Design-based ensemble learning for individual prediction

L.-C. Zhang[1] and D. Lee[2]

[1] University of Southampton, UK and Statistics Norway, Norway
e-mail: L.Zhang@soton.ac.uk

[2] University of Alabama, US
e-mail: dlee84@cba.ua.edu

Abstract

Valid inference of the unobserved individual prediction errors is a fundamental issue to supervised machine learning, no matter how confident one is about the obtained predictor. An independent and identically distributed model of the prediction errors is commonly assumed for algorithm-based learning, such as random forest, support vector machine or neural network, which could be misleading in situations where the available observations are not obtained in a completely random fashion.

Survey sampling has a long tradition for estimating various aggregates of a given population. The inference of the associated uncertainty is based on the known samling design by which the sample of observations are obtained, "irrespectively of the unknown properties of the target population studied" (Neyman, 1934). But there has never been a design-based theory for prediction at the individual level.

We shall define and develop for the first time a general design-based approach to *individual prediction*, given the sampling design and the sample-splitting design for cross-validation, while the outcomes and features are treated as constants associated with the given population. Whether the predictor for the out-of-sample units is selected from an ensemble of models or a weighted average of them, the proposed approach can provide valid inference of the associated risk with respect to the known sampling design.

Keywords: Probability sampling; Ensemble learning; Rao-Blackwellisation.

References

Neyman, J. (1934) On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97:558-606.