Categorical Encoding Methods

1 Categorical Encoding

Categorical encoding is the process of converting categorical data (such as labels or strings) into numerical form so that machine learning models can process them. Since most machine learning algorithms require numerical inputs, encoding categorical features is essential. Below are some common encoding techniques.

2 One-Hot Encoding

In one-hot encoding, each unique category is converted into a new binary column. For example, if you have three categories (Red, Green, Blue), they will become three new columns with binary values indicating the presence (1) or absence (0) of the category.

Color	Red	Green	Blue
Red	1	0	0
Green	0	1	0
Blue	0	0	1

Table 1: One-Hot Encoding Example

3 Label Encoding

In label encoding, each unique category is assigned an integer. This method is simple but can introduce unintended ordinal relationships in the data.

Color	Encoded Label
Red	0
Green	1
Blue	2

Table 2: Label Encoding Example

4 Target Encoding (Mean Encoding)

Target encoding replaces each category with the mean of the target variable for that category. Below is an example where the target variable is sales.

Example:

Consider the following data:

Color	Sales
Red	10
Green	7
Red	12
Blue	9
Green	8
Blue	6
Red	9

Table 3: Original Data

Now, calculate the mean sales for each color:

$$\text{Mean Sales (Red)} = \frac{10 + 12 + 9}{3} = 10.33$$

$$\text{Mean Sales (Green)} = \frac{7 + 8}{2} = 7.5$$

$$\text{Mean Sales (Blue)} = \frac{9 + 6}{2} = 7.5$$

Replace each color with the mean sales:

Color	Sales	Encoded Color (Mean Sales)
Red	10	10.33
Green	7	7.5
Red	12	10.33
Blue	9	7.5
Green	8	7.5
Blue	6	7.5
Red	9	10.33

Table 4: Target Encoding Example

5 Binary Encoding

Binary encoding converts categories into binary numbers and then splits the binary digits into different columns.

Example:

Let's consider the categories A, B, and C:

Category	Binary Encoding
A	01
B	10
C	11

Table 5: Binary Encoding Example

6 Frequency Encoding

In frequency encoding, each category is replaced with its frequency of occurrence in the dataset.

Example:

Consider the following data:

Color	Frequency
Red	100
Green	150
Blue	200

Table 6: Frequency Encoding Example

7 Ordinal Encoding

Ordinal encoding assigns an integer to each category based on the natural order of the categories.

Example:

Consider the categories Small, Medium, and Large:

Size	Encoded
Small	1
Medium	2
Large	3

Table 7: Ordinal Encoding Example