How Sure are You?

Ulf Johansson[1] and Cecilia Sönströd[2]

[1] Dept. of Computer Science and Informatics, Jönköping University, Sweden
ulf.johansson@ju.se
[2] Dept. of Information Technology, University of Borås, Sweden
cecilia.sonstrod@hb.se

Abstract. The purpose of this extended abstract is two-fold: (i) to propose that FAT and XAI should include algorithmic confidence as a requirement for accountable and explainable AI and (ii) to argue that conformal prediction and Venn prediction, with their strong validity guarantees, are very good tools for building modules that measure and communicate algorithmic confidence.

1 Background

As AI is increasingly used not only for decision support, but also automated decision making, trust in the resulting decisions or recommendations becomes vital. Consequently, how to make AI solutions trustworthy is today a key question addressed by researchers from many disciplines. The importance of trust in AI is also strongly manifested in the two vibrant areas Explainable AI (XAI) and Fairness, Accountability and Transparency (FAT).

Associations such as the ACM [1], FAT/ML [3] and IEEE [4] have proposed guidelines and frameworks for FAT/XAI, incorporating demands to be placed on AI solutions, as well as evaluation criteria for explainability and FAT. Humans interacting with AI need to be able to make informed judgments about when to trust the system. This is, of course, nothing new, but has been present in the AI discourse since the era of expert systems. In fact, we argue that modern AI research should have learned more from the expert systems period. More specifically, accountability was studied intensively, and the "how" and "why" questions of expert systems, are now the basis for XAI. Furthermore, the inability to handle uncertainties correctly, which was one of the main issues for expert systems, is a problem also for modern machine learning.

Representing a major research initiative within FATE/XAI, the DARPA Research Programme on Explainable AI [2], is aimed at developing new AI solutions that produce "more explainable models, while maintaining a high level of learning performance (e.g., prediction accuracy)". Similarly, The FAT/ML Principles for Accountable Algorithms and a Social Impact Statement for Algorithms [3], list *responsibility*, *explainability*, *accuracy* and *auditability* as the components of an accountable algorithm. One guiding question here is: "How confident are the decisions output by your system?" Thus, demands on explainability and accountability must include some capacity for reporting uncertainty. In fact, algorithmic ability to reason about its own competence, specifically about confidence in individual recommendations, is vital. With this in mind, an explainable algorithm should be able to assess and clearly communicate a confidence measure for each prediction or recommendation, that can be easily understood by a human user. Consequently, we propose that the questions an algorithm should be able to answer are amended to include “How sure are you?” Unfortunately, this is a surprisingly hard question for current machine learning systems!

In predictive regression, the models typically lack the ability to quantify their confidence in individual predictions. Similarly, in classification, probabilistic predictors are notorious for being poorly calibrated, in effect often becoming outright misleading.

2 Prediction with confidence

The framework *Prediction with confidence*, in our opinion, equips any existing predictive model with the ability to answer the “How sure are you?” question. Simply put, this is achieved by turning the model into a *confidence predictor* through calibration on a data set that must be i.i.d. with future test instances.

In regression, *conformal prediction* [5] enables the user to specify a desired significance level ϵ and the prediction intervals produced by the algorithm are mathematically guaranteed to contain the true target with the probability $1 - \epsilon$. Consequently, it is possible to choose a suitable level of confidence based on, e.g., legal, ethical or financial constraints of a particular task. Furthermore, in many situations where a decision is based on predictive modeling, using the two endpoints of the interval will produce very robust worst-cases and best-cases.

In classification, *Venn predictors* [5], will produce probabilistic predictions that are automatically perfectly calibrated, even in the small. Obviously, well-calibrated probabilistic predictors are very strong tools for automated decision making or decision support. Specifically, if combined with utilities, such predictors will produce optimal decisions, according to Bayesian decision theory.

References

1. ACM: New statement on algorithmic transparency and accountability by ACM U.S. Public Policy Council (2017), https://techpolicy.acm.org/
2. DARPA: Explainable Artificial Intelligence (XAI) (2016), https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf
3. FAT/ML: Principles for accountable algorithms and a social impact statement for algorithms (2017), http://www.fatml.org/resources/principles-for-accountable-algorithms
4. IEEE: The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems (2017), http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html
5. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic Learning in a Random World. Springer-Verlag New York, Inc. (2005)