Conformal Prediction with Missing Values

Margaux Zaffran*[1,2,3], Aymeric Dieuleveut[3], Julie Josse[2], and Yaniv Romano[4]

[1]EDF – EDF Recherche et Développement – France
[2]Inria – L'Institut National de Recherche en Informatique et e n Automatique (INRIA) – France
[3]Centre de Mathématiques Appliquées - Ecole Polytechnique – Ecole Polytechnique, Centre National de la Recherche Scientifique – France
[4]Departments of Computer Science and Electrical Engineering - Technion, Israel Institute of Technology – Israel

Abstract

Conformal prediction is a theoretically grounded framework for constructing predictive intervals. We study conformal prediction with missing values in the covariates – a setting that brings new challenges to uncertainty quantification. We first show that the marginal coverage guarantee of conformal prediction holds on imputed data for any missingness distribution and almost all imputation functions. However, we emphasize that the average coverage varies depending on the pattern of missing values: conformal methods tend to construct prediction intervals that under-cover the response conditionally to some missing patterns. This motivates our novel generalized conformalized quantile regression framework, missing data augmentation, which yields prediction intervals that are valid conditionally to the patterns of missing values, despite their exponential number. We then show that a universally consistent quantile regression algorithm trained on the imputed data is Bayes optimal for the pinball risk, thus achieving valid coverage conditionally to any given data point. Moreover, we examine the case of a linear model, which demonstrates the importance of our proposal in overcoming the heteroskedasticity induced by missing values. Using synthetic and data from critical care, we corroborate our theory and report improved performance of our methods.

Keywords: Missing Values, Conformal Prediction, Distribution, Free Uncertainty Quantification

*Speaker