Data Preprocessing-Data Encoding

Data Encoding

Data encoding is a crucial step in data preprocessing that involves converting categorical variables into numerical variables. The goal of encoding is to transform non-numerical data into a format that can be processed by machine learning algorithms, which typically require numerical inputs.

Why Encode Categorical Variables?

Categorical variables are often encoded using one-hot encoding (OHE) or label encoding (LE), as they provide valuable information about the relationships between different categories. By encoding categorical variables, we enable machine learning models to:

1. Understand the relationships between different categories.
2. Make predictions based on these relationships.

Types of Data Encoding

There are several types of data encoding:

1. One-Hot Encoding (OHE)

In OHE, each category is represented as a binary vector where only one element is set to 1 and the rest are 0s. This method is commonly used for categorical variables with multiple categories.

Example

Suppose we have a categorical variable `color` with values `red`, `green`, and `blue`. Using OHE, we get:

color	red	green	blue
red	1	0	0
green	0	1	0
blue	0	0	1

2. Label Encoding (LE)

In LE, each category is assigned a unique integer value. This method is commonly used for categorical variables with two categories.

Example

Suppose we have a categorical variable `gender` with values `male` and `female`. Using LE, we get:

gender	0	1
male	0	
female		1

3. Ordinal Encoding

In ordinal encoding, each category is assigned a unique integer value based on its position in the order.

Example

Suppose we have a categorical variable `education` with values `high school`, `college`, and `masters`. Using ordinal encoding, we get:

education	0	1	2
high school	0		
college	1		
masters		2	

4. Binary Encoding

In binary encoding, each category is represented as a binary vector.

Example

Suppose we have a categorical variable `color` with values `red`, `green`, and `blue`. Using binary encoding, we get:

color	binary
red	101
green	010
blue	011

Choosing the Right Encoding

The choice of encoding depends on the type of categorical variable and the specific problem being addressed. One-hot encoding is generally used for categorical variables with multiple categories, while label encoding is commonly used for binary categorical variables.

Remember to always normalize or scale the encoded data after applying any encoding technique to ensure that it's in a suitable format for machine learning algorithms.

Curated by Brajesh Kumar