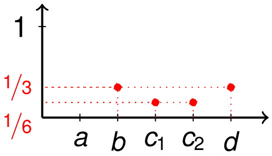


Issue in representing incompleteness

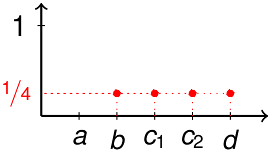
For various reasons, degree c is divided into c_1, c_2. What should



become? Two possibilities are



Coherent with initial model

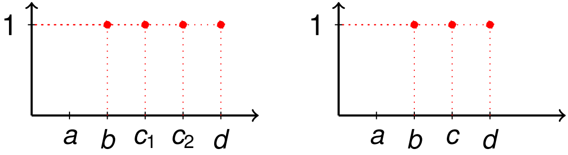


Uniform on $\{b, c_1, c_2, d\}$

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)

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



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) probabilities?

Some pros:

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

Some cons:

- Loss function adaptation requires some thinking
- Limited expressiveness (yes/no model)

Limited expressiveness

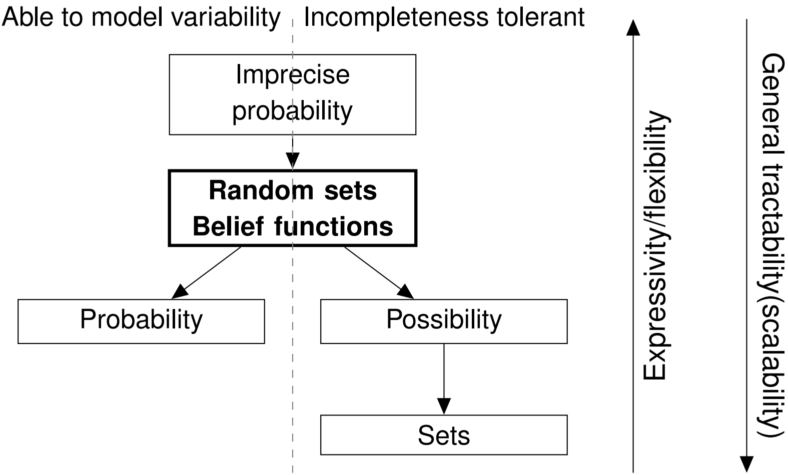
Remember this?

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 richer frameworks.

A not completely accurate but useful picture [9]

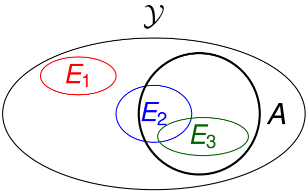


Random sets and belief functions [9]

Basic tool

A positive distribution $m : 2^{\mathcal{Y}} \to [0,1]$, with $\sum_E m(E) = 1$ and usually $m(\emptyset) = 0$, from which

- $\overline{P}(A) = \sum_{E \cap A \neq \emptyset} m(E)$ (Plausibility measure)
- $\underline{P}(A) = \sum_{E \subseteq A} m(E) = 1 - \overline{P}(A^c)$ (Belief measure)



$$\overline{P}(A) = m(E_2) + m(E_3)$$

$$\underline{P}(A) = m(E_3)$$

- Probabilities p: mass $m(\{y\}) = p(y)$ on atoms/singletons only
- Sets: $E \to$ mass $m(E) = 1$

Revisiting our example

Information: rather a 4 than a 9

Modelling by RS: $m(\{4\}) = 0.5, m(\{4,9\}) = 0.5$

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

Another practically useful example [1, 12]

A set E of most plausible values

A confidence degree $\alpha = \underline{P}(E)$

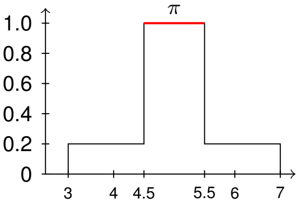
Corresponding mass:

- $m(E) = \alpha$
- $m(\mathcal{Y}) = 1 - \alpha$

Known as simple support function

pH value $\in [4.5, 5.5]$ with

$\alpha = 0.8$ ($\sim$ "quite probable")



Outline

1. Appetizer: on the nature and origins of uncertain data
   - On the nature of data uncertainty
   - On the modelling of data uncertainty
   - On the origins of data uncertainty
2. Main course: learning with data uncertainty, challenges and opportunities
   - Challenges of learning under uncertain data
   - Leveraging uncertain data opportunities
3. Dessert: conclusions and beyond

Data uncertainty as something to deal with

- Previously provided expert labels:
  - Ranked labels by likelihoods [16];
  - Imprecise quantiles in ordered settings [10];
  - Subsets with confidence degree [1];
  - Combination of multiple opinions;
  - ...
- Imperfect measurements:
  - Measurement errors as intervals;
  - Measurement errors as noise;
  - ...

