INITIATIVE FOR APPLIED ARTIFICIAL INTELLIGENCE

Aleatoric and Epistemic Uncertainty in ML

Michael Panchenko, 02.11.2022

Aleatoric and Epistemic: quick recap

- Aleatoric uncertainty is "irreducible uncertainty" that is intrinsic to the data. Example: uncertainty in the **outcome of a fair coin toss**
- Epistemic uncertainty is uncertainty in the suitability of the predictive model for a certain input. Example: **uncertainty in the fairness of the coin** that is being tossed
- Unfortunately, these two are generally difficult to separate and distinguish
- Bayesian inference: $p(y \mid x_q) = \int_{\mathcal{H}} p(y \mid x_q, h)\, dP(h \mid D)$, where $\mathcal{H}, D$ are the hypothesis space and the data, x_q stands for "query"
- Bayesian coin toss: $\mathcal{H}$ is parametrized by fairness of coin α; $p(y = 1 \mid \alpha) = \alpha$
- The distributions $p(\alpha) = \delta\left(\alpha - \frac{1}{2}\right)$ (no epistemic uncertainty) and $p(\alpha) = 1$ ("maximal" epistemic uncertainty) result in **the same total uncertainty**!

Probability theory might not be sufficient

- Roland Fisher: "not knowing the chance of mutually exclusive events and knowing the chance to be equal are two quite different states of knowledge."

- For binary classification: let $q := p(+1 \mid x_q)$. From the Bayesian framework we can derive the probability distribution of q by marginalizing the posterior $p\left(q \mid x_q \right) = \int_{\mathcal{H}} \mathbb{I}_{\{h(x_q)=q\}} \, dP(h \mid D)$. This is analogous to $p(\alpha)$ in the coin toss.

- The epistemic uncertainty is a **"measure of variability"** of this (second order probability) distribution. A peaked distribution corresponds to low epistemic uncertainty, a smeared distribution to high epistemic uncertainty.

- Aleatoric uncertainty is related to where the peaks of $p\left(q \mid x_q \right)$ are

- Which measure to take for epistemic uncertainty is far from clear, natural candidates like entropy or variance both have downsides

Intuition on the different uncertainties

Formulated in terms of data, aleatoric uncertainty is due to overlapping data points and epistemic uncertainty is due to lack of data similar to the queried point

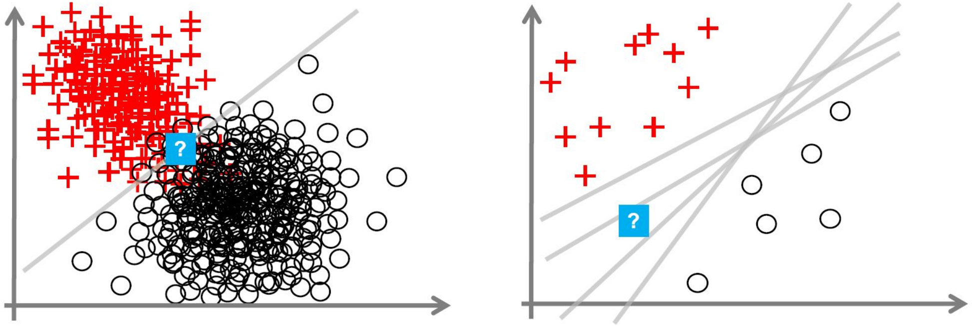


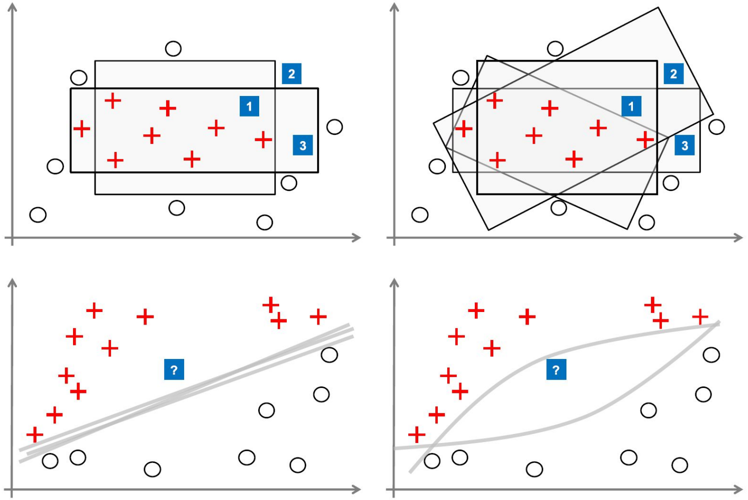
Fig. 6 Left: Even with precise knowledge about the optimal hypothesis, the prediction at the query point (indicated by a question mark) is aleatorically uncertain, because the two classes are overlapping in that region. Right: A case of epistemic uncertainty due to a lack of knowledge about the right hypothesis, which is in turn caused by a lack of data

Sources of epistemic uncertainty: **model uncertainty** (controlled by the richness of the hypothesis $\mathcal{H}$) and **approximation uncertainty**, controlled by the data and the learning algorithms. The former tends to be of a subordinate significance

Epistemic uncertainty in version-spaces

Version space learning: maintain a **set of hypothesis that is consistent with data**. The prediction is a subset of the possible labels, produced by all elements of the version space.

