Aleatoric and Epistemic Uncertainty in Machine Learning: An Ensemble-based Approach

Eyke Hüllermeier

Paderborn University
Heinz Nixdorf Institute
Department of Computer Science

18–SEP–2020

Introduction

- Machine learning is inseparably connected with **uncertainty**.
- **Learning** in the sense of generalizing beyond the data seen so far is necessarily based on a process of **induction**.
- **Models** induced from data are never provably correct, but only hypothetical and therefore uncertain, and the same holds true for the **predictions** produced by a model.
- Other sources of uncertainty: incorrect model assumptions, noisy or imprecise data, etc.
- **Trustworthy representation** of uncertainty is desirable and should be considered as a key feature of any machine learning method, all the more in safety-critical application domains.

Introduction

- Many applications require safe and reliable predictions, and hence a certain level of **self-awareness** of ML systems:
  - equip predictions with an appropriate **quantification of uncertainty**,
  - **reject** a decision in cases of high uncertainty (abstention) ,
  - deliver a credible **set-valued prediction** (partial abstention),
  - ...

Driver assistance systems: a safety-critical application

Introduction

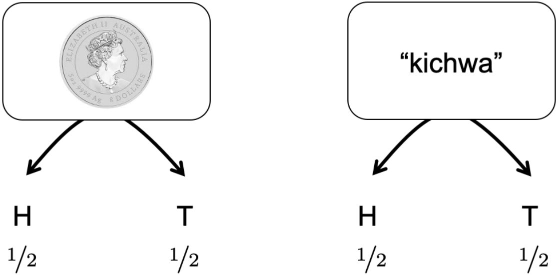
Example of a lack of "uncertainty-awareness":

Predictions by EfficientNet (Tan and Le, 2019) on test images from ImageNet: For the left image, the neural network predicts "typewriter keyboard" with certainty 83.14 %, for the right image "stone wall" with certainty 87.63 %.

Aleatoric versus epistemic uncertainty

- Traditional approaches in ML fail to distinguish inherently different sources of uncertainty, often referred to as **aleatoric** and **epistemic** uncertainty (Hora, 1996; Der Kiureghian and Ditlevsen, 2009).
- Motivated in the context of ML for medical diagnosis by Senge et al. (2014), increasing attention more recently due to interest by the deep learning community (Kendall and Gal, 2017).
- **Aleatoric** (*aka* statistical) uncertainty refers to the notion of **randomness**, that is, the variability in the outcome of an experiment which is due to inherently random effects.
- **Epistemic** (*aka* systematic) uncertainty refers to uncertainty caused by a **lack of knowledge**, i.e., to the epistemic state of the agent.
- As opposed to aleatoric uncertainty, epistemic uncertainty can in principle be reduced on the basis of additional information.

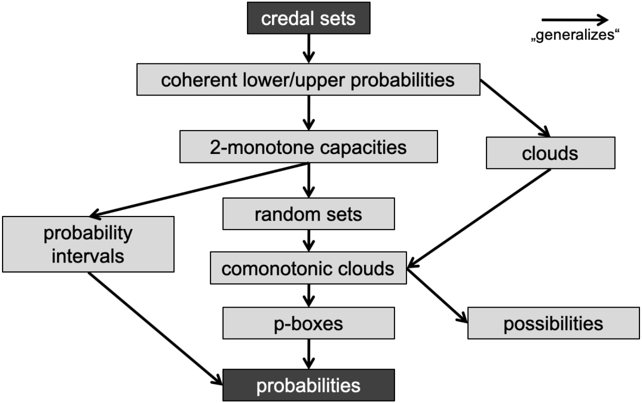
Aleatoric versus epistemic uncertainty



"Not knowing the chance of mutually exclusive events and knowing the chance to be equal are two quite different states of knowledge"

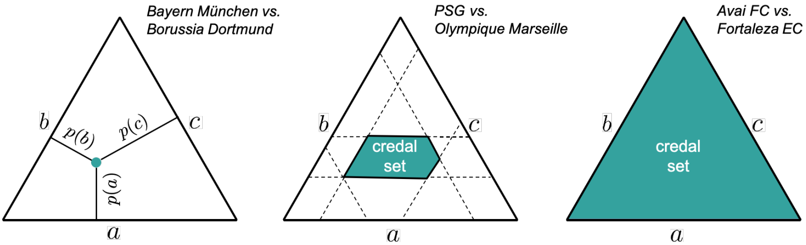
Ronald Fisher (1890-1962)

