DATA IMPUTATION TECHNIQUES

BY AMARRAM MADHU MALVEETIL

ROLL NO.:21111009

"Department of BIOMEDICAL ENGINEERING"

"NATIONAL INSTITUTE OF TECHNOLOGY, RAIPUR", CHATTISGARH"

BATCH:2026 SEMESTER:V

Assignment 1 of "ARTIFICIAL INTELLIGENCE"

SUBMITTED ON JULY 12, 2024

INTRODUCTION

In data analysis and machine learning, data imputation is a crucial preprocessing procedure that addresses the ubiquitous problem of missing values in datasets. The quality and dependability of analytical results can be severely impacted by missing data, which can result in skewed estimations and decreased statistical power. Through the use of imputation techniques, analysts can ensure that following analyses are thorough and resilient by substituting reasonable estimates for missing values, thus maintaining the integrity of the dataset.

This paper explores a variety of techniques for data imputation, from straightforward plans like mean imputation to more complex plans like Multiple Imputation by Chained Equations (MICE). To give readers a thorough grasp of how to manage missing data successfully, each technique's theoretical foundations, practical implementations, and efficacy in various contexts will be addressed. These will be bolstered by real-world case studies and informative Python code examples.

1. MEAN/MEDIAN/MODE IMPUTATION

Technique: Replace missing values with the mean, median, or mode of the respective feature.

Case Study:

Dataset: Healthcare dataset with patient ages.

Scenario: Missing ages in the dataset.

Approach: Impute missing ages with the median age of the patients since age data can be skewed.

Result: This method maintains the central tendency of the data, though it can underestimate the variability.

2. K-NEAREST NEIGHBOURS (KNN) IMPUTATION

Technique: Use the k-nearest neighbors to impute missing values. The missing value is replaced by the average (or majority) of the k-nearest neighbors.

Case Study:

Dataset: Housing prices dataset with missing values in the "number of bedrooms" column.

Scenario: Some entries are missing the number of bedrooms.

Approach: For each missing value, find the k-nearest houses (based on other features like size, location, price) and impute the missing number of bedrooms with the average number from these neighbors.

Result: This method takes into account the relationships between features, providing more accurate imputations in complex datasets.

3. MULTIPLE IMPUTATIONS BY CHAINED EQUATIONS

Technique: Create multiple imputations for missing data, which are then combined to produce a single set of estimates.

Case Study:

Dataset: Clinical trial data with missing values in blood pressure measurements.

Scenario: Several patients have missing blood pressure readings.

Approach: Use MICE to generate multiple imputed datasets by predicting missing values based on other observed data. The results from these datasets are then pooled to form a final, robust estimate.

Result: MICE accounts for the uncertainty around missing values, providing more reliable and valid statistical inferences.

4 PREDICTIVE MODEL IMPUTATION

Technique: Use regression or classification models to predict and impute missing values.

Case Study:

Dataset: Credit scoring dataset with missing values in the "income" column.

Scenario: Some records have missing income information.

Approach: Train a regression model using other features (like age, occupation, credit score) to predict the missing income values.

Result: Predictive modeling can accurately estimate missing values if the relationships between features are well understood and captured by the model.

5. INTERPOLATION

Technique: Estimate missing values within the range of observed data points using linear or polynomial interpolation.

Case Study:

Dataset: Time series data of monthly sales.

Scenario: Missing sales data for a few months.

Approach: Use linear interpolation to estimate missing sales data by connecting the dots between known sales values before and after the missing periods.

Result: Interpolation is particularly useful for time-series data, providing smooth transitions between known data points.

CONCLUSION

These techniques illustrate various ways to handle missing data depending on the dataset characteristics and the nature of the missing values. Each method has its strengths and weaknesses, and the choice of technique should consider the specific context and the potential impact on subsequent data analysis.