Documentations

Prosenjit Mondol

August 7, 2025

1 Data Preprocessing

- **Sampling**
  Sampling techniques are used to select a representative subset of data from a large population to reduce the computational complexity and improve the efficiency of the analysis.
- **Transformation**
  Transformation techniques involve manipulating raw data to create a single input, such as scaling, normalization, or encoding categorical data.
- **Denoising**
  Denoising techniques remove unwanted noise from the data that can lead to inaccurate results.
- **Imputation**
  Imputation techniques are used to fill in missing values in the data using statistical methods.
- **Feature extraction**
  Feature extraction techniques help to identify and extract relevant features from the data that are significant in a particular context.
- **Normalization**
  Normalization techniques are used to organize data for more efficient access and processing.

2 Handle Categorical Data

Categorical data is a type of data that represents qualitative or nominal characteristics, such as gender, occupation, Categorical data cannot be measured or compared using mathematical operations like addition or subtraction.

2.1 Different Encoding Methods for Categorical Data

- **One-Hot Encoding**
  One-Hot Encoding creates a new binary column for each category.

```
X = pd.get_dummies(X)
print(X)
```

- **Label Encoding**
  Label Encoding assigns a numerical value to each category.

```
from sklearn.preprocessing import LabelEncoder

lencoders = {}

for col in data[features].columns:
    lencoders[col] = LabelEncoder()
    data[col] = lencoders[col].fit_transform(data[col])

data[features].nunique()
```

- **Binary Encoding**
  Binary Encoding creates new columns representing each category.