Generalized uncertainty calculi



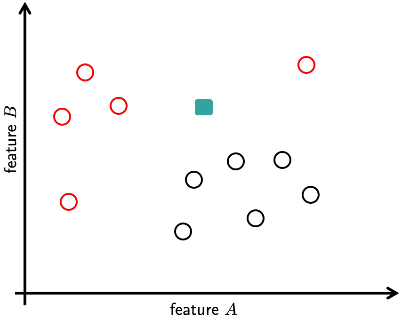
Largely motivated by the observation that standard probability is (arguably) inappropriate to represent ignorance, to weaken closed world assumption, etc.

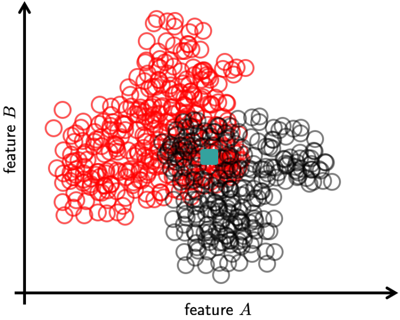
Credal sets



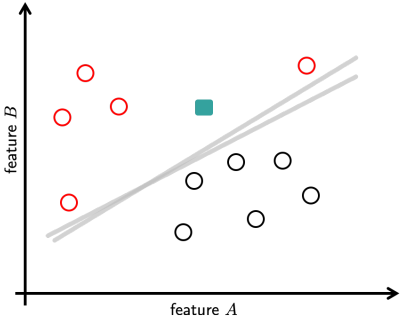
Probability distributions $p = (p(a), p(b), p(c))$ on $\Omega = \{a, b, c\}$, for example $\Omega = \{\text{home wins}, \text{draw}, \text{away wins}\}$, as points in a Barycentric coordinate system.

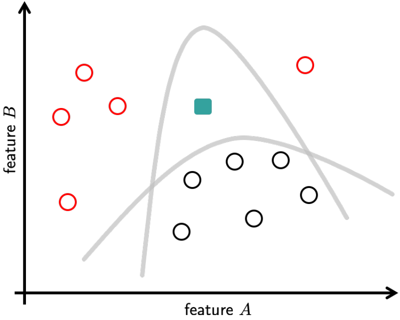
Aleatoric versus epistemic uncertainty in ML





Aleatoric versus epistemic uncertainty in ML

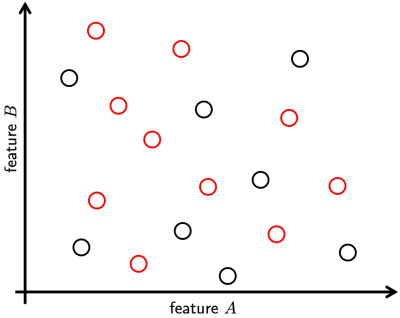


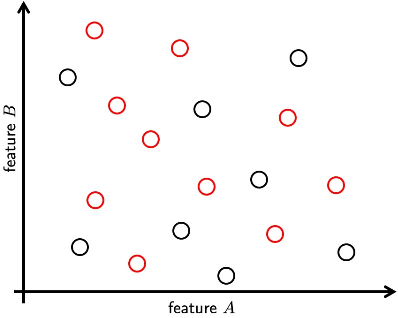


strong prior

weaker prior

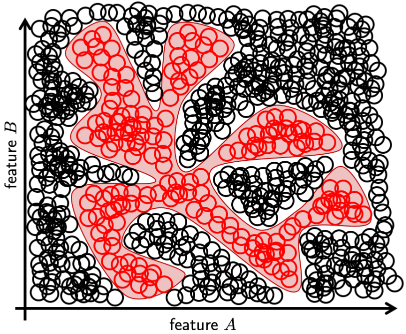
Aleatoric versus epistemic uncertainty in ML

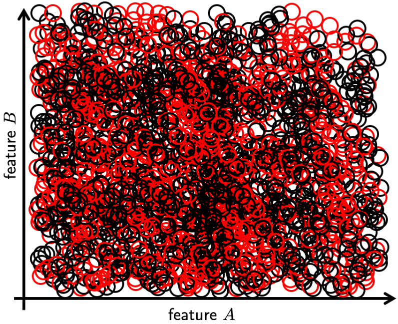




Is the uncertainty aleatoric or epistemic?

