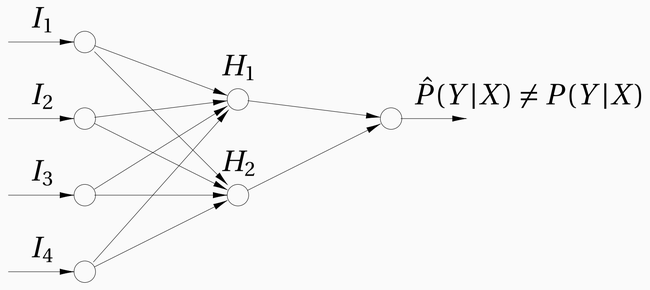
Learning calibrated belief functions from conformal predictions

Vitor Martin Bordini, Sebastien Destercke and Benjamin Quost

CNRS, UMR 7253 Heudiasyc, UTC

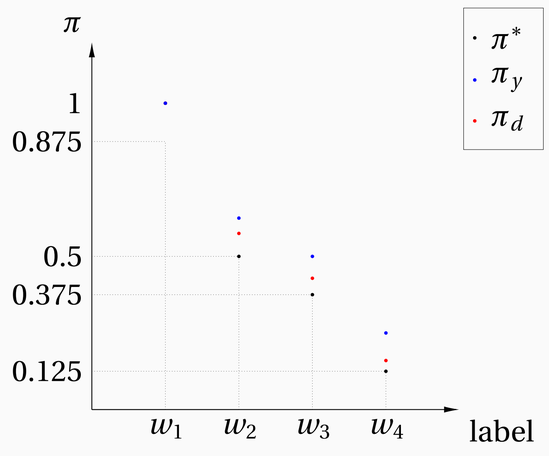
Problem

- Common problem on machine learning predictions: poor calibration.
- Calibration definition: The probability $\hat{P}$ given by a model h for a class w and input x is the true probability, i.e, $\mathbb{P}(y = w|h(x) = \hat{P}) = \hat{P}$.
- Inductive Conformal Prediction (ICP) [1] is a possible solution to this problem.
- What is the relation (if any) between ICPs and Imprecise Probabilities?



Possibility theory

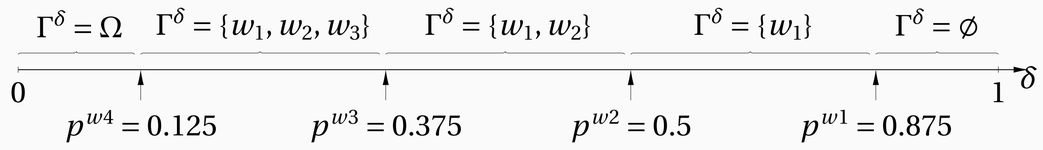
- It is proven that **ICP output p is equal to a possibility distribution** $\pi : \Omega \mapsto [0, 1]$.
- We want to normalize π such that $\max \pi^* = 1$ where π^* is the normalized distribution.
- Limitation: we can only extract imprecise probabilities from a possibility distribution.
- Normalisation: Ours $(\pi^*(w^*) = 1, \pi^*(w) = \pi(w), w^* = \arg\max(\pi(w)) \quad \text{for all } w \neq w^*)$, Dempster's or Yager's?
- We can build a Belief function $Bel(B)$ with a Necessity Measure N such that $Bel(B) := N(B) = 1 - \max_{x \in \neg B} \pi(x), B \subseteq \Omega$.
- In the example below, we can compute the Belief functions as $Bel(\{w_1\}) = 1 - \max_{w \in \{w_2, w_3, w_4\}} \pi^*(w) = 0.5$, $Bel(\{w_1, w_2\}) = 1 - \max_{w \in \{w_3, w_4\}} \pi^*(w) = 0.625$, $Bel(\{w_1, w_2, w_3\}) = 1 - \max_{w \in \{w_4\}} \pi^*(w) = 0.875$, and so on.



Inductive Conformal Prediction

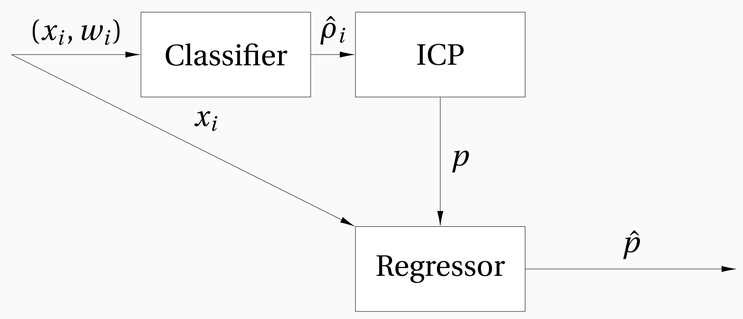
- Assumes the dataset $Z = \{(x_i, w_i), w_i \in \Omega | i = 1, \ldots, n\}$ is exchangeable.
- Compute non-conformity scores β_i.
- Computes p-values, ICP output, by comparing the non-conformity scores of a single example and the ones of the calibration set.
- In the example below, w_1 is the best prediction because p_{w_1} has the highest value among all (β_{w_1} is the smallest).

