Quantifying Aleatoric and Epistemic Uncertainty in Machine Learning

Are Conditional Entropy and Mutual Information Appropriate Measures?

Lisa Wimmer[1,3], Yusuf Sale[2,3], Paul Hofman[2,3], Bernd Bischl[1,3], Eyke Hüllermeier[2,3]

[1] Department of Statistics, LMU Munich
[2] Institute of Informatics, LMU Munich
[3] Munich Center for Machine Learning (MCML)

Predictive uncertainty

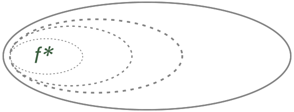
Common understanding

Uncertainty arising from[1]

⊢ imperfect data

⊢ limited knowledge

Implicit assumption



$f : \mathcal{X} \rightarrow \mathcal{Y}$

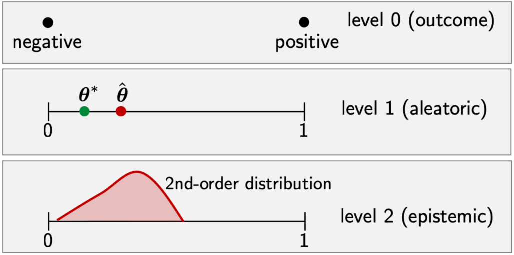
Uncertainty components

$$\underbrace{\text{Total uncertainty}}_{\textbf{TU}} \stackrel{?}{=} \underbrace{\text{aleatoric uncertainty}}_{\textbf{AU}} + \underbrace{\text{epistemic uncertainty}}_{\textbf{EU}}$$

[1] Hüllermeier and Waegeman (2021), Kendall and Gal (2017)

3+ levels

- ⊢ **Level 0** $y \in \mathcal{Y}$ // no uncertainty
- ⊢ **Level 1** $\boldsymbol{\theta} \in \mathbb{P}(\mathcal{Y})$ // uncertainty about $y \mid \mathbf{x} \rightsquigarrow$ AU
- ⊢ **Level 2** $Q \in \mathbb{P}(\mathbb{P}(\mathcal{Y}))$ // uncertainty about $\boldsymbol{\theta} \rightsquigarrow$ EU



Intuition

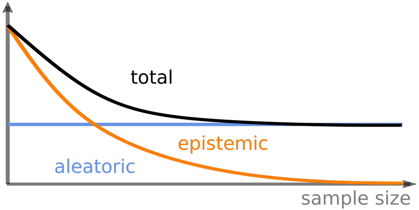
Uncertainty components

$$\underbrace{\text{Total uncertainty}}_{\textbf{TU}} \stackrel{?}{=} \underbrace{\text{aleatoric uncertainty}}_{\textbf{AU}} + \underbrace{\text{epistemic uncertainty}}_{\textbf{EU}}$$

$\vdash$ **TU**, **EU** maximal for total ignorance

$\vdash$ **EU** $\rightarrow 0$ for $n \rightarrow \infty$

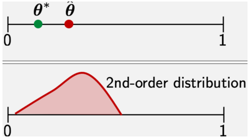
$\vdash$ **AU** $\equiv c \in \mathbb{R}$



Entropy-based measures

TU – Shannon entropy[2]

$$H(Y) = H(\mathbb{E}_Q[Y \mid \theta]) = -\sum_y p(y) \cdot \log p(y)$$



AU – conditional entropy

$$H(Y \mid \Theta) = \mathbb{E}_Q[H(Y \mid \theta)] = \mathbb{E}_Q\left[-\sum_y p(y|\theta) \log p(y|\theta)\right]$$

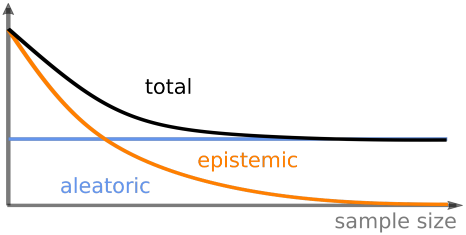
EU – mutual information

$I(Y, \Theta) = H(Y) - H(Y \mid \Theta)$ // uncertainty reduction in Y

[2] Shannon (1948), Houlsby et al. (2011), Cover and Thomas (2006)