Aleatoric versus epistemic uncertainty in ML

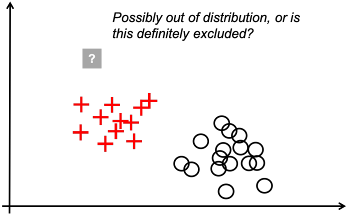




Like random versus pseudo-random numbers ...

Problem setting and assumptions

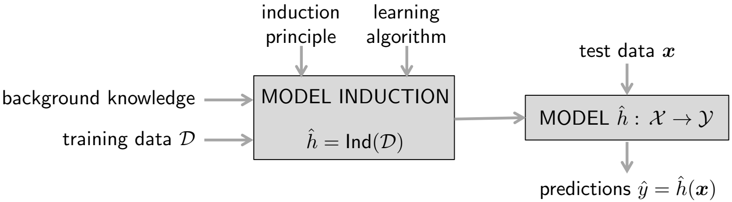
- A precise specification of the **problem setting** and **underlying assumptions** is an important prerequisite, not only for providing learning guarantees, but also for **uncertainty quantification**.



- Here, one might be quite sure about the class of the query under standard assumptions of binary classification, but much less so in a setting of **novelty detection**, where new classes may emerge.
- Likewise, assumptions such as **i.i.d. data generation** are really crucial (the past should be representative of the future).

Supervised learning and predictive uncertainty

- Uncertainty occurs in various facets in machine learning, and different **settings and learning problems** will usually require a different handling from an uncertainty modeling point of view.
- Here, we focus on the standard setting of **supervised learning** and **predictive uncertainty**.



- Assuming probabilistic data generation $\mathbf{P}(\boldsymbol{x}, y) = \mathbf{P}(\boldsymbol{x})\mathbf{P}(y \mid \boldsymbol{x})$, **probabilistic predictors** (estimating $\mathbf{P}(y \mid \boldsymbol{x})$) are natural primitives.

Supervised learning and predictive uncertainty

- A learner is given access to a set of (i.i.d.) **training data**

$$\mathcal{D} := \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\} \subset \mathcal{X} \times \mathcal{Y} \ ,$$

  where $\mathcal{X}$ is an instance space and $\mathcal{Y}$ the set of outcomes.
- Given a **hypothesis space** $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ and a loss function

$$\ell : \mathcal{Y} \times \mathcal{Y} \longrightarrow \mathbb{R} \, ,$$

  the goal of the learner is to induce a hypothesis $h^* \in \mathcal{H}$ with low risk

$$R(h) := \int_{\mathcal{X} \times \mathcal{Y}} \ell(h(\boldsymbol{x}), y) \, d\, \mathbf{P}(\boldsymbol{x}, y) \ .$$

- The learner's choice is commonly guided by the **empirical risk**

$$R_{emp}(h) := \frac{1}{N} \sum_{i=1}^{N} \ell(h(\boldsymbol{x}_i), y_i) \ .$$

Supervised learning and predictive uncertainty

- Yet, since $R_{emp}(h)$ is only an estimation of the true risk $R(h)$, the hypothesis (**empirical risk minimizer**)

$$\hat{h} := \arg\min_{h \in \mathcal{H}} R_{emp}(h)$$

will normally not coincide with the true **risk minimizer**

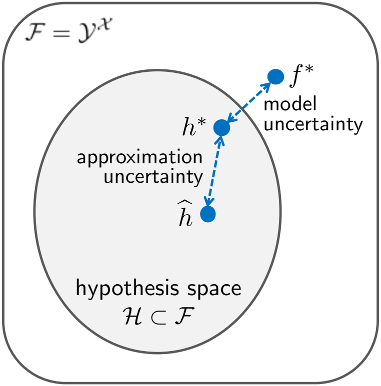
$$h^* := \arg\min_{h \in \mathcal{H}} R(h) \, .$$

- Correspondingly, there remains **uncertainty** regarding h^* as well as the approximation quality of $\hat{h}$ (in the sense of its proximity to h^*) and its true risk $R(\hat{h})$.
- Eventually, one is often interested in the **predictive uncertainty**, i.e., the uncertainty related to the prediction $\hat{y}_q$ for an **individual** (query) instance $\boldsymbol{x}_q \in \mathcal{X}$.

