How can I convert a categorical variable to numeric in Pandas?

Authored by

stats writer

July 2, 2024

Converting a categorical variable to a numeric variable in Pandas involves using the "pd.get_dummies" function to create dummy variables for each category, and then dropping one of the dummy variables to avoid the "dummy variable trap." This process allows for the conversion of categorical data into a numerical format that can be used for analysis and modeling in Pandas. By converting categorical variables to numeric, it becomes easier to perform mathematical operations and use statistical methods on the data.

Convert Categorical Variable to Numeric in Pandas

You can use the following basic syntax to convert a categorical variable to a numeric variable in a pandas DataFrame:

df = pd.factorize(df)

You can also use the following syntax to convert every categorical variable in a DataFrame to a numeric variable:

#identify all categorical variables
cat_columns = df.select_dtypes().columns#convert all categorical variables to numeric
df = df.apply(lambda x: pd.factorize(x))

The following examples show how to use this syntax in practice.

Example 1: Convert One Categorical Variable to Numeric

Suppose we have the following pandas DataFrame:

import pandas as pd

#create DataFrame
df = pd.DataFrame({'team': ,
'position': ,
'points': ,
'rebounds': })

#view DataFrame
df

team positionpoints rebounds
0 A G 5 11
1 A G 7 8
2 A F 7 10
3 B G 9 6
4 B F 12 6
5 B C 9 5
6 C G 9 9
7 C F 4 12
8 C C 13 10

We can use the following syntax to convert the 'team' column to numeric:

```
#convert 'team' column to numeric
df = pd.factorize(df)

#view updated DataFrame
df team positionpoints rebounds
0 0 G 5 11
1 0 G 7 8
2 0 F 7 10
3 1 G 9 6
4 1 F 12 6
5 1 C 9 5
6 2 G 9 9
7 2 F 4 12
8 2 C 13 10
```

Here is how the conversion worked:

Each team that had a value of 'A' was converted to 0.Each team that had a value of 'B' was converted to 1.Each team that had a value of 'C' was converted to 2.

Example 2: Convert Multiple Categorical Variables to Numeric

Once again suppose we have the following pandas DataFrame:

```
import pandas as pd

#create DataFrame
df = pd.DataFrame({'team': ,
'position': ,
'points': ,
'rebounds': })

#view DataFrame
df
```

team position points rebounds
0 A G 5 11
1 A G 7 8
2 A F 7 10
3 B G 9 6
4 B F 12 6
5 B C 9 5
6 C G 9 9
7 C F 4 12
8 C C 13 10

We can use the following syntax to convert every categorical variable in the DataFrame to a numeric variable:

```
#get all categorical columns
cat_columns = df.select_dtypes().columns#convert all categorical columns to numeric
df = df.apply(lambda x: pd.factorize(x))

#view updated DataFrame
df

team position points rebounds
0 0 0 5 11
1 0 0 7 8
2 0 1 7 10
3 1 0 9 6
4 1 1 12 6
5 1 2 9 5
6 2 0 9 9
7 2 1 4 12
8 2 2 13 10
```

Note: You can find the complete documentation for the pandas factorize() function .

Additional Resources

The following tutorials explain how to perform other common operations in pandas: