On the consistency of supervised learning with missing values

Julie Josse (CMAP, XPOP), Nicolas Prost (CMAP, XPOP, PARIETAL), Erwan Scornet (X), Gaël Varoquaux (PARIETAL)

In many application settings, the data have missing features which make data analysis challenging. An abundant literature addresses missing data in an inferential framework: estimating parameters and their variance from incomplete tables. Here, we consider supervised-learning settings: predicting a target when missing values appear in both training and testing data. We show the consistency of two approaches in prediction. A striking result is that the widely-used method of imputing with the mean prior to learning is consistent when missing values are not informative. This contrasts with inferential settings where mean imputation is pointed at for distorting the distribution of the data. That such a simple approach can be consistent is important in practice. We also show that a predictor suited for complete observations can predict optimally on incomplete data, through multiple imputation. We analyze further decision trees. These can naturally tackle empirical risk minimization with missing values, due to their ability to handle the half-discrete nature of incomplete variables. After comparing theoretically and empirically different missing values strategies in trees, we recommend using the ``missing incorporated in attribute" method as it can handle both non-informative and informative missing values.