Agenda

1. Introduction
2. **Sources of uncertainty in supervised learning**
3. Modeling approximation uncertainty
4. Ensemble methods for uncertainty quantification
5. Conclusion and outlook

Sources of uncertainty



	point prediction	probability
ground truth	$f^*(\boldsymbol{x})$	$\mathbf{p}(\cdot \mid \boldsymbol{x})$
best possible	$h^*(\boldsymbol{x})$	$\mathbf{p}(\cdot \mid \boldsymbol{x}, h^*)$
induced predictor	$\hat{h}(\boldsymbol{x})$	$\mathbf{p}(\cdot \mid \boldsymbol{x}, \hat{h})$

Sources of uncertainty

- A query instance $\boldsymbol{x}_q$ gives rise to a conditional probability on $\mathcal{Y}$:

$$\mathbf{p}(y \mid \boldsymbol{x}_q) = \frac{\mathbf{p}(\boldsymbol{x}_q, y)}{\mathbf{p}(\boldsymbol{x}_q)}$$

- Thus, even given full information in the form of the measure $\mathbf{P}$ (and its density $\mathbf{p}$), uncertainty about the actual outcome y remains.
- This uncertainty is of an **aleatoric** nature.
- The best point predictions (minimizing expected loss) are prescribed by the **pointwise Bayes predictor** f^*:

$$f^*(\boldsymbol{x}) := \arg\min_{\hat{y} \in \mathcal{Y}} \int_{\mathcal{Y}} \ell(y, \hat{y}) \, d\mathbf{P}(y \mid \boldsymbol{x}) \, .$$

Sources of uncertainty

- The **Bayes predictor** does not necessarily coincide with the pointwise Bayes predictor.
- This discrepancy between h^* and f^* is connected to the uncertainty regarding the **right type of model** to be fit, and hence the choice of the hypothesis space $\mathcal{H}$.
- We shall refer to this uncertainty as **model uncertainty**.
- Due to model uncertainty, one cannot guarantee

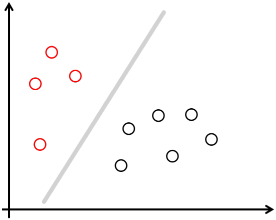
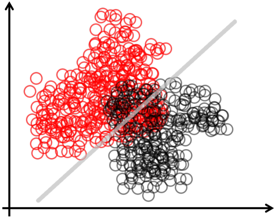
$$h^*(\boldsymbol{x}) = f^*(\boldsymbol{x}),$$

or, in the case of probabilistic predictions $\mathbf{p}(y \,|\, \boldsymbol{x}, h^*)$, that

$$\mathbf{p}(\cdot \,|\, \boldsymbol{x}, h^*) = \mathbf{p}(\cdot \,|\, \boldsymbol{x}).$$

Sources of uncertainty

- Hypothesis $\hat{h}$ produced by the learner is an estimate of h^*.
- The quality of this estimate strongly depends on the quality and the amount of training data.

- We refer to the uncertainty about the discrepancy between $\hat{h}$ and h^* as **approximation uncertainty**.
- Both model and approximation uncertainty are of **epistemic** nature.

Reducible versus irreducible uncertainty

- One way to characterize uncertainty as **aleatoric** or **epistemic** is to ask whether or not it can be reduced through additional information.
- Aleatoric uncertainty refers to the **irreducible** part of the uncertainty, which is due to the stochastic dependency between instances $\boldsymbol{x}$ and outcomes y.

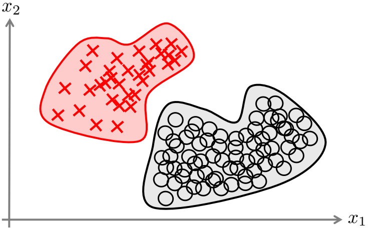
flipping a biased coin

- Model uncertainty and approximation uncertainty are subsumed under the notion of epistemic uncertainty, that is, uncertainty due to a **lack of knowledge** about the perfect predictor.
- In principle, these uncertainties are **reducible**.