- P-values property: $P(\{p(w_i) \geq \delta\}) \geq 1 - \delta, w_i \in \Omega, \delta in(0, 1)$. We have thus Conformal Regions:$\Gamma^\delta(x) = \{w_j : \pi_x(w_j) \geq \delta\}$ (see example below).
- Advantages: Simple to implement/understand and with a rigorous theory behind it.
- Drawbacks: Calibration set (needs more data) and a bit slower.



Our solution

- Our hypothesis: ICP outputs can be learned directly from a machine learning model.
- We need p-values as labels, which doesn't exist in any public datasets.
- We train a model that estimates probability distributions and then we apply the ICP on this model output to compute p-values.
- This p-values are the labels to train a regressor.
- P-value vector p can be interpreted as a possibility distribution π.
- Estimation of calibrated belief functions via possibility distribution.



Experiments

- Classifier: EfficientNet [3] or Logistic Regression depending on the dataset.
- Regressor: EfficientNet feature extractor + 4 linear layers or Random Forest depending on the dataset.

data set	# training instances	# test instances	Input Shape	# classes
Digits	1437	360	(8,8,1)	10
Heart disease	771	147	(9,1)	2
Titanic	748	143	(7,1)	2
Symptom2Disease	960	240	(384,1)	24
SVHN	1437	360	(32,32,3)	10
Cifar10	50,000	10,000	(32,32,3)	10
Cifar100	50,000	10,000	(32,32,3)	100
Artists	6700	1676	(512,512,3)	50

P-values comparison

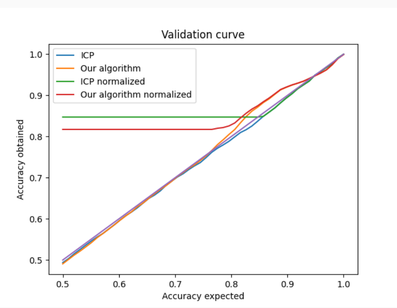
- Comparison between ICP and the regressor outputs.
- Calibration dataset = 10% of the train dataset.
- The Mean Square Root(MSR) and the R2 coefficient are presented on table.

data set	Classifier	Regressor	Classifier accuracy	RMSR (10^{-3})	R2 coeff.
Digits	logistic regression	random forest	96	3.6	0.84
Heart			82	3.6	0.96
Titanic			80	0.9	0.99
Symptom2Disease	EfficientNet.v2	feature extractor + 4 dense linear layers	96	7	0.83
SVHN			94	0.03	0.99
Cifar10			85	0.34	0.98
Cifar100			62	9.80	0.17
Artists			89	0.80	0.78

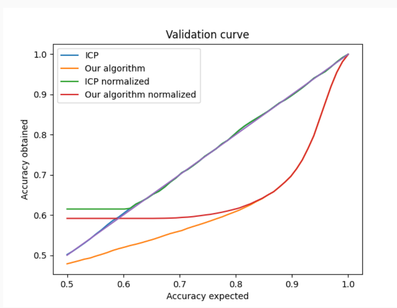
Calibration curves

- Check whether calibration properties are maintained.
- Regressor does not scale well with the number of classes.

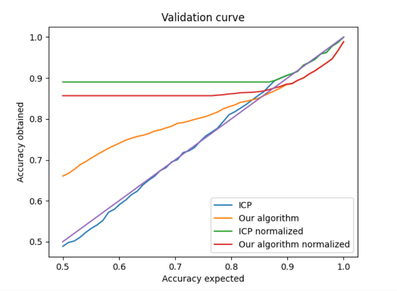
Cifar 10



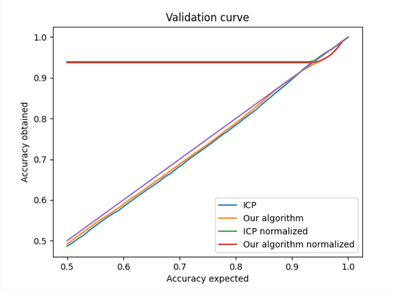
Cifar100



Artist



SVHN



Conclusion

- Calibration techniques make model predictions statistically valid.
- The ICP is a popular calibration technique but it is slower and requires more data.
- Our algorithm decrease the dependence of ICP on the calibration dataset while also being less computationally expensive and having similar performance.
- However, it still requires a minimum amount of data and takes more time to learn.
- Future works may solve this problem using co-learning techniques [4] [2].

References

[1] Harris Papadopoulos. Inductive conformal prediction: Theory and application to neural networks. In Paula Fritzsche, editor, *Tools in Artificial Intelligence*, chapter 18. IntechOpen, Rijeka, 2008.

[2] Yann Soullard, Sebastien Destercke, and Indira Thouvenin. Co-training with credal models. In *Artificial Neural Networks in Pattern Recognition: 7th IAPR TC3 Workshop, ANNPR 2016, Ulm, Germany, September 28–30, 2016, Proceedings 7*, volume 9896, pages 92–104, 09 2016.

[3] Mingxing Tan and Quoc V. Le. Efficientnetv2: Smaller models and faster training. *International Conference on Machine Learning*, 2021.

[4] Yingda Xia, Dong Yang, Zhiding Yu, Fengze Liu, Jinzheng Cai, Lequan Yu, Zhuotun Zhu, Daguang Xu, Alan Yuille, and Holger Roth. Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation. *Medical image analysis*, 65:101766, 2020.