Python Feature Engineering Cookbook

Stephen CUI[1]

2022 年 12 月 19 日

[1] cuixuanStephen@gmail.com

目录

Chapter 1

Encoding Categorical Variables

Categorical variables are those whose values are selected from a group of categories or labels. In some categorical variables, the labels have an intrinsic order. These are called ordinal categorical variables. Variables in which the categories do not have an intrinsic order are called nominal categorical variables。

The values of categorical variables are often encoded as strings. To train mathematical or machine learning models, we need to transform those strings into numbers. The act of replacing strings with numbers is called categorical encoding.

补充： 数据的四个等级

1. nominal level（定类等级）
2. ordinal level（定序等级）
3. interval level（定距等级）
4. ratio level（定比等级）

The values of categorical variables are often encoded as strings. To train mathematical or machine learning models, we need to transform those strings into numbers. The act of replacing strings with numbers is called categorical encoding.

1.1 Technical requirements

We will also use the open-source Category Encoders Python library, which can be installed using pip:

```
pip install category_encoders
```

We will also use the Credit Approval dataset(**?? ??**).

```python
import random
import numpy as np
import pandas as pd

data = pd.read_csv('../data/credit_approval_uci.csv')

cat_cols = [
    c for c in data.columns if data[c].dtypes == 'O'
]
num_cols = [
    c for c in data.columns if data[c].dtypes != 'O'
]

data[num_cols] = data[num_cols].fillna(0)
data[cat_cols] = data[cat_cols].fillna('Missing')

if(not data.isnull().any().any()):
    print('not exist missing value')
```

1.2 Creating binary variables through one-hot encoding

In one-hot encoding, we represent a categorical variable as a group of binary variables, where each binary variable represents one category. The binary variable takes a value of 1 if the category is present in an observation, or 0 otherwise.

A categorical variable with k unique categories can be encoded using $k-1$ binary variables. For the Color variable, which has three categories ($k = 3$; red, blue, and green), we need to create two ($k-1 = 2$) binary variables to capture all the information so that the following occurs:

- If the observation is red, it will be captured by the red variable (red = 1, blue = 0).
- If the observation is blue, it will be captured by the blue variable (red = 0, blue = 1).
- If the observation is green, it will be captured by the combination of red and blue (red = 0, blue = 0).

Encoding into $k-1$ binary variables is well-suited for linear models. **There are a few occasions in which we may prefer to encode the categorical variables with k binary variables:**

- When training decision trees since they do not evaluate the entire feature space at the same time.
- When selecting features recursively.
- When determining the importance of each category within a variable.

1.2.1 How to do it...

In this recipe, we will compare the one-hot encoding implementations of pandas, scikit-learn, Feature-engine, and Category Encoders.