Preprocessing Categorical Features

20 Dec 2017

Often, machine learning methods (e.g. logistic regression, SVM with a linear kernel, etc) will require that categorical variables be converted into dummy variables (also called OneHot encoding). For example, a single feature `Fruit` would be converted into three features, `Apples` , `Oranges` , and `Bananas` , one for each category in the categorical feature.

There are common ways to preprocess categorical features: using pandas or scikit-learn.

Preliminaries

```
from sklearn import preprocessing
from sklearn.pipeline import Pipeline
import pandas as pd
```

Create Data

```
raw_data = {'first_name': ['Jason', 'Molly', 'Tina', 'Jake',
'Amy'],
        'last_name': ['Miller', 'Jacobson', 'Ali', 'Milner',
'Cooze'],
        'age': [42, 52, 36, 24, 73],
        'city': ['San Francisco', 'Baltimore', 'Miami', 'Douglas',
'Boston']}
df = pd.DataFrame(raw_data, columns = ['first_name', 'last_name',
'age', 'city'])
df
```

	first_name	last_name	age	city
0	Jason	Miller	42	San Francisco