Uncertainty Quantification 360: A Hands-on Tutorial

Soumya Ghosh, Q. Vera Liao, Karthikeyan Natesan Ramamurthy, Jiří Navrátil, Prasanna Sattigeri
Kush R. Varshney, Yunfeng Zhang*
IBM Research
USA

ABSTRACT

This tutorial presents an open source Python package (https://github.com/IBM/UQ360) named Uncertainty Quantification 360 (UQ360), a toolkit that provides a broad range of capabilities for quantifying, evaluating, improving, and communicating uncertainty in the AI application development lifecycle. We will first introduce the concepts in uncertainty quantification through an interactive experience (http://uq360.mybluemix.net) followed by use cases with different quantification algorithms and evaluation metrics. The hands-on experience gained from tutorial will aid researchers and developers in producing and evaluating high-quality uncertainties from AI models in an efficient manner.

CCS CONCEPTS

• **Software and its engineering** → *Software libraries and repositories.*

KEYWORDS

AI, Uncertainty Quantification, Trust, Opensource

ACM Reference Format:
Soumya Ghosh, Q. Vera Liao, Karthikeyan Natesan Ramamurthy, Jiří Navrátil, Prasanna Sattigeri and Kush R. Varshney, Yunfeng Zhang. 2022. Uncertainty Quantification 360: A Hands-on Tutorial. In *5th Joint International Conference on Data Science & Management of Data (9th ACM IKDD CODS and 27th COMAD) (CODS-COMAD 2022), January 8–10, 2022, Bangalore, India.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3493700.3493767

1 MOTIVATION AND GOALS

Success stories of AI models are plentiful, but we have also seen prominent examples where the models behave in unexpected ways. For example, a typical failure mode of state-of-the-art prediction models is the inability to abstain from making predictions when the test data violate assumptions made during training, potentially resulting in highly confident but incorrect predictions. Hence, there is a renewed interest in improving the reliability and transparency of AI models [1].

A typical AI lifecycle process consists of collecting data, preprocessing it, selecting a model to learn from the data, choosing a learning algorithm to train the selected model, and performing inference using the learned model. There are inherent uncertainties associated with each of these steps. For example, data uncertainty may arise from the inability to collect or represent real-world data reliably. Flaws in data pre-processing—whether during curation, cleaning, or labeling—also create data uncertainty. Similarly, models are only proxies for the real world and their learning and inference algorithms rely on various simplifying assumptions and thus introduce modeling and inferential uncertainties. The predictions made by an AI system are susceptible to all these sources of uncertainty.

Reliable uncertainty quantification provides a vital diagnostic for both developers and users of an AI system. For developers, it can suggest strategies for improving the system. For example, high data uncertainty may point towards improving the data representation process, while a high model uncertainty may suggest the need to collect more data. For users, accurate uncertainties, especially when combined with effective communication strategies, can add a critical layer of transparency and trust, crucial for better AI-assisted decision making [17]. Trust in AI systems is a necessary condition for their successful deployment in high-stakes applications spanning health care, finance, and the social sciences.

The research on uncertainty quantification (UQ) is long-standing but has enjoyed ever increasing interest in recent years due to the observations made above. Numerous approaches for improved UQ in AI models have been proposed. However, choosing a particular UQ method depends on many factors: the underlying model, type of machine learning task (regression vs. classification), characteristics of the data, and the user's goal. If inappropriately used, a particular UQ method may produce poor uncertainty estimates and mislead users. Moreover, even a highly accurate uncertainty estimate may be misleading if poorly communicated. The main goals of this tutorial are:

- to introduce a diverse set of algorithms to quantify uncertainties, metrics to measure them, methods to improve the quality of estimated uncertainties, and approaches to communicate the uncertainties effectively;
- to discuss a taxonomy and guidance for choosing these capabilities based on the user's needs; and
- to encourage further exploration of connections between UQ and other pillars of trustworthy AI such as fairness and transparency.