Reducible versus irreducible uncertainty

- But what does "reducible" actually mean?
- An obvious source of additional information is the **training data** $\mathcal{D}$: Uncertainty can be reduced by observing more data, ...
- ... while the problem setting $(\mathcal{X}, \mathcal{Y}, \mathcal{H}, \mathbf{P})$ remains fixed.
- In practice, this is of course not always the case.
- For example, a learner may decide to extend the description of instances by **additional features**, thereby replacing the current instance space $\mathcal{X}$ by another space $\mathcal{X}'$.
- Thus, aleatoric and epistemic uncertainty should not be seen as absolute notions. Instead, they are **context-dependent** in the sense of depending on the setting $(\mathcal{X}, \mathcal{Y}, \mathcal{H}, \mathbf{P})$.

Reducible versus irreducible uncertainty

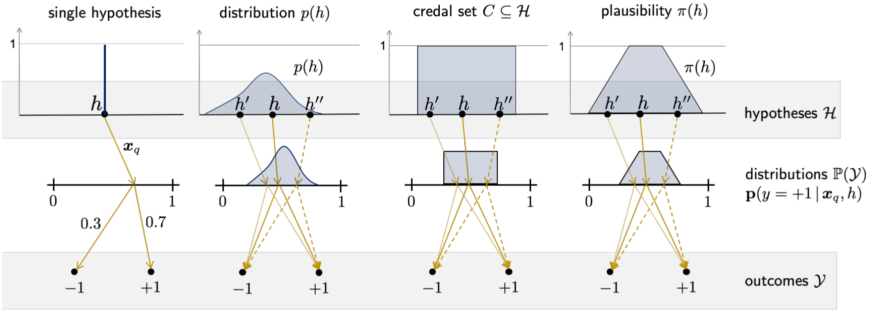


Left: The two classes are overlapping, which causes (aleatoric) uncertainty in a certain region of the instance space. Right: By adding a second feature, and hence embedding the data in a higher-dimensional space, the two classes become separable, and the uncertainty can be resolved.

Agenda

1. Introduction
2. Sources of uncertainty in supervised learning
3. **Modeling approximation uncertainty**
4. Ensemble methods for uncertainty quantification
5. Conclusion and outlook

Approaches for representing uncertainty in ML



probabilistic
single distribution

Bayesian
distribution of distributions

Levi
set of distributions

Gärdefors-Sahlin (GS)
graded set of distributions

Bayesian agents

- An explicit attempt at **uncertainty quantification**, i.e., separating and measuring aleatoric and epistemic uncertainty (in the context of regression with DNNs) is made by Depeweg et al. (2018).
- Here, epistemic uncertainty corresponds to uncertainty about network weights $\boldsymbol{w}$ (playing the role of hypotheses h).
- The idea is to model **epistemic uncertainty** as **mutual information** between outcomes and hypotheses:

$$\underbrace{H[Y]}_{total\ \ uncertainty} = \underbrace{I(Y;W)}_{epistemic} + \underbrace{H[Y \mid W]}_{aleatoric}$$

- Intuitively, epistemic uncertainty thus captures the amount of information about the model parameters $\boldsymbol{w}$ that would be gained through knowledge of the true outcome y.

Bayesian agents

- **Total uncertainty** = entropy of the predictive posterior distribution, in the case of discrete $\mathcal{Y}$ given by

$$H\big[\, \mathbf{p}(y \,|\, \boldsymbol{x}) \,\big] = - \sum_{y \in \mathcal{Y}} \mathbf{p}(y \,|\, \boldsymbol{x}) \log_2 \mathbf{p}(y \,|\, \boldsymbol{x}) \,.$$

- This uncertainty also includes the (epistemic) uncertainty about the network weights $\boldsymbol{w}$, but fixing a set of weights, i.e., considering a distribution $\mathbf{p}(y \,|\, \boldsymbol{w}, \boldsymbol{x})$, removes the epistemic uncertainty.
- Therefore, the expectation over the entropies of these distributions,

$$\begin{aligned} & \mathbf{E}_{p(\boldsymbol{w} \,|\, \mathcal{D})} H\big[\mathbf{p}(y \,|\, \boldsymbol{w}, \boldsymbol{x})\big] = \\ & \qquad = - \int p(\boldsymbol{w} \,|\, \mathcal{D}) \left(\sum_{y \in \mathcal{Y}} \mathbf{p}(y \,|\, \boldsymbol{w}, \boldsymbol{x}) \log_2 \mathbf{p}(y \,|\, \boldsymbol{w}, \boldsymbol{x}) \right) \, d\, \boldsymbol{w} \;\; , \end{aligned}$$

is a measure of the **aleatoric uncertainty** (conditional entropy).