How should we integrate those in learning procedures?

Data uncertainty as an opportunity

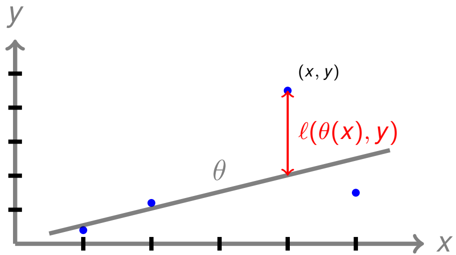
- Actively sought expert labels
  - Active learning;
  - Optimal sampling/experiment design;
  - ...
- External model providing labels for unlabelled data:
  - Probabilistic classifiers;
  - Classifiers returning sets of classes [5];
  - "Stacked" conformal predictors providing possibility distributions [3];
  - ...

How can we use those to improve upon our learning process?

Plan

How good is a model?

- Learning $\theta : \mathcal{X} \to \mathcal{Y}$ from observations (x, y)
- $\ell(\theta(x) = \hat{y}, y)$: 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

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

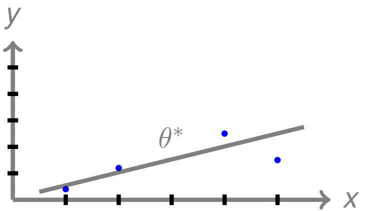
the one with lowest possible average loss

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

Prototypical cases

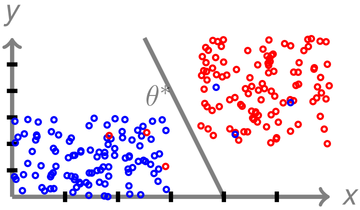
Regression

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



Classification (binary log reg)

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

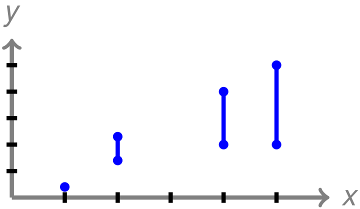


Outline

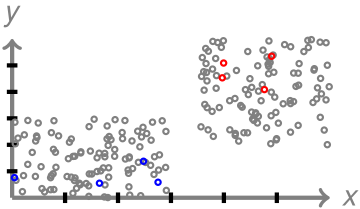
1. Appetizer: on the nature and origins of uncertain data
   - On the nature of data uncertainty
   - On the modelling of data uncertainty
   - On the origins of data uncertainty
2. Main course: learning with data uncertainty, challenges and opportunities
   - Challenges of learning under uncertain data
   - Leveraging uncertain data opportunities
3. Dessert: conclusions and beyond

The imprecise setting illustrated

Regression



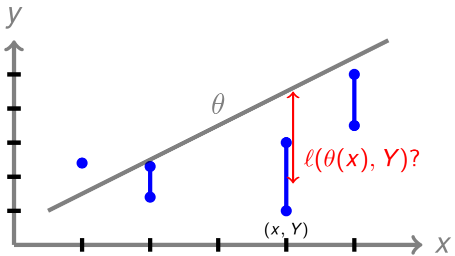
Classification (binary log reg)



How to define h_{θ^*}?

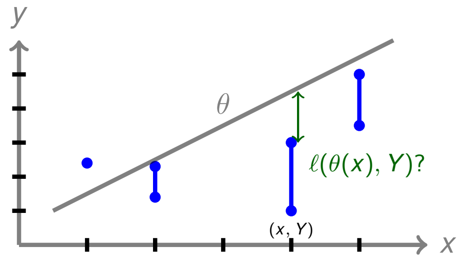
Focusing on regression

Regression



Focusing on regression

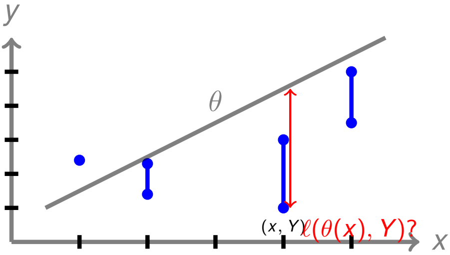
Regression



- Minimum?