2 DESCRIPTION

This 2 hour long tutorial is organized in 3 parts. We will begin with a gentle introduction to the concepts in uncertainty quantification and evaluation. We will walk through an interactive experience that will demonstrate these concepts with a use-case. This will be

*The work was done while the author was at IBM Research. Now the author is with Twitter.

CODS-COMAD 2022, January 8–10, 2022, Bangalore, India

ACM ISBN 978-1-4503-8582-4/22/01…$15.00
https://doi.org/10.1145/3493700.3493767

followed by deep-dives into different types of uncertainty quantification and metrics. We discuss below the main components of the tutorial.

2.1 UQ Algorithms

UQ algorithms can be broadly classified as intrinsic or extrinsic depending on how the uncertainties are obtained from the AI models. We refer to methods as intrinsic if they were explicitly designed to produce uncertainties along with predictions. This includes variational Bayesian neural networks (BNNs) [2], Gaussian processes [14], quantile regression [9] and hetero/homo-scedastic neural networks [8] which are models that fall in this category. We will discuss the Horseshoe BNNs [7] that use sparsity promoting priors and can lead to better-calibrated uncertainties, especially in the small data regime.

For prediction algorithms that do not have an inherent notion of uncertainty built into them, we use *extrinsic* approaches to extract uncertainties post-hoc. We will present meta-models [3] that generate reliable confidence measures (in classification), prediction intervals (in regression) [11], and predict performance metrics such as accuracy on unseen and unlabeled data [4]. We will also present the Infinitesimal Jackknife (IJ) based algorithm [6], provided in the toolkit. This perturbation-based approach performs uncertainty quantification by estimating model parameters under different perturbations of the original data while only requiring the model to be trained once on the unperturbed dataset.

2.2 Evaluation Metrics

The quality of estimation generated by a UQ algorithm also needs to be evaluated. A poorly calibrated UQ estimation should not be trusted nor should it be presented to a user. We will discuss the standard calibration metrics for classification and regression tasks. For classification, these include Expected Calibration Error [10], Brier Score, etc. Regression metrics include Prediction Interval Coverage Probability (PICP) and Mean Prediction Interval Width (MPIW) among others. We will present a novel operation-point agnostic approach for the assessment of prediction uncertainty estimates called the Uncertainty Characteristic Curve (UCC) [12].

We will also present algorithms for improving the uncertainty estimation quality. This includes isotonic regression [16], Platt-scaling [13], auxiliary interval predictors [15], and UCC Recalibration [12].

2.3 Toolkit Architecture

The toolkit has been engineered with a common interface for all of the different UQ capabilities and is extensible to accelerate innovation by the community advancing trustworthy and responsible AI. We will discuss the architecture of the toolkit that is designed to be scikit-learn compatible so that they can fit into developers' existing workflow. We will present ways to score and select models using uncertainty quality using *Lale*, an AutoAI framework.

2.4 UQ Visualization and Communication

We will discuss the choice of styles of communication methods, from concise descriptions to detailed visualizations. We will provide guidance for communicating UQ to help practitioners make the choice, as informed by psychology and human-computer interaction research. For classification tasks, UQ360 provides functions to generate confidence scores. For regression tasks, UQ360 provides functions to generate the numerical ranges, visual confidence intervals, density plots, and quantile dot plots [5]. We will discuss several metrics and diagnosis tools such as reliability diagram and risk-vs-rejection rate curves which also support analysis by subgroups in the population to study fairness implications of acting on given uncertainty estimates.

3 TARGET AUDIENCE

The tutorial is suitable for wide range of audience with different backgrounds. The interactive experience is most useful for consumers of AI predictions or decision makers which provides introductions to the concepts and expounds on UQ communication techniques. The hands on tutorials notebooks provides in depth look into the toolkit capabilities. We believe the UQ360 toolkit streamlines and fosters the common practices of quantifying, evaluating, improving, and communicating uncertainty in the AI lifecycle. This tutorial will be a helpful demonstration of that.

4 PRESENTERS

The presenters will be Prasanna Sattigeri, Soumya Ghosh and Jiří Navrátil. The team involved in preparing this tutorial additionally includes Q. Vera Liao, Karthikeyan Natesan Ramamurthy, Kush R. Varshney and Yunfeng Zhang.