Bayesian agents

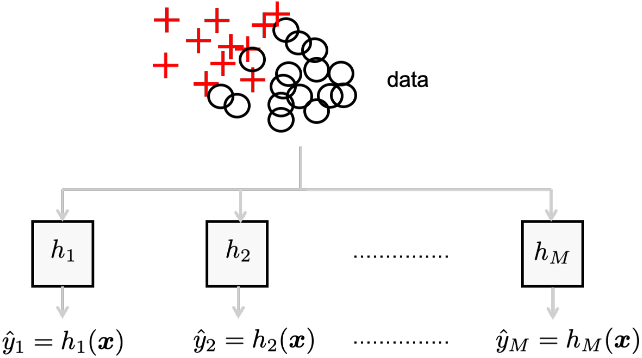
- Finally, the **epistemic uncertainty** is obtained as the difference

$$\mathrm{EU}(\boldsymbol{x}) := H\big[\mathbf{p}(y \,|\, \boldsymbol{x})\big] - \mathbf{E}_{p(\boldsymbol{w} \,|\, \mathcal{D})} H\big[\mathbf{p}(y \,|\, \boldsymbol{w}, \boldsymbol{x})\big] \ ,$$

  which equals the **mutual information** between y and $\boldsymbol{w}$.
- A similar approach was recently adopted by Mobiny et al. (2017).

Agenda

Ensemble methods for uncertainty quantification



- Ensemble can be seen as an approximation of a distribution.
- Intuitively, diversity is an indicator for epistemic uncertainty.

Bayesian agents: Ensemble-based approximation

- Based on an ensemble $H = \{h_1, \ldots, h_M\}$ of hypotheses, an approximation of conditional entropy can be obtained by

$$\mathrm{AU}(\boldsymbol{x}) := -\frac{1}{M} \sum_{i=1}^{M} \sum_{y \in \mathcal{Y}} \mathbf{p}(y \mid h_i, \boldsymbol{x}) \log_2 \mathbf{p}(y \mid h_i, \boldsymbol{x}) \,,$$

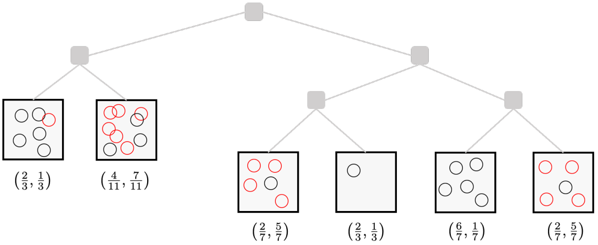
an approximation of total uncertainty (Shannon entropy) by

$$\mathrm{U}(\boldsymbol{x}) := -\sum_{y \in \mathcal{Y}} \underbrace{\left(\frac{1}{M} \sum_{i=1}^{M} \mathbf{p}(y \mid h_i, \boldsymbol{x}) \right)}_{\bar{\mathbf{p}}(y \mid h_i, \boldsymbol{x})} \log_2 \underbrace{\left(\frac{1}{M} \sum_{i=1}^{M} \mathbf{p}(y \mid h_i, \boldsymbol{x}) \right)}_{\bar{\mathbf{p}}(y \mid h_i, \boldsymbol{x})} \,,$$

and an approximation of epistemic uncertainty (mutual information) by the difference, which is equivalent to **Jensen-Shannon divergence** of the distributions $\mathbf{p}(y \mid h_i, \boldsymbol{x})$, $i = 1, \ldots, M$.

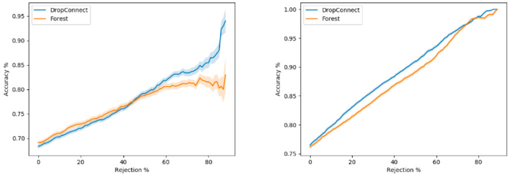
Bayesian agents: Ensemble-based approximation

- For neural networks, it has been shown that techniques such as **Dropout** (Gal and Ghahramani, 2016) and **DropConnect** (Mobiny et al., 2017) can be interpreted as (implicit) ensemble methods, and can hence be used to implement this approach.
- Of course, any other ensemble technique could be used as well.
- We proposed an implementation based on **Random Forests**, using decision trees that predict probabilities in terms of (Laplace-corrected) relative frequencies (Shaker and Hüllermeier, 2020).