Focusing on regression

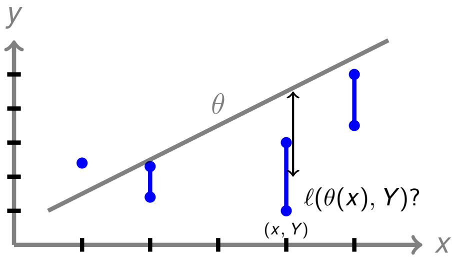
Regression



- Minimum?
- Maximum?

Focusing on regression

Regression



- Minimum?
- Maximum?
- Other? Average?

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 [7]:

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

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

[2] And likelihoods/posteriors alike

Formally speaking

- If we know the "imprecisiation" process $P_{obs}((X, Y)|(x, y))$, no theoretical problem → "merely" a computational one
- If not, common approaches are to redefine a precise criterion:
  - Optimistic (Maximax/Minimin) approach [13, 6]:

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

  - Pessimistic (Maximin/Minimax) approach [11]:

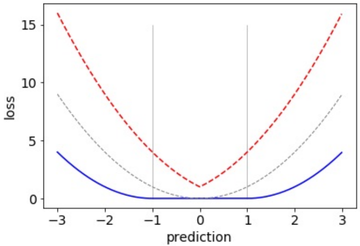
$$\ell_{pes}(\theta(x), Y) = \max\{\ell(\theta(x), Y)|y \in Y\}$$

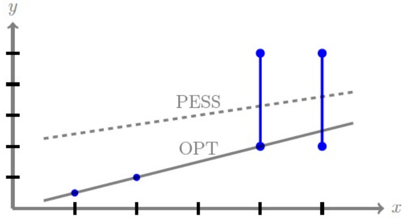
  - "EM-like" or averaging/weighting approaches[3]

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

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

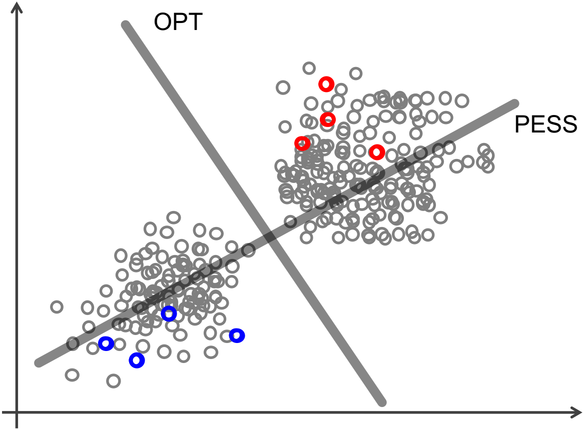
Not a trivial choice: regression example





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

A logistic regression example

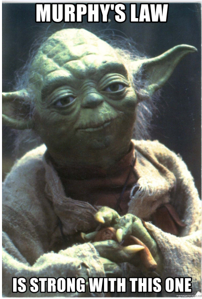


Which one should I be?

Optimist . . .

or. . .

Pessimist?



$\rightarrow$ 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

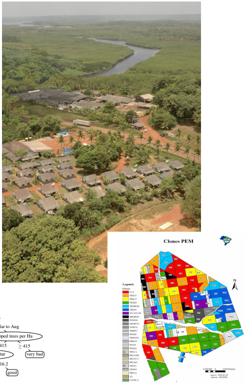
Optimism $\simeq$ semi-sup. learning if imprecision=missingness.

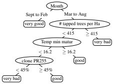
When to be pessimist?

- 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)

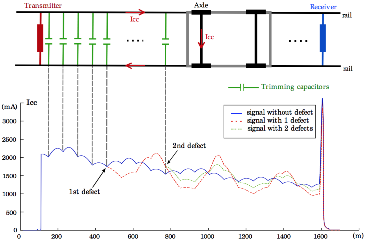
Some applications

Rubber quality prediction [18]





Railway default detection [4]



Outline

Possible utilities of uncertain data

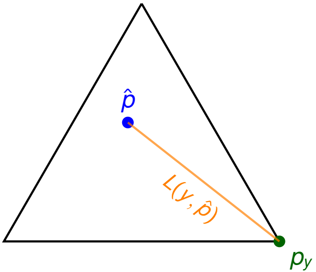
By being more cautious about the label certainty, uncertain data can:

