What's a good imputation to predict with missing values?

Marine Le Morvan*[1]

[1]Le Morvan Marine – L'Institut National de Recherche en Informatique et e n Automatique (INRIA) – France

Abstract

How to learn a good predictor on data with missing values? Most efforts focus on first imputing as well as possible and second learning on the completed data to predict the outcome. Yet, this widespread practice has no theoretical grounding. Here we show that for almost all imputation functions, an impute-then-regress procedure with a powerful learner is Bayes optimal. This result holds for all missing-values mechanisms, in contrast with the classic statistical results that require missing-at-random settings to use imputation in probabilistic modeling. Moreover, it implies that perfect conditional imputation is not needed for good prediction asymptotically. All imputation functions are not equal though. Some can lead to harder learning problems, suggesting that it is beneficial to adapt the imputation function to the prediction function and vice versa. We propose such a procedure by relying on NeuMiss, a principled neural network capturing the conditional links across observed and unobserved variables whatever the missing-value pattern. Its originality and strength comes from the use of a new type of non-linearity: the multiplication by the missingness indicator. Simulations confirm that joint imputation and regression through NeuMiss is better than various two step procedures with finite number of samples, including in difficult MNAR settings such as self-masking.

*Speaker