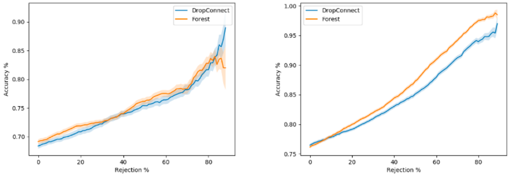


Experimental results

- Quality of uncertainty quantification was evaluated (indirectly) in terms of **accuracy-rejection curves**.
- Results for two approaches, DNN with DropConnect and Random Forests, both for aleatoric (above) and epistemic (below) uncertainty:



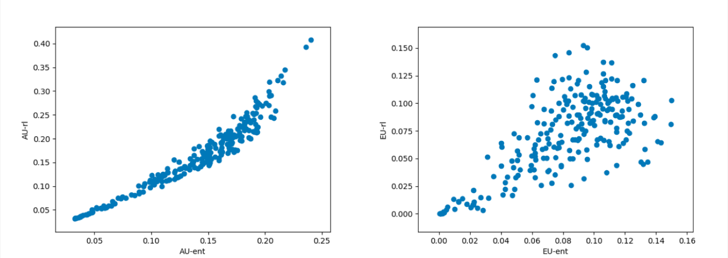
(a) spect (b) diabetes



(c) spect (d) diabetes

Experimental results

- Relationship between uncertainty degrees from neural networks and Random Forests (aleatoric left, epistemic right, diabetes data):



Conclusion and outlook

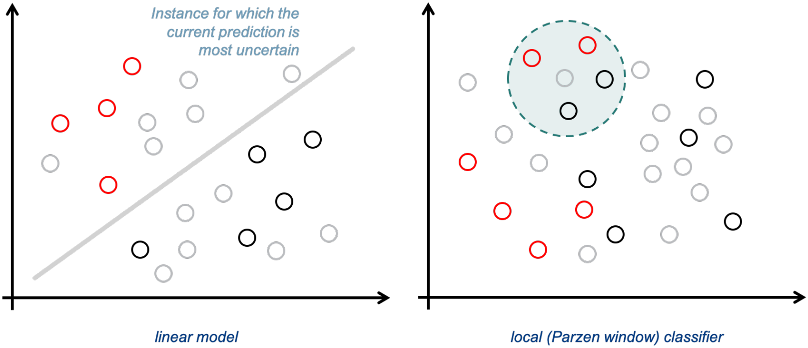
- We highlighted the benefits of distinguishing between different types of uncertainty in ML, notably **aleatoric and epistemic uncertainty**.
- In the setting of **supervised learning**, **aleatoric** uncertainty in a prediction is due to the inherently stochastic dependency between instances and outcomes (and hence **irreducible**).
- **Epistemic** uncertainty is naturally associated with the **lack of knowledge** about the true (or Bayes-optimal) hypothesis.
- In a Bayesian setting, epistemic uncertainty is reflected by the posterior $\mathbf{p}(h \,|\, \mathcal{D})$ on $\mathcal{H}$: The less peaked, the less informed the learner is, the higher its (epistemic) uncertainty.
- We considered an information-theoretic approach to uncertainty quantification and its realization by means of **ensemble learning**.
- Ongoing work on generalizing this for **Levi and GS agents**, building on uncertainty quantification for generalized representations.

Conclusion and outlook

- We also highlighted the fact that **uncertainty quantification**, like any other form of statistical inference, requires a precise specification of the underlying **assumptions**.
- This might be one reason for why **model uncertainty** is so difficult to deal with (and actually neglected by most approaches so far).
- On the other side, **model misspecification** is a common problem in practice, and should therefore not be ignored.
- Indeed, **conflict and inconsistency** can be seen as another source of uncertainty, in addition to randomness and a lack of knowledge.
- In addition to foundational work of that kind, there are many interesting **applications** that can benefit from "uncertainty-informed" decisions ...

