实用python编程 第8讲

数据建模

2017-12-04

本节内容

- 数据分析（pivot table）
- 数据建模（“预测”titantic乘客是否幸存）

pivot table 数据透视表

- A ***pivot table*** is a table that summarizes data in another table
  - made by applying an operation such as sorting, averaging, or summing to data in the first table, typically including grouping of the data.

	A	B	C	D
0	foo	one	small	1
1	foo	one	large	2
2	foo	one	large	2
3	foo	two	small	3
4	foo	two	small	3
5	bar	one	large	4
6	bar	one	small	5
7	bar	two	small	6
8	bar	two	large	7

```
pd.pivot_table(df, values='D',
index=['A', 'B'], columns=['C'],
aggfunc=[np.sum, np.mean])
```

		sum		mean	
	C	large	small	large	small
A	B				
bar	one	4.0	5.0	4.0	5.0
	two	7.0	6.0	7.0	6.0
foo	one	4.0	1.0	2.0	1.0
	two	NaN	6.0	NaN	3.0

作业4回顾

根据titanic的乘客数据（train.xlsx），分别给出3种舱位（头等舱／二等舱／三等舱 Pclass=1, 2, 3)的男性、女性在本次海难中的存活率。按如下格式在屏幕中输出：

某同学的方案（用了布尔索引）

```
import pandas as pd
df=pd.read_excel('train.xlsx',sheetname='train')
df.head()

df_0=df[['Name','Pclass','Sex','Survived']]
for i in range (3):
    df_1=df_0[(df_0.Sex=='male')&(df_0.Pclass==i+1)]
    print i+1,'male survival_rate', df_1.Survived.mean()
    df_2=df_0[(df_0.Sex=='female')&(df_0.Pclass==i+1)]
    print i+1,'female survival_rate',df_2.Survived.mean()
```

```
1 male survival_rate 0.368852459016
1 female survival_rate 0.968085106383
2 male survival_rate 0.157407407407
2 female survival_rate 0.921052631579
3 male survival_rate 0.135446685879
3 female survival_rate 0.5
```

pivot_table方案

```
df = pd.read_excel('data/titanic/train.xlsx')
table = pd.pivot_table(df, values="Survived", index=['Pclass', 'Sex'], aggfunc=np.mean)
```

		Survived
Pclass	**Sex**	
1	**female**	0.968085
	male	0.368852
2	**female**	0.921053
	male	0.157407
3	**female**	0.500000
	male	0.135447

pivot_table方案

- 再看不同港口登船的存活率
  - 在pivot_table方法中指定参数 `columns=['Embarked']`

		Survived
Pclass	**Sex**	
1	**female**	0.968085
	male	0.368852
2	**female**	0.921053
	male	0.157407
3	**female**	0.500000
	male	0.135447

	Embarked	**C**	**Q**	**S**
Pclass	**Sex**			
1	**female**	0.976744	1.000000	0.958333
	male	0.404762	0.000000	0.354430
2	**female**	1.000000	1.000000	0.910448
	male	0.200000	0.000000	0.154639
3	**female**	0.652174	0.727273	0.375000
	male	0.232558	0.076923	0.128302

C: Cherbourg, France
Q: Queenstown, Ireland
S: Southampton, England

数据建模

- 两个阶段
  - 训练阶段：根据有限的带有答案的样本训练一个模型（分类器）
  - 测试阶段：用训练好的模型对测试样本进行预测

一个训练样本由两部分组成：观察值、目标值
一个测试样本只有观察值

训练样本 (观察值) → 训练 → loss(预测值, 目标值）

↓模型

测试样本 (观察值) → 预测 → 预测值

预测哪些乘客能存活

- 测试数据 test_with_label.xlsx
  - 共有418名乘客，其中男性266名、女性152名

```
testset = pd.read_excel('data/titanic/test_with_label.xlsx')
print testset.shape
print testset.Sex.value_counts()
```

```
(418, 12)
male      266
female    152
Name: Sex, dtype: int64
```

预测哪些乘客能存活

- 基于规则
  - 如果是女性，则survived=1， 否则survived=0

```
def naive_predict(x):
    return x.Sex == 'female'

res = testset.apply(naive_predict, axis=1)
matched = testset.Survived == res
accuracy = sum(matched) / float(testset.shape[0])
print 'accuracy %g' % accuracy
```

```
accuracy 0.76555
```

利用训练数据“学习”决策过程

- 对于每个训练样本，我们能观察到哪些值？

如何处理缺失值？

			Age					Cabin	
3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q

统计缺失

- 统计每列值为空的单元个数
  - 891条乘客纪录中有177条（约20%）没有Age值

```
df.apply(lambda x: sum(x.isnull()), axis=0)
```

```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

填充缺失值

- 用数据估计

```
pd.pivot_table(df, values='Age', index=['Sex'], aggfunc=[np.mean, np.median])
```

	mean	median
	Age	Age
Sex		
female	27.915709	27.0
male	30.726645	29.0

二分类问题：幸存或遇难

1. 特征提取
2. 分类器训练（使用sklearn机器学习库）
3. 预测
4. 评分

具体内容见notebook/data-modeling.ipynb