k-Nearest Neighbor Imputation Under Monotonicity Constraints

TU/e supervisor: dr. W. Duivesteijn
Additional supervision by: R. M. Schouten, TU/e

Introduction

Imputation is the act of filling in the missing values in a dataset. Traditional, statistical solutions to imputation delineate three specific underlying mechanisms for what caused the missingness: MCAR, MAR, and MNAR. The correct choice of imputation method hinges on which of these three situations we find ourselves in. However, recently, machine learning based approaches to imputation has become popular. In this project, we concern ourselves with k-Nearest Neighbor imputation [1], the use of a k-NN classifier to impute missing values.

It has been shown [2] that k-NN classifiers can benefit from properly modeling and handling monotonicity constraints. Often we have domain knowledge about our datasets that comes in the form of such constraints: if a bank considers two applicants A and B for a loan, who are identical except that B has a higher income, it would be profoundly strange if A would get the loan and B would not. This would be a violation of a monotonicity constraint, and the paper [2] shows how k-NN classifiers can be adapted to satisfy such constraints.

It stands to reason that k-NN imputation of missing data could also benefit from adhering to monotonicity constraints. In this project we will explore this idea, along with the impact that the three classical missingness mechanisms have on the performance of k-NN imputation.

Current Status

Nothing more exists about this project than the idea described in this document. We're in the pioneering stage; boldly go where noone has gone before!

Requirements

Missing data mechanisms have a strong presence in the statistics community; you will need some affinity with statistics while doing the literature review for this project. When done well, we expect this project to lead to a publication at a top-level data mining or machine learning conference.

References

[1] L. Beretta, A. Santaniello: Nearest neighbor imputation algorithms: a critical evaluation. BMC Medical Informatics and Decision Making 16, article number 74, 2016.

[2] W. Duivesteijn, A. Feelders: Nearest Neighbour Classification with Monotonicity Constraints. ECML/PKDD (1) 2008: 301–316