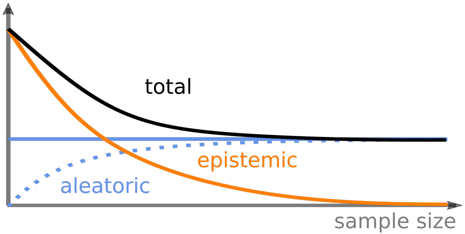
Fundamental relationship

$$\underbrace{\text{Shannon entropy}}_{\text{TU}} = \underbrace{\text{conditional entropy}}_{\text{AU}} + \underbrace{\text{mutual information}}_{\text{EU}}$$



Fundamental relationship

$$\underbrace{\text{Shannon entropy}}_{\text{TU}} = \underbrace{\text{conditional entropy}}_{\text{AU}} + \underbrace{\text{mutual information}}_{\text{EU}}$$



Proposition 5. *If EU and TU attain their respective maxima at the beginning of learning, and they are constructed to be on the same scale, then TU cannot decompose additively into EU and AU if AU is positive.*

Desired formal properties

A0 TU, AU and EU are non-negative.

A1 EU vanishes for Dirac measures $Q = \delta_{\boldsymbol{\theta}}$.

A2 **EU** and **TU** are maximal for Q being the uniform distribution.

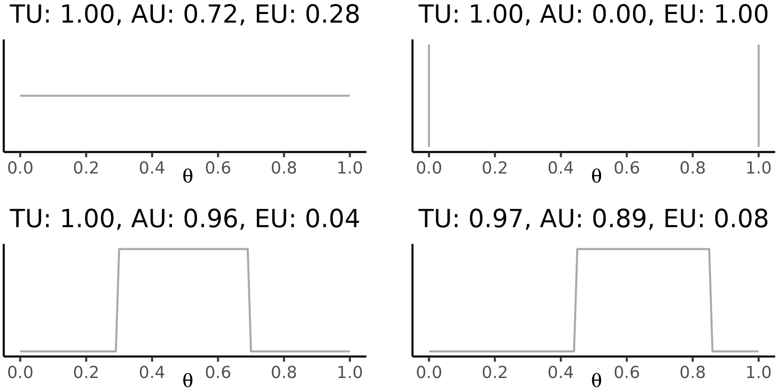
A3 If Q' is a mean-preserving spread of Q, then $\text{EU}(Q') \geq \text{EU}(Q)$ (weak version) or $\text{EU}(Q') > \text{EU}(Q)$ (strict version); the same holds for **TU**.

A4 If Q' is a center-shift of Q, then $\text{AU}(Q') \geq \text{AU}(Q)$ (weak version) or $\text{AU}(Q') > \text{AU}(Q)$ (strict version); the same holds for TU.

A5 If Q' is a spread-preserving location shift of Q, then $\textbf{EU}(Q') = \textbf{EU}(Q)$.

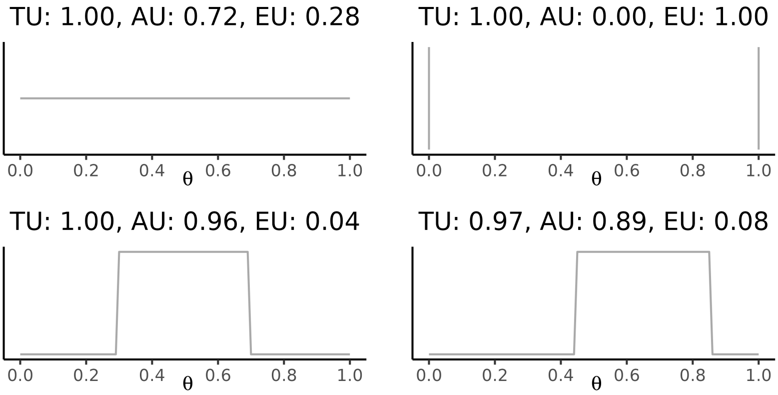
A paradoxical example

A2 **EU** and **TU** are maximal for Q being the uniform distribution.

