On the consistency of supervised learning with missing values

Nicolas Prost CMAP, Ecole Polytechnique
Parietal, INRIA

Abstract

In many application settings, the data are plagued with missing features. These hinder data analysis. An abundant literature addresses missing values in an inferential framework, where the aim is to estimate parameters and their variance from incomplete tables. Here, we consider supervised-learning settings where the objective is to best predict a target when missing values appear in both training and test sets. We analyze which missing-values strategies lead to good prediction. We show the consistency of two approaches to estimating the prediction function. The most striking one shows that the widely-used mean imputation prior to learning method is consistent when missing values are not informative. This is in contrast with inferential settings as mean imputation is known to have serious drawbacks in terms of deformation of the joint and marginal distribution of the data. That such a simple approach can be consistent has important consequences in practice. This result holds asymptotically when the learning algorithm is consistent in itself. We contribute additional analysis on decision trees as they can naturally tackle empirical risk minimization with missing values. This is due to their ability to handle the half-discreteness of variables with missing values. After comparing theoretically and empirically different missing-values strategies in trees, we recommend using the missing incorporated in attributes method as it can handle both non-informative and informative missing values.