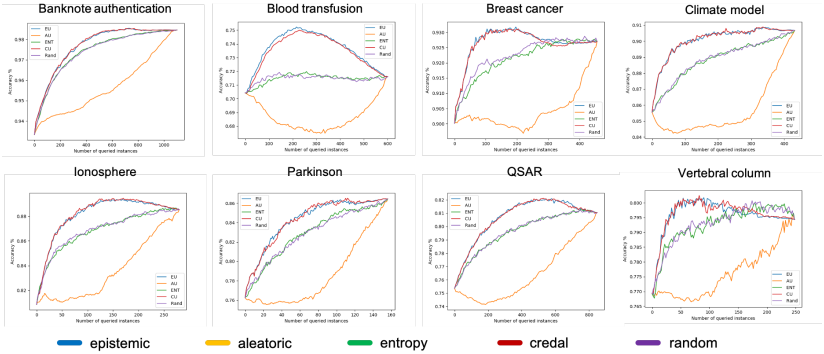
Epistemic uncertainty sampling

- The idea of **epistemic uncertainty sampling** is to use a measure of epistemic (instead of total) uncertainty in uncertainty sampling for active learning (Nguyen et al., 2019).



Epistemic uncertainty sampling

- Performance curves for different uncertainty sampling techniques using decision trees as base learners:



References

Abellan J, Klir J, Moral S (2006) Disaggregated total uncertainty measure for credal sets. International Journal of General Systems 35(1)

Depeweg S, Hernandez-Lobato J, Doshi-Velez F, Udluft S (2018) Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. In: Proc. ICML, Stockholm, Sweden

Der Kiureghian A, Ditlevsen O (2009) Aleatory or epistemic? does it matter? Structural Safety 31:105–112

Gal Y, Ghahramani Z (2016) Bayesian convolutional neural networks with Bernoulli approximate variational inference. In: Proc. of the ICLR Workshop Track

Hora S (1996) Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. Reliability Engineering and System Safety 54(2–3):217–223

Kendall A, Gal Y (2017) What uncertainties do we need in Bayesian deep learning for computer vision? In: Proc. NIPS, pp. 5574–5584

Klir G, Mariano M (1987) On the uniqueness of possibilistic measure of uncertainty and information. Fuzzy Sets and Systems 24(2):197–219

Mobiny A, Nguyen H, Moulik S, Garg N, Wu C (2017) DropConnect is effective in modeling uncertainty of Bayesian networks. CoRR abs/1906.04569

Nguyen V, Destercke S, Hüllermeier E (2019) Epistemic uncertainty sampling. In: Proc. DS 2019, 22nd International Conference on Discovery Science, Split, Croatia

Senge R, Bösner S, Dembczynski K, Haasenritter J, Hirsch O, Donner-Banzhoff N, Hüllermeier E (2014) Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty. Information Sciences 255:16–29

Shaker M, Hüllermeier E (2020) Aleatoric and epistemic uncertainty with random forests. In: Proc. IDA 2020, 18th Int. Symposium on Intelligent Data Analysis, Springer, Konstanz, Germany, pp. 444–456

Tan M, Le Q (2019) EfficientNet: Rethinking model scaling for convolutional neural networks. In: Proc. ICML, Long Beach, California

Zaffalon M (2002) The naive credal classifier. Journal of Statistical Planning and Inference 105(1):5–21

References

arXiv.org > cs > arXiv:1910.09457

Search... All fields Search

Help | Advanced Search

Computer Science > Machine Learning

[Submitted on 21 Oct 2019 (v1), last revised 16 Sep 2020 (this version, v3)]

Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods

Eyke Hüllermeier, Willem Waegeman

The notion of uncertainty is of major importance in machine learning and constitutes a key element of machine learning methodology. In line with the statistical tradition, uncertainty has long been perceived as almost synonymous with standard probability and probabilistic predictions. Yet, due to the steadily increasing relevance of machine learning for practical applications and related issues such as safety requirements, new problems and challenges have recently been identified by machine learning scholars, and these problems may call for new methodological developments. In particular, this includes the importance of distinguishing between (at least) two different types of uncertainty, often referred to as aleatoric and epistemic. In this paper, we provide an introduction to the topic of uncertainty in machine learning as well as an overview of attempts so far at handling uncertainty in general and formalizing this distinction in particular.

Comments: 59 pages
Subjects: **Machine Learning (cs.LG)**; Machine Learning (stat.ML)
Cite as: arXiv:1910.09457 **[cs.LG]**
(or arXiv:1910.09457v3 **[cs.LG]** for this version)

Submission history

From: Eyke Hüllermeier [view email]
[v1] Mon, 21 Oct 2019 15:41:28 UTC (5,161 KB)
[v2] Fri, 3 Apr 2020 14:55:47 UTC (5,546 KB)
[v3] Wed, 16 Sep 2020 06:34:50 UTC (4,713 KB)