Data Imputation Techniques

Simran Dewangan(22111056)

July 12, 2024

Introduction

Data imputation is used to fill missing data in a dataset using various techniques. This solves the issues of an incomplete dataset. I have explored 5 different data imputation techniques with different case studies and datasets for specific tasks. I have first explained what are the different techniques I used, how do they work and the case studies I used in the process. I have uploaded the datasets with missing values shown as 'NaN', the imputed datasets and the python codes used for the same.

1. Mean/Median/Mode Imputation

This is one of the most easiest data imputation technique where the missing value is replaced by the Mean(average of the available data for a particular variable), Median or Mode of the given data. I have used the Mean Imputation Technique which is discussed in the next section.

Case Study: Housing Area

In a dataset of housing area, we have the longitude, latitude, median age of the housing area, total rooms and bedrooms in that area. In this particular dataset, some values are missing shown by 'NaN'. This is imputed using calculating the averages in each column and replacing it with the missing values.

2. K-Nearest Neighbors Imputation

In this method, we assume that the missing data has a value closer to the nearest data which is why we call it, 'K-nearest neighbors'. The missing value is replaced by a value close to the nearest neighbour. This can be used for both numeric and categorized datasets.

Case Study: Heart Health Analysis

In a dataset of Heart Health Analysis, the data regarding age, sex, resting BP, cholesterol and maximum heart rate is shown. Some data is missing which is imputed by using the KNN imputer, assuming the missing values are close to the nearest values.

3. Interpolation Imputation

It is a technique used to calculate unknown values assuming that it falls right between the known values. There are different types of interpolation techniques. The one I have used is Linear Interpolation, where the missing values are estimated by drawing a straight line between the two closest known points.

Case Study: Diabetes

In a Diabetes dataset of women, their no. of pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI and Age is mentioned. This is a numeric dataset which can be solved using interpolation techniques. I used the linear interpolation methos here.

4. Random Forest Regression

This method uses Machine Learning to Handle the missing data. It uses the random forest algorithm to predict the missing data by making decision trees and comparing the data available to it.

Case Study: Iris Species and Characteristics

A dataset of iris is used where some of its features are given such as its Sepal Length, Sepal Width, Petal Length, Petal Width and Species. A random forest regression algorithm is used to impute the missing data values. This can be used in imputing categorized data.

5. MICE Technique

Multiple Imputation by Chained Equations is one of the most powerful methods that imputes missing values by combining multiple imputations and creates multiple imputed datasets using regression models.

Case Study: Red Wine Quality

There is a record of the quality and constituents of a Red Wine, where its properties are shown. I have applied MICE technique to impute the missing data. First, the data is divided into numeric and categorized dataframes and then interative imputer is used.