Data Mining Lab 1

Divyanshu Khandelwal

Course title

Introduction To Data Mining

#Lab-1: Data Pre-processing Techniques (Handling Missing Data, Categorical Data Encoding, Normalization)
Session Date: 22nd August 2023

/ / Prerequisite:

- Managing Incomplete Data.

PRE-LAB:

1. Strategies for Handling Missing Data: Explained in Detail

ANSWER:-
Addressing missing data is a pivotal aspect of both data analysis and machine learning, primarily due to the prevalence of incomplete or absent information in real-world datasets. Several approaches can be adopted to tackle this issue, each with its advantages and limitations. Let's explore various scenarios and the corresponding techniques:

1. Eliminating Rows with Missing Data (Complete Case Analysis):
In cases where the amount of missing data is minimal and doesn't heavily impact the analysis, a viable strategy is to remove rows containing missing values. Although this method is straightforward, it can lead to the loss of valuable insights and potential bias if the missing data follows a non-random pattern.

2. Using Default Values or Placeholders:
For categorical data gaps, default categories or placeholders can be assigned to indicate missing data. Numeric data can be replaced with common values like 0 or special values such as -999. While this approach is simple, it might unintentionally introduce patterns into the data.

3. Imputing with Mean, Median, or Mode:
This technique involves substituting missing numeric values with the mean, median, or mode of observed values. While it maintains the central tendencies of the data, it could result in an underestimation of variability and distortion of the distribution.

4. Regression-based Imputation:
When the missing values in certain columns can be predicted using other variables, employing regression analysis to estimate those values can be effective. This approach is particularly valuable when strong correlations exist between the variables and a stable linear relationship can be assumed.

5. Interpolation and Extrapolation:
Interpolation estimates values between observed points, while extrapolation estimates values beyond the observed range, primarily suitable for time-series data following specific trends. These methods work well with trend-based data, but intricate underlying patterns might compromise their accuracy.

6. Multiple Imputation:
Multiple Imputation involves generating multiple imputed datasets, each with different estimations for missing values. Analysis is performed on each dataset, and results are aggregated to account for the uncertainty introduced by missing data. This approach yields more accurate estimates and valid statistical insights.

7. Advanced Techniques (Machine Learning):
Leveraging machine learning algorithms like k-nearest neighbors, decision trees, and random forests can predict missing values based on inherent data relationships. However, these methods are intricate and require careful tuning and validation to avoid overfitting.

8. Utilizing Domain-specific Knowledge:
Domain expertise can aid in inferring missing values. For instance, in medical data analysis, a physician's input could be used to estimate absent medical measurements.

Choosing the appropriate strategy depends on data characteristics, the extent of missing data, analysis objectives, and the potential impact of different approaches on outcomes. It's crucial to thoroughly evaluate the consequences of each method and document the handling of missing data for transparency and reproducibility in analyses.

2. Explanation of Label Encoding and One Hot Encoding with an example.

ANSWER:-

One-Hot Encoding:
Encoding involves converting data. One-hot encoding transforms categorical data into numeric format by creating new columns. Each column represents a category and is filled with 1s and 0s, indicating the presence or absence of that category. To illustrate, let's consider an example involving four countries: India, Australia, Russia, and America.

Label Encoding:
Label encoding is a simple technique that involves converting each categorical value into a numerical value. For instance, consider a dataset with multiple columns. To illustrate label encoding, let's focus on a single categorical column: "State." This column includes different states as its values.

3. Explanation of Various Normalization Techniques with their Formulas.

ANSWER:-

Data normalization is a crucial step in data mining to bring data values onto a common scale. This is essential as many machine learning algorithms are sensitive to the scale of input features, and normalizing the data can lead to improved results. Different normalization techniques include:

1. Min-Max Normalization:
Formula: (x - min(x)) / (max(x) - min(x))
This technique scales feature values between 0 and 1 by subtracting the minimum value and dividing by the range.

2. Z-Score Normalization (Standardization):
Formula: (x - mean(x)) / standard deviation(x)
This method standardizes feature values to have a mean of 0 and a standard deviation of 1 by subtracting the mean and dividing by the standard deviation.

3. Decimal Scaling:
Formula: x / 10^d
Feature values are divided by a power of 10 (d) to ensure values are within a certain range.

4. Logarithmic Transformation:
Formula: log(x)
A logarithmic transformation is applied to feature values, which can help manage data with a wide range of values, reducing the impact of outliers.

