Deep Kernel Aalen-Johansen Estimator: An Interpretable and Flexible Neural Net Framework for Competing Risks

Xiaobin Shen[*], **George H. Chen**[*,†]

Heinz College of Information Systems and Public Policy, Carnegie Mellon University

** equal contribution*

† presenting author

ABSTRACT

We propose an interpretable deep competing risks model called the Deep Kernel Aalen-Johansen (DKAJ) estimator, which generalizes the classical Aalen-Johansen nonparametric estimate of cumulative incidence functions (CIFs). Each data point (e.g., patient) is represented as a weighted combination of clusters. If a data point has nonzero weight only for one cluster, then its predicted CIFs correspond to those of the classical Aalen-Johansen estimator restricted to data points from that cluster. These weights come from an automatically learned kernel function that measures how similar any two data points are. On four standard competing risks datasets, we show that DKAJ is competitive with state-of-the-art baselines (that are not interpretable) while being able to provide visualizations to assist model interpretation.

Keywords: survival analysis, competing risks, neural networks, interpretability

Quantifying Uncertainty in Answers from any Language Model and Enhancing their Trustworthiness

Jonas Mueller

Cleanlab

ABSTRACT

We introduce methods for detecting bad and speculative answers from a pretrained Large Language Model by estimating a numeric confidence score for any output it generated. Our uncertainty quantification techniques work for any LLM accessible only via a black-box API, whose training data remains unknown. Experiments on both closed and open-form Question-Answer benchmarks reveal that our approach more accurately identifies incorrect LLM responses than alternative uncertainty estimation procedures (across many frontier models). By sampling multiple responses from the LLM and considering the one with the highest confidence score, we can additionally obtain more accurate responses from the same LLM, without any extra training steps.

Keywords: Large Language Model; Uncertainty Quantification; Trustworthy AI

Accessing the Impact of Data Alteration: When Can R-Squared from Synthetic (or Corrupted) Data Be Trusted?

Xiao-Li Meng, James Bailie, Mohammad Kakooei, Adel Daoud

Department of Statistics, Harvard University

ABSTRACT

Motivated by a seemingly counterintuitive finding in a poverty study that linked privacy-protected household data with satellite imagery via deep learning, we set out—and invite readers to join—to explore the mathematical relationships between an algorithm's performance on two data sets: the original, intended data and an altered version of it. We propose a comparative framework that is agnostic to the alteration mechanism, whether the modification reflects beneficent intent (e.g., synthetic data for privacy protection) or maleficent intervention (e.g., corruption or distortion). This agnostic stance directs attention to the inherent complexity of a "three-body" problem: the relationships between the unaltered and altered data sets, and between how an algorithm interacts with each. To analyze this complexity in a general and coherent way, we introduce the neologism *data alterity*—a notion extending beyond syntheticity—and decompose it qualitatively into *target alterity* and *residual alterity* for a given predictive algorithm and evaluative metric.

The commonly used metric R-squared provides a low-hanging fruit for demonstrating the potential of this framework, through an algebraic relationship linking the unaltered R-squared, the altered R-squared, and an alterity-impact score. This relationship reveals how alterity can help or harm performance by affecting target and residual components differently, regardless of whether the alteration is beneficent or maleficent. We further show that the altered R-squared is typically conservative when alteration behaves like independent noise added to the target variable, as in differential privacy, and we provide a computable adjustment to correct for this case. We also identify a necessary and sufficient condition under which residual alterity exceeds one, implying conservativeness of the altered R-squared whenever the target variable remains unaltered. These initial insights are illustrated in a proxy study predicting ground-level features from Earth observations whose locations have been synthetically perturbed for privacy protection.

Keywords: Alterity-Impact Score; Data Alterity; Data Privacy; Deep Learning; Digital Twins; Residual Alterity; Target Alterity.