Erasmus University Rotterdam

Missing Data Imputation Techniques

Tatev Karen Aslanyan

KNN Imputation

Missing data is a widely-known problem since most of the statistical methods assume that the data used in the analysis is complete while often this assumption does not hold. Even a small amount of missing values in the data can cause serious problems leading to wrong estimation results and conclusions. There exist numerous methods for solving this problem. These techniques for imputation, estimation of missing values, are divided into univariate and multivariate methods. Univariate imputation is usually used because of its simplicity. More specifically, for a continuous variable many researchers replace the missing data with the mean or median of the observed values, and for a categorical variable, they replace missing values with the mode of the observed values. However, univariate imputation often destroys the multivariate structure of the data, which introduces bias. In addition, it leads to eliminating the variability in the imputed values. Therefore, multivariate approaches perform better and reflect sampling variability.

One single imputation approach for handling missing values is K-Nearest Neighbors(KNN) which leads to more accurate results since it estimates the missing values by their closest neighbors and not all observations. KNN has been introduced by Troyanskaya et. al. (2001) which aims to find for missing observation their k most similar neighbors(donors) in the observed data and use them to estimate the value of missing point. This method falls in the category of donor-based methods since it uses K most similar donors to estimate the missing value. This method can be used for the data which contains continuous, discrete, ordinal and categorical variables which makes it particularly useful for dealing with all kind of missing data. The idea behind KNN is that the missing value of a point can be approximated by the values of the points that are closest to it, assuming that if these observations are similar so should be their corresponding values. It is a simple method that can predict both quantitative and qualitative attributes and as a nonparametric method, it does not make an assumption of any particular model. Besides, it is not necessary to create a predictive model for each feature with missing data. However, KNN has also a few disadvantages. The method is quite sensitive to outliers, and it requires to pre-specify the distance and the number of neighbors (k). A small k produces deterioration in the performance of classifier after imputation in the estimation process. Moreover, KNN does not take into account the possible negative correlations between variables. Finally, the most accurate way of imputing data is multiple imputation (MI) firstly introduced by (Robin, 1975) which instead performing imputation once it imputes the same data for multiple times. This imputation method takes into account the uncertainty of imputed missing values by leading to more accurate estimates for the standard error which is not the case for single imputation. However, MI's greatest disadvantage is its complexity especially when it is applied to a large amount of data like this analysis. Therefore, we will use KNN imputation approach because it can handle data of large size while providing more accurate results than for instance mean or mode univariate imputation.

When implementing KNN imputation, we need to consider following parameters:

- The number of neighbors: Taking a low k will increase the influence of noise, and the results might be biased. On the other hand, taking a high k will tend to blur local impacts which are what we are looking for. It is also recommended to take an odd k for binary classes to avoid ties.

- The aggregation method: We allow for the arithmetic mean, median and mode for numeric variables (continuous) and mode for categorical ones (both ordinal and nominal).

- Normalizing the data: This will give every attribute the same influence in identifying neighbors when computing certain distances like the Euclidean one. The algorithm normalizes the data when both numeric and categorical variable are provided.

- Similarity distance: among the various distance metrics available, we will focus on the main ones, Euclidean, Hamming, and Manhattan.
  - Euclidean distance $d(x_k, x_l) = \sqrt{(\sum_{j=1}^{n}(x_{kj} - x_{lj})^2)}$ is good to use if the input variables are continuous.
  - Hamming distance $d(x_k, x_l) = \sum_{j=1}^{n} I(x_{kj} \neq x_{lj})$ can be used when there are nominal variables.
  - Manhattan distance $d(x_k, x_l) = \sum_{j=1}^{n} |x_{kj} - x_{lj}|$ is a good measure if the input variables are ordinal.

Algorithm 1 KNN basic algorithm

```
Input: Data that contain missing values
Output: Imputed data ($x_{ij}$)

• For every $x_j$ where $j = 1, ..., n$
    – Compute pairwise distances between observations, leaving out variable $x_j$ and using only
      pairwise complete information
    – For each observation $x_i$ , $i = 1, ..., n$, with $x_{ij}$ missing:
        ∗ Find k donors with observed value in variable $x_j$ that have the smallest distances
          from unit $x_i$
        ∗ Set $x_{ij}$ to aggregated value of variable $x_j$ from donors
• Return imputed data matrix $(x_{ij})_{1 \leq i \leq n, 1 \leq j \leq p}$
```