CSE 4020 – MACHINE LEARNIG LAB
FACULTY : PROF. SUGANYA G
DHRUBANKA DUTTA – 17BCE1019
L – 11,12 , 10.9.19

Handling missing values

1. Abstract

Brief explanation of the concept

In statistics, imputation is the process of replacing missing data with substituted values. Data in real world are rarely clean and homogeneous. Typically, they tend to be incomplete, noisy, and inconsistent and it is an important task of a Data scientist to prepossess the data by filling missing values. It is important to be handled as they could lead to wrong prediction or classification for any given model being used.

There are three main problems that missing data causes: missing data can introduce a substantial amount of bias, make the handling and analysis of the data more arduous, and create reductions in efficiency.

2. Methodology

a) The first step of multiple imputation for missing data is to impute the missing values by using an appropriate model which incorporates random variation.

b) The second step of multiple imputation for missing data is to repeat the first step 3-5 times.

c) The third step of multiple imputation for missing data is to perform the desired analysis on each data set by using standard, complete data methods.

d) The fourth step of multiple imputation for missing data is to average the values of the parameter estimates across the missing value samples in order to obtain a single point estimate.

e) The fifth step of multiple imputation for missing data is to calculate the standard errors by averaging the squared standard errors of the missing value estimates. After this, the researcher must calculate the variance of the missing value parameter across the samples. Finally, the researcher must combine the two quantities in multiple imputation for missing data to calculate the standard errors.

3. Algorithm

In order to deal with the problem of increased noise due to imputation, Rubin (1987) developed a method for averaging the outcomes across multiple imputed data sets to account for this. All multiple imputation methods follow three steps.

a) Imputation – Similar to single imputation, missing values are imputed. However, the imputed values are drawn m times from a distribution rather than just once. At the end of this step, there should be m completed datasets.

b) Analysis – Each of the m datasets is analyzed. At the end of this step there should be m analyses.

c) Pooling – The m results are consolidated into one result by calculating the mean, variance, and confidence interval of the variable of concern.

4. Results and Analysis

The acccuracy of the model improved by a considerable amount and the cost of the model comes out to be lesses than the time it contained missing values.

5. Conclusion

The handling of missing values makes the model more accurate.