Model interpretability in Bayesian framework

Homayun Afrabandpey

Nokia Technologies, Finland, homayun.afrabandpey@nokia.come

A salient approach to obtain interpretability in the Bayesian framework is to restrict the prior to favor interpretable models. This approach, however has shortcomings including difficulties in formulating interpretability prior for certain classes of models such as neural networks and obtaining good trade-offs between accuracy and interpretability. In this talk, I will explain our recent approach towards interpreting Bayesian predictive models through a decision theoretic approach without constraining priors. We developed a two-step strategy to interpretability in the Bayesian framework. In the first step, we fit a highly accurate Bayesian predictive model, which we call reference model, to the training data without constraining it to be interpretable. In the second phase, we construct an interpretable proxy model that best describes locally and/or globally the behavior of the reference model. In our experiments, we show that for the same level of interpretability, our approach finds better trade-off and generates more accurate models than the earlier alternative of restricting the prior.

This is a joint work with Tomi Peltola, Juho Piironen, Aki Vehtari and Samuel Kaski, available in https://link.springer.com/article/10.1007/s10994-020-05901-8.