Encoding of Categorical Variables

EMPOWERING High Performance Technology Teams

OVERVIEW

Overview

- Total Session (2 hours)
- Focus majorly on Encoding of categorical variables

Agenda

- Mapping Method
- Ordinary Encoding
- Label Encoding
- Pandas Dummies

ENCODING

Categorical Data Encoding

- Most of the data in real life come with categorical string values while machines require all input and output features to be numeric
- Encoding categorical data is a process of converting categorical data into integer format so that the data with converted categorical values can be provided to the models to give and improve the predictions

Height
Tall
Medium
Short

Height
0
1
2

Mapping Method

Mapping is a technique used in data preprocessing to transform categorical variables into numerical representations

It is useful when there is an inherent order or hierarchy among the categories

It enables the conversion of qualitative data into a format that can be processed by machine learning algorithms

Involves creating a mapping dictionary that assigns unique numerical values to each category

Ordinal Encoding

Transforms categorical value into numerical value in ordered sets

The encoding starts from 0 or 1 and increments by one for the succeeding category

Customer feedback - 5 point Likert scale

Feedback	Assign numerical code
Poor	1
Fair	2
Good	3
Very Good	4
Excellent	5

Label Encoding

In label encoding, the categorical value is replaced with a numeric value between 0 and the number of classes minus 1. If the categorical variable value contains 5 distinct classes, we use (0, 1, 2, 3, and 4).

State (Nominal Scale)	State (Label Encoding)
Maharashtra	3
Tamil Nadu	4
Delhi	0
Karnataka	2
Gujarat	1
Uttar Pradesh	5

One-Hot Encoding / Pandas Dummies

Pandas Dummies creates binary columns for each category, representing the presence or absence of a category

Each category becomes a new column, and the value is 1 if the category is present and 0 if not

It is suitable when there is no inherent order among the categories and all categories are independent

Color
Red
Green
Blue

One-hot encoding →

d1	d2	d3
1	0	0
0	1	0
0	0	1

Curse of Dimensionality

- A potential drawback of this method is a significant increase in the dimensionality of the dataset (which is called Curse of Dimensionality)

- We are creating additional columns, one for each unique value in the set of the categorical attribute we'd like to encode. Suppose we have a categorical attribute that contains 1000 unique values, one-hot encoding will generate 1,000 additional new attributes and this is not desirable.

DEMO

Summary

- Categorical encoding is imperative since machines require all dependent and independent variables to be numeric
- Mapping method involves creating a mapping dictionary that assigns unique numerical values to each category
- Ordinal encoding transforms categorical value into numerical value in ordered sets
- Label encoding replaces categorical data with a numeric value between 0 and the number of classes minus 1
- One-Hot Encoding creates a new column (also called a dummy variable) with binary encoding (0 or 1) to denote whether a particular row belongs to this category or not