Missing data imputation and prediction

Antoine Ogier
Supervisor: Julie Josse
Academic supervisor: Geoff Nicholls

June 13, 2018

Contents

Acknowledgements

Introduction

Missing values in data is a prominent issue that has been much discussed in statistical literature. In the eighties, Donald Rubin devised many of the tools that are still used today to handle missing data: the expectation-maximisation algorithm, the definition of the three missing data patterns (MCAR, MAR, MNAR), multiple imputation. Two main approaches have been extensively researched to handle missing data: developing an ad-hoc algorithm that is capable of handling missing data itself (this is often based on the EM algorithm); or filling in the missing observations with imputed values.

However, there is in this regard a significant gap between the fields of statistical inference and machine learning: while the former has been actively developing and evaluating methods to handle missing data — mostly for parameter and confidence intervals estimation —, these methods are rarely used in the context of prediction, where replacing missing values with the mean of the observed data is standard practice.

This can be explained by the fact that machine learning practitioners enjoy having access to the whole range of standard algorithms they are familiar with, and are thus reluctant to use algorithms made for missing data rather than fill in the values. Conversely, filling by the mean is generally considered as 'good enough' for prediction purposes. Astonishingly in this regard, there is no thorough assessment of the way that missing values impact prediction performance, or comparison of imputation methods in this regard: whenever a new imputation method is published, a comparison with existing methods is usually conducted, but only regarding its performance for statistical inference. This means that we currently do not know whether it is worthwhile, when working on prediction, to turn to more elaborate (but also more computationally intensive) methods than mean imputation.

The goal of this work is to lay the groundwork for a review of imputation methods in the context of predictions. The final goal is twofold:

- Compare the predictive performance of some machine learning algorithms applied to datasets filled in with various imputation methods. This will be done both on real-world datasets and simulated ones with different missingness patterns.

- Investigate the relevance of multiple imputation methods (where multiple possible values of each missing observation are imputed) for prediction: in theory, having multiple imputed datasets gives us more

information on the certainty of the imputation (and so, of the resulting prediction).

However, conducting this investigation raises a major methodological issue related to the way that current imputation methods are implemented. We will start by addressing this issue before moving on to our investigation in itself.

1 Methodological investigation: the train-validation split

1.1 The issue with current imputation methods

1.2 An imputation method compliant with ML methodology

2 The impact of missing values on prediction

2.1 Presentation of some imputation methods

2.2 Is the mean good enough?

2.3 Application: the Traumabase data

3 Multiple imputation: uses in prediction

3.1 Presentation

3.2 Aggregating predictions: principle and performance

3.3 Application the Traumabase data