Imputation Techniques

Data imputation refers to the process of filling in missing values in a dataset. In machine learning, imputing missing values is important because many algorithms are sensitive to missing values and cannot handle them effectively. If missing values are not handled, they can lead to biased or incorrect results, especially when using methods like regression, decision trees, and clustering.

There are several techniques used for data imputation in machine learning, including:

Mean/Median/Mode imputation: In this method, missing values are replaced by the mean, median, or mode of the non-missing values in the same column. Mean imputation is used for continuous variables, median imputation for ordinal variables, and mode imputation for categorical variables. This method is simple and easy to implement, but it can result in loss of information and introduce bias if the distribution of the data is not symmetrical.

Regression imputation: In this method, a regression model is used to predict the missing values based on the values of the other variables in the dataset. The regression model can be linear, polynomial, or any other type of regression model. The advantage of this method is that it can capture the relationship between the variables, but it requires a well-defined regression model and can result in overfitting if the model is too complex.

Multivariate Imputation by Chained Equation (MICE)

MICE algorithm is probably one of the most used imputation techniques and also a popular interview question.

Multivariate Imputation by Chained Equation (MICE) is a statistical technique used to impute missing values in a dataset with multiple variables. MICE is a flexible and powerful imputation method that can handle missing data in both continuous and categorical variables and variables with complex relationships.

MICE creates multiple imputed datasets, where missing values are imputed based on the observed values of other variables in the dataset. The imputation process is performed sequentially, where each variable is imputed conditional on the observed values of the other variables in the dataset. This process is repeated multiple times, with each iteration generating a new set of imputed values for the missing data.

The mathematical formulation of MICE can be described as follows:

Suppose we have a dataset Y with n observations and p variables, where some of the values are missing. Let Y_hat denote the imputed dataset, where imputed values replace missing values.

The MICE algorithm can be broken down into the following steps:

- Initialization: Initialize the imputed dataset Y_hat by using a simple imputation method, such as mean imputation or regression imputation.
- To apply MICE algorithm, we will use IterativeImputer from scikit-learn. This estimator is still under experimental, so we must import enable_iterative_imputer.
- Iteration: For each variable i, impute the missing values using the conditional distribution of i given the observed values of the other variables in Y_hat. Let Y_hat_i denote the imputed values for variable i, where imputed values replace missing values. Repeat this step for all variables in the dataset.
- Convergence: Repeat the iteration step multiple times until the imputed values converge to a stable solution. This can be assessed by examining the convergence of the imputed values across iterations.
- Combining: After multiple imputed datasets have been created, combine the results using a formula that considers the uncertainty in the imputed values. For example, the final imputed value for a variable could be the average of the imputed values across multiple datasets. The MICE algorithm can be expressed mathematically using the following formula:

$$Y_i^* = f_i(Y_1, Y_2, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_p)$$

where Y_i^* is the imputed value for variable i, f_i is the imputation model for variable i, and Y_1, Y_2, ..., Y_{i-1}, Y_{i+1}, ..., Y_p are the observed values of the other variables in the dataset.

MICE is a powerful imputation method that can handle missing data in complex datasets. It has been shown to be more accurate than other imputation methods, such as listwise deletion and mean imputation.