- **Regularize/better calibrate the learning procedure $\rightarrow$ smoother learning with more trustworthy probabilistic outputs**
- Help in self- or co-supervised learning, by being more cautious about automatically labelled examples

Cross-entropy: standard labels

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

with $p_y(y) = 1$

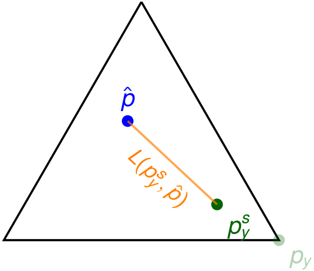


- Model is encouraged to strongly correct the prediction towards p_y
- p_y not equal to the distribution $p(|x)$

Cross-entropy: soft labels

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

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

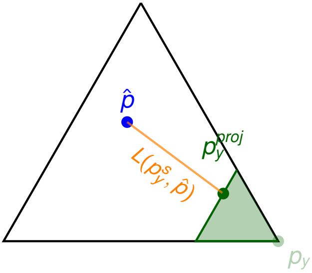


- Model still correct itself, but less strongly (regularise)
- p_y^s may be closer to $p(|x)$

Cross-entropy: credal/evidential labels

$$L(m^s_y, \hat{p}) = -\inf_{\underline{P} \leq P \leq \overline{P}} \sum_y p_y \log(\hat{p}(y)) = \begin{cases} 0 & \text{if } \underline{P} \leq \hat{P} \leq \overline{P} \\ L(p^{proj}_y, \hat{p}) \end{cases}$$

with $m^s_y = \alpha p_y + (1-\alpha)\delta$ with $\delta =$sets of all probabilities



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

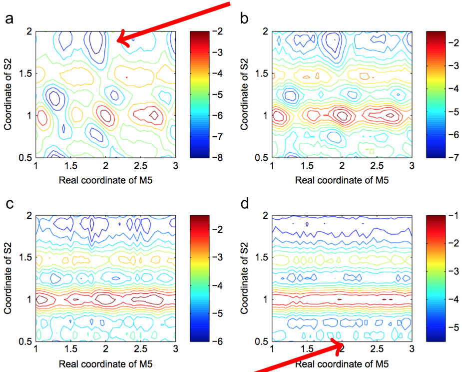
Example of results [15]

Data set	Probabilist		Credal	
	Accuracy	Calib. (ECE)	Accuracy	Calib. (ECE)
MNIST	0.98	0.11	0.98	0.01
Fashion-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:

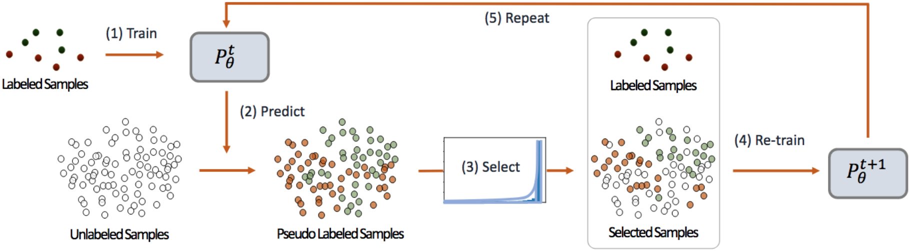
- Regularize/better calibrate the learning procedure $\rightarrow$ smoother learning with more trustworthy probabilistic outputs
- **Help in self- or co-supervised learning, by being more cautious about automatically labelled examples**

Issue

Labelled points

Unlabelled points

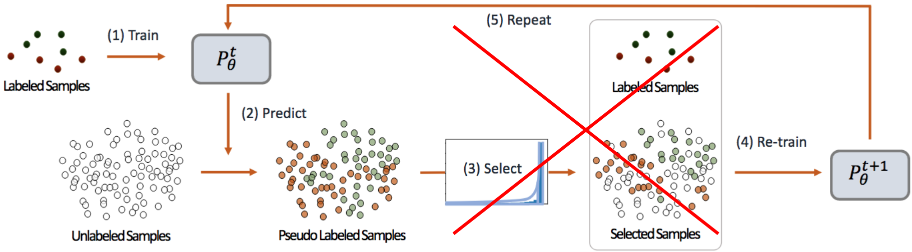
Self-labelling process [2]



