Filling Survey Missing Data With Generative Adversarial Imputation Networks

CYVG

09 December 2020

Abstract

In this paper we propose an alternative method for dealing with missing values on survey data sets. The method leverages two known techniques, categorical encoding and generative adversarial imputation networks. We test this approach on the "Kaggle Data Science Survey" and the "Stack Overflow Annual Developer Survey", we experiment with different proportions of missing values and sample sizes and find the technique to yield high quality data imputations.

Keywords: Generative Adversarial Networks; Imputation Algorithms; Surveys

1. Introduction

Survey response rates have seen a decline in recent years and more often than not we see incomplete survey data sets. Often this is because survey designers give the option to skip a question, or simply a respondent decides to not finish the survey. In the literature this is known as data missing completely at random (MCAR) because in most cases there is no dependency on any of the variables. This pervasive problem has also been the cause for multiple solutions to emerge. An imputation algorithm, for example, can be used to estimate missing values based on data that was observed/measured. A substantial amount of research has been dedicated to developing imputation algorithms for medical data but it is also commonly used in image concealment, data compression, and counterfactual estimation.

Often imputation algorithms work very well when the data is continuous, and or contains a mixture of continuous and categorical responses, however it is common to only observe categorical and text responses in survey data sets. Text data is an almost impossible problem to solve because we can't just simply create an algorithm that will write an opinion on behalf of another person. There are both ethical and technical problems associated. Categorical responses on the other hand are simpler to use because having a finite amount of categories allows us to encoded the data. The most popular encoding technique is known in the statistics literature as dummy variable encoding or in the computer science and machine learning literature as one-hot encoding. This popular technique also comes with its limitations since a substantial amount of information is lost by turning variables into vectors of 0 and 1s.

Moreover this technique requires us to increase the dimensions of our data set which results in a loss of computational efficiency.

Hence, we address the problem of data imputation when the data set consists of only categorical variables, and in particular when the data comes from a survey. In this paper we propose an alternative method for data imputation in survey data which comprises of combining two known methods, categorical encoding and a state of the art imputation algorithm. Specifically, we encode categorical variables with a technique based on the weight of evidence (WOE) method and then use the imputation algorithm known as generative adversarial imputation networks (GAIN) to fill missing values.

The paper is divided in the following manner, first in section 2 we elaborate on generative adversarial imputation networks and how they are applied in this context by discussing the proposed encoding technique based on the weight of evidence method. In section 3 we discuss the experiments we conducted on the "*Kaggle Data Science Survey*" and the "*Stack Overflow Annual Developer Survey*" to test the effectiveness of this method. In this section we comment on the surveys, network architectures and hyperparameters and our empirical results. Finally, in the last section we comment on our results and the implications, how this method could be applied in practice, limitations and areas of future work.

2. Survey Generative Adversarial Imputation Networks

2.1 Generative Adversarial Imputation Networks

2.1.1 Data Imputation Problem

2.1.2 GAIN Methodology

2.2 Variable Encoding

2.3 GAINs on Survey Data

3. Experiment

3.1 Data

3.2 Pre-processing and Network Architecture

3.3 Results

4. Discussion

References

Bhalla, Deepanshu. 2015. “Weight of Evidence (Woe) and Information Value (Iv) Explained.” https://www.listendata.com/2015/03/weight-of-evidence-woe-and-information.html.

Goodfellow, Ian J., Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. “Generative Adversarial Networks.” http://arxiv.org/abs/1406.2661.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D’Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. “Welcome to the tidyverse.” *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Xie, Yihui. 2020. *Knitr: A General-Purpose Package for Dynamic Report Generation in R*. https://yihui.org/knitr/.

Yoon, Jinsung, James Jordon, and Mihaela van der Schaar. 2018. “GAIN: Missing Data Imputation Using Generative Adversarial Nets.” http://arxiv.org/abs/1806.02920.