In this approach there is no aleatoric uncertainty, epistemic uncertainty can be unambiguously captured by the size of the set of predictions.



While rather impractical by itself, this setting is useful for obtaining intuition on how methods for (epistemic) uncertainty estimation can be developed.

This kind of approach is also related to various methods for **estimation of uncertainty from ensembles**.

Estimating the two uncertainties in ML

- The Fisher information matrix $-\mathbb{E}[\frac{\partial^2 l}{\partial\theta_i \partial\theta_j}]$ gives a measure of certainty in the parameters obtained through MLE (it is expensive though).
- In Gaussian Processes, aleatoric uncertainty is modeled explicitly and can be simply subtracted from total uncertainty.
- The mutual information between predicted labels and weights captures epistemic uncertainty. It can be estimated with ensemble techniques.
- Outlier detection or classification with rejection can also be viewed as methods for estimating epistemic uncertainty. As mentioned before, it is caused by sparsity of data. In particular, outlier scores can be used as numerical proxies for this kind of uncertainty.

Beyond probability theory

- Set-prediction methods are somehow more amenable to an analysis of epistemic uncertainty than purely probabilistic ones.

- Example: **Conformal Predictions**. Given a function that predicts **a strangeness score on input-output tuples** $f : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$, one associates a "p-value" to a query x_{N+1} with prediction y_{N+1} by counting $p(y_{N+1}) := \frac{\#\{i | f(x_i, y_i) \geq f(x_{N+1}, y_{N+1})\}}{N+1}$. All values of y with a sufficiently high p-value are reported.

- Intuition: the more points from the original dataset are stranger than the new tuple, the better is the predicted label y_{N+1} of the tuple

- The difference between highest and second highest p-value for a query is a measure of epistemic uncertainty

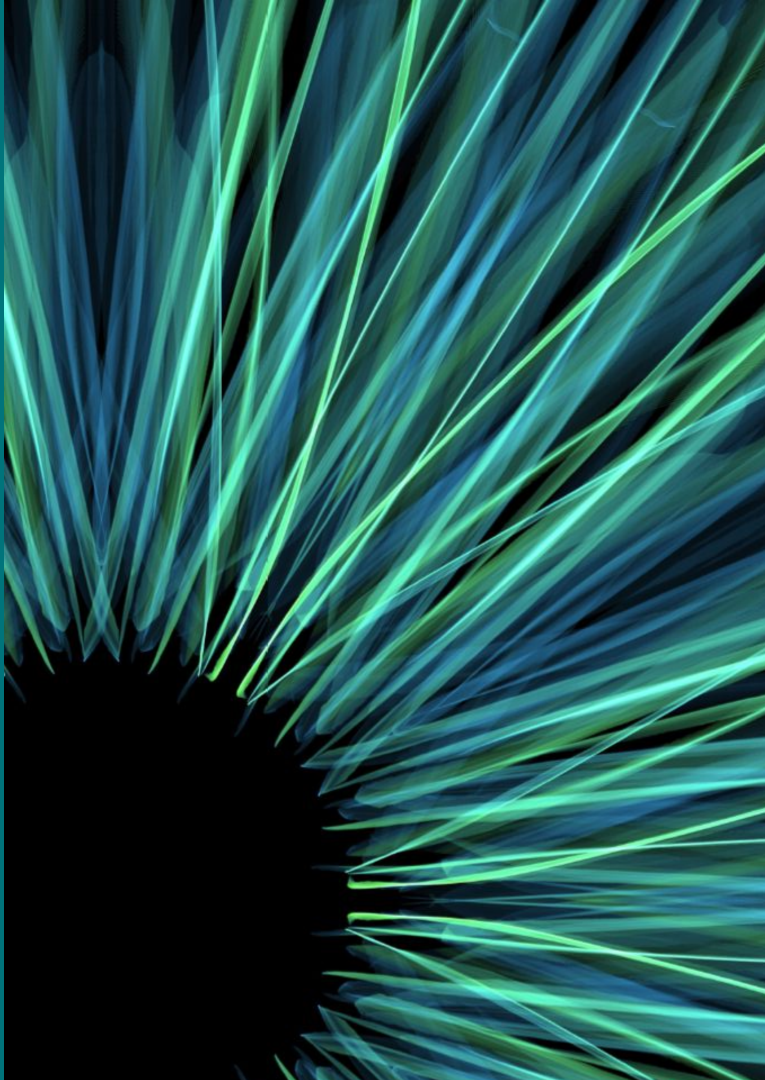
- Orthogonal approach: plausibility scores – introducing non-probabilistic quantities, like ratios of likelihood functions.

Conclusion

- The total uncertainty is a sum of the epistemic and aleatoric uncertainties
- Separating these two from each other is not trivial, there is no consensus on the best way to do so
- Purely probabilistic methods might not be sufficient for this task, it is an active area of research

INITIATIVE FOR APPLIED ARTIFICIAL INTELLIGENCE

Thank you!



appliedAI Initative GmbH
Freddie-Mercury-Str. 5
80797 Munich, Germany

Tel.+49 (89) 18 94 69-0
www.appliedai.de