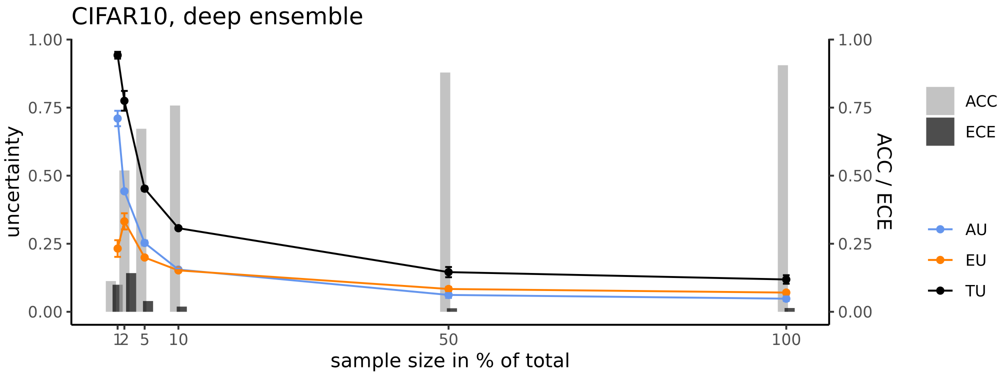


A paradoxical example

A5 If Q' is a spread-preserving location shift of Q, then $\mathbf{EU}(Q') = \mathbf{EU}(Q)$.

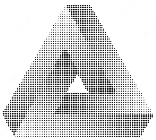


Empirical evidence



Main criticism

- Inadequacy of standard uncertainty measures
  - **EU**: counter-intuitive behavior; measure of conflict rather than ignorance
  - **AU**: estimation under intrinsic uncertainty from level 2
  - **TU**: loss of information due to marginalization
- **Additivity**: not possible for finite n under **A0–A5**



A way forward

Better measures[3]

- **Axiomatic** foundation
- Inter-level uncertainty **propagation**

Other representational frameworks[3]

- Beyond classical **probability** theory

[3] Hüllermeier et al. (2022), Sale et al. (2023), Dubois et al. (1996)

Stop by

Poster #374

Questions?

References

Figure on last page. https://www.pexels.com/photo/white-and-black-mountain-wallpaper-933054/, 2023.

Figure of Penrose triangle. https://www.freepik.com/premium-vector/penrose-triangle-halftone-geometric-minimalism-print_5069990.htm, 2023.

Figure on title page. https://losslandscape.com/, 2023.

T. M. Cover and J. A. Thomas. Elements of Information Theory. Wiley, 2 edition, 2006.

D. Dubois, H. M. Prade, and P. Smets. Representing partial ignorance. 26(3):361–377, 1996.

N. Houlsby, F. Huszár, Z. Ghahramani, and M. Lengyel. Bayesian Active Learning for Classification and Preference Learning, 2011.

E. Hüllermeier and W. Waegeman. Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods. 2021.

E. Hüllermeier, S. Destercke, and M. H. Shaker. Quantification of Credal Uncertainty in Machine Learning: A Critical Analysis and Empirical Comparison. In Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence, 2022.

A. Kendall and Y. Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? 2017.

Y. Sale, M. Caprio, and E. Hüllermeier. Is the volume of a credal set a good measure for epistemic uncertainty? In R. J. Evans and I. Shpitser, editors, Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence, volume 216 of Proceedings of Machine Learning Research. PMLR, 2023.

C. E. Shannon. A Mathematical Theory of Communication. 27:379–423, 1948.