Prasanna Sattigeri is a Research Staff Member at IBM Research. His research interests include Bayesian inference, deep generative modeling, uncertainty quantification, and related subareas of machine learning and AI. His current work focuses on developing theory and practical systems for machine learning applications that demand constraints such as reliability, fairness, and interpretability. He is core-contributor to several open source trustworthy AI toolkits - AI Fairness 360, AI Explainability 360 and Uncertainty Quantification 360.

Soumya Ghosh is a Research Staff Member at IBM Research. His research focuses on the design of flexible statistical models and efficient inference algorithms for reasoning about noisy, high-dimensional data. His recent work has examined approaches for combining the complementary strengths of Bayesian methods, graphical models, and deep neural networks for modeling the progression of neurodegenerative diseases and procedures for assessing the sensitivity and robustness of inferences drawn from such statistical methods.

Jiří Navrátil is a Principal Research Staff Member at IBM Research. He is currently involved in research efforts in the area of trusted AI and machine learning for industrial applications. Previously at IBM, Jiří worked on statistical question answering, machine translation, voice-based authentication, and spoken language recognition.

REFERENCES

[1] Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q. Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Gauthier Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, Lama Nachman, Rumi Chunara, Madhulika Srikumar, Adrian Weller, and Alice Xiang. 2021. Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty. In *Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society.*

[2] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. 2015. Weight Uncertainty in Neural Network. In *Proceedings of the 32nd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 37)*, Francis Bach and David Blei (Eds.). PMLR, Lille, France, 1613–1622. http://proceedings.mlr.press/v37/blundell15.html

[3] Tongfei Chen, Jirí Navrátil, Vijay Iyengar, and Karthikeyan Shanmugam. 2019. Confidence scoring using whitebox meta-models with linear classifier probes. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*. 1467–1475.

[4] Benjamin Elder, Matthew Arnold, Anupama Murthi, and Jiri Navratil. 2021. Learning Prediction Intervals for Model Performance. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[5] Michael Fernandes, Logan Walls, Sean Munson, Jessica Hullman, and Matthew Kay. 2018. Uncertainty displays using quantile dotplots or cdfs improve transit decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 144.

[6] Soumya Ghosh, William T. Stephenson, Tin D. Nguyen, Sameer Deshpande, and Tamara Broderick. 2020. Approximate Cross-Validation for Structured Models. In *Advances in Neural Information Processing Systems*, Vol. 33.

[7] Soumya Ghosh, Jiayu Yao, and Finale Doshi-Velez. 2019. Model Selection in Bayesian Neural Networks via Horseshoe Priors. *Journal of Machine Learning Research* 20, 182 (2019), 1–46.

[8] Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in Bayesian deep learning for computer vision?. In *Advances in Neural Information Processing Systems*, Vol. 30. 5580–5590.

[9] Roger Koenker and Gilbert Bassett, Jr. 1978. Regression Quantiles. *Econometrica* 46, 1 (Jan. 1978), 33–50.

[10] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 29.

[11] Jiri Navratil, Matthew Arnold, and Benjamin Elder. 2020. Uncertainty Prediction for Deep Sequential Regression Using Meta Models. arXiv:2007.01350 [cs.LG]

[12] Jiri Navratil, Benjamin Elder, Matthew Arnold, Soumya Ghosh, and Prasanna Sattigeri. 2021. Uncertainty Characteristics Curves: A Systematic Assessment of Prediction Intervals. arXiv:2106.00858 [cs.LG]

[13] John C. Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*. MIT Press, Cambridge, MA, 61–74.

[14] Carl Edward Rasmussen and Christopher K. I. Williams. 2006. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA.

[15] Jayaraman J Thiagarajan, Bindya Venkatesh, Prasanna Sattigeri, and Peer-Timo Bremer. 2020. Building calibrated deep models via uncertainty matching with auxiliary interval predictors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 6005–6012.

[16] Bianca Zadrozny and Charles Elkan. 2001. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In *Proceedings of the International Conference on Machine Learning*. 609–616.

[17] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 295–305.