5. Root Transformation:
Formula: sqrt(x)
Feature values are transformed using the square root function, useful for managing data with a wide range of values and mitigating the effect of outliers.

6. Normalization is vital for ensuring that features are on the same scale, preventing dominance of certain features and enhancing model accuracy. Each technique suits different data types and model requirements. Remember that normalization should be applied to input features, not the target variable, to avoid introducing bias.

4. Why Data Cleaning, Scrubbing, and Normalization Constitute Over 70% of a Data Scientist's Workload?

ANSWER:-

The assertion that "Data cleaning, scrubbing, and normalization make up over 70% of a data scientist's workload" underscores the substantial time and effort data scientists dedicate to preparing and preprocessing data before meaningful analysis or model development can occur. This is due to several reasons:

1. Data Quality and Integrity:
Real-world data is often messy, incomplete, and prone to errors. Prior to analysis or modeling, data scientists must identify and rectify issues such as missing values, inconsistent formats, duplicates, and outliers. Ensuring data quality and integrity is paramount for obtaining reliable and accurate results.

2. Diverse Data Sources:
Data can originate from diverse sources, including databases, spreadsheets, APIs, and unstructured text. Each source may have its own structure and format, necessitating standardization and transformation into a consistent format for analysis.

3. Handling Missing Data:
Numerous datasets contain missing values, which can hinder analysis and modeling. Data scientists need to determine how to handle missing data through techniques like imputation, demanding careful consideration to avoid bias and erroneous conclusions.

4. Feature Engineering:
Effective feature engineering is critical for building accurate and robust models. Data scientists often create new features or derive insights from existing ones. This involves transformations, aggregations, and combining variables to enhance model predictive power.

5. Managing Outliers:
Outliers, data points deviating significantly from the rest, can skew statistics and affect model performance. Data scientists must decide whether to remove, transform, or address outliers, ensuring the data is representative.

6. Normalization and Scaling:
Feature normalization or scaling ensures that different features have similar scales, preventing certain features from dominating others. Algorithms sensitive to input scales benefit from this step, leading to more reliable models.

7. Reproducibility and Documentation:

Thorough data preprocessing is crucial for reproducibility and transparency. Documenting steps taken during data cleaning and preprocessing ensures work can be replicated and understood by others.

Due to these factors, data cleaning, scrubbing, and normalization are time-intensive but necessary steps. Neglecting them can lead to biased results, unreliable models, and erroneous conclusions. Therefore, a substantial portion of a data scientist's efforts is dedicated to ensuring data quality and readiness for analysis.

IN-LAB:

You are very much interested in Data Science, and you thought of doing a project on Data Mining. So, you want to predict the IMDB rating of a movie based on various features. But before that you realized that the

Data is not clean. So, you need to perform following operations:

a. Impute the columns containing more than 100 NaN values with suitable central tendency measure.

b. Remove the rows with NaN values in remaining columns.
c. Label Encode the columns ‘language’, ‘country’, ‘content rating’. d. OneHotEncode the column ‘country’

Data Set:

https://www.kaggle.com/carolzhangdc/imdb-5000-movie-dataset

SOLUTION LINK: Click Here

LINK TO THE CODE:-

https://colab.research.google.com/drive/1OQNYQzx8kdkOWQ0wg7EYav8wze-RT-jq?usp=sharing

POST-LAB:

1. You are planning to travel in the airplane, and you have data which contains origin place, destination, time, distance. As a Data Scientist, you are curious whether there is any correlation between hours of journey and distance. But you realized that data is not clean. So, you need to perform following steps:

1. Find the missing values of hours and distance and replace with the mean and find the correlation and visualize it. And find the correlation by normalizing the data using decimal scaling technique. And find the difference in both correlations.
2. Find the missing values of hours and distance and replace with the median and find the correlation and visualize it. And find the correlation by normalizing the data usingmin-max normalization technique. And find the difference in both correlations.
3. Find the missing values of hours and distance and replace with the mode and find the correlation and visualize it. And find the correlation by normalizing the data using z-score normalization technique. And find the difference in both correlations.
4. At which mean/median/mode do you find the maximum correlation.

Data Set:

https://drive.google.com/file/d/1z93XaFkM1mErUsazRh1GWwaAFllL8DhO/view?usp=sharing

solution:- Click Here

LINK TO THE CODE:-

https://colab.research.google.com/drive/1UguAiHzHBvxnfuhdknDHu
w8R3lALkp8O?usp=sharing