Classical approach

- Replace unlabelled examples by hard labels
- Potential bias

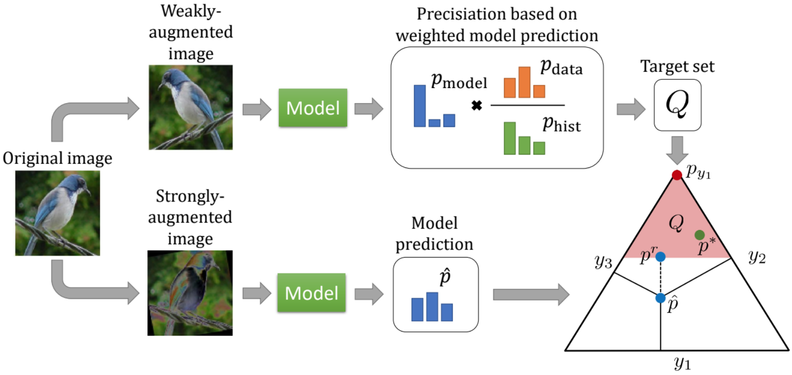
Self-labelling process [2]



Credal approach

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

Recent use in self-supervised deep learning [14]



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

Plan

Conclusions

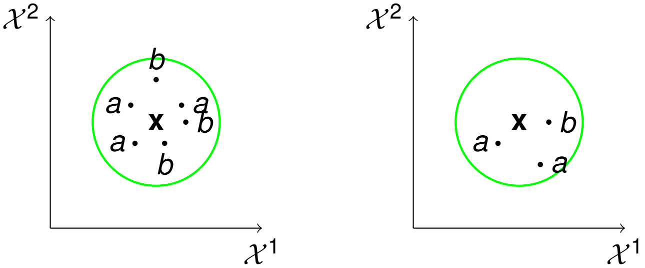
Uncertain data as a constraint:

- Need to adapt standard learning;
- Way to do so heavily impact result.

Uncertain data as an opportunity:

- Modelling uncertainty as a means to regularise obtained model
- Uncertainty-aware labels as an improvement to self-supervised, automatic-labelling training

Uncertainty quantification



$P(\mathbf{x} = a) \simeq 0.5$ — Aleatoric uncertainty

$P(\mathbf{x} = a) \in [0.35, 0.95]$ — Epistemic uncertainty

Differentiating these two aspects useful in:

- Active learning (lack of knowledge vs decision border) [17]
- Reasons to doubt a classification result (explainability, reject)

Recognition of conflicting examples

Belief functions allow[4] $m(\emptyset) > 0$,

Two possible interpretations leading to possible use:

- $m(\emptyset)$= degree of conflict $\rightarrow$ analysing the sources of this conflict to explain its origins (XAI)
- $m(\emptyset)$ =probability that the class is unknown $\rightarrow$ use it to detect novelties/unknown anomalies

[4] Can be assimilated to conformal prediction giving $\emptyset$

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

[4] Zohra L Cherfi, Latifa Oukhellou, Etienne Côme, Thierry Denoeux, and Patrice Aknin.
Partially supervised independent factor analysis using soft labels elicited from multiple experts: Application to railway track circuit diagnosis.
Soft computing, 16(5):741–754, 2012.

[5] Giorgio Corani, Alessandro Antonucci, and Marco Zaffalon.
Bayesian networks with imprecise probabilities: Theory and application to classification.
In *Data Mining: Foundations and Intelligent Paradigms*, pages 49–93. Springer, 2012.

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

[7] 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.

References II

[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.

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

[10] Scott Ferson, VLADIK KREINOVICK, Lev Ginzburg, and FARI SENTZ.
Constructing probability boxes and dempster-shafer structures.
Technical report, Sandia National Lab.(SNL-NM), Albuquerque, NM (United States); Sandia . . . , 2003.

[11] 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.

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

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

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

References III

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

[16] Enrique Miranda and Sébastien Destercke.
Extreme points of the credal sets generated by comparative probabilities.
Journal of Mathematical Psychology, 64:44–57, 2015.

[17] Vu-Linh Nguyen, Sébastien Destercke, and Eyke Hüllermeier.
Epistemic uncertainty sampling.
In *International Conference on Discovery Science*, pages 72–86. Springer, 2019.

[18] Nicolas Sutton-Charani, Sébastien Destercke, and Thierry Denoeux.
Application of e 2 m decision trees to rubber quality prediction.
In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 107–116. Springer, 2014.

[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.