Probabilistic supervised learning, uncertainty prediction, and predictive inference

Dr. Franz J. Kiraly
(University College London)

May 16, 2018

Suppose we would like to predict a patient's time to recovery, the weather tomorrow, a manufacturing processes' efficiency, the score in a basketball game, or the return on an investment. In all these domain applications, it is of crucial practical importance not only to have a good estimate of the average, but also of likely value ranges in which the observation falls.
While in machine learning, a wealth of supervised learning methods produce a prediction, these usually do not come with predictive intervals.
And while Bayesian approaches do produce a predictive (posterior) distribution, off-shelf methods will not necessarily yield a good proxy for the observed spread (e.g., a "belief" distribution of predictive average, or of homoscedastic variance).
In this talk, we present an overarching statistical and methodological framework for supervised learning methods which produce such a distribution, and discuss some important existing approaches to the problem from this general viewpoint.
We further present a number of theoretical observations in the general set-up, including results on predictive inference which establish parallels between the tasks of predictive model selection, estimation of entropy, and conditional independence testing. On the practical side, we present a number of composition strategies which allow the leveraging of classical supervised learning algorithms, as well as advanced meta-modelling strategies for the probabilistic case.
Furthermore, we discuss how frequentist, Bayesian, and other approaches to produce the uncertainty prediction may be considered on equal footing in the given framework, and fairly compared to each other. Finally, we briefly showcase a new python package, skpro, which provides a unified interface for the probabilistic supervised learning workflow (https://arxiv.org/abs/1801.00753).