Python Tricks

Stephen CUI[1], Ten-Gallon-Head

2022 年 12 月 13 日

[1]junocuixuan@163.com

Chapter 1

Python包管理

`No modle named linearmodels`，如果linearmodels已经安装，很可能是Jupyter不是默认的python路径。

Chapter 2

数据清洗Data Wrangling

2.1 读入数据

2.1.1 CSV

CSV是以逗号分割的文本文件，CSV中的文字可以采用不同的编码，ANSI，GBK，UTF8，UNICODE等，一定会有一个，可能会有一个默认的，WINDOWS下默认的好像是ANSI，文件另存在确定按钮边上可以选个字符集存储，存储后的文本字符集就是你选的。

注意：用excel导出有UTF8和直接逗号选择，建议选择直接逗号，因为选UTF8，python默认读入会出现首列乱码。

2.1.2 memory error

扩充虚拟内存的操作解决问题，读的是1.04GB的csv，算损失函数的时候。

建议不解决memory error：因为报错源头在于csv过大了，所以改进读入的值。

dask 库是处理大批量数据的pandas； pandas只能在内存上运算； dask可以在硬盘加内存上运算；

2.2 缺失值

2.2.1 缺失值判断

判断一个数据框是否有缺失值可以连续使用两次any()函数：

```
1 df.isnull().any(axis='index').any(axis='index')
```

2.2.2 缺失值处理

Chapter 3

建模

3.1 逻辑回归Logistic Regression

1. logit回归建模的时候，做数据清洗时，要避免解释变量与被解释变量之间存在共线性，那样会导致后面的回归形成奇异矩阵。
2. 在数据清洗时，后面清洗的方法必须依托于前面方法，避免产生由于前后清洗方法不一致导致数据产生偏差。例如：我们第一开始对于数据进行线性内插时，用的forward方法，会留下数据前段的空值，所以后面再次填充时应用bfill而不是fbill。
3. logit回归的时候，need covariance of paremeters for computing（unnormalized）covariances

Part I

pandas

Chapter 4

User Guide

4.1 Selection by label

When using .loc with slices, if both the start and the stop labels are present in the index, then elements located between the two (including them) are returned:

```
s = pd.Series(list('abcde'), index=[0, 3, 2, 5, 4])
s.loc[3: 5]
```

If at least one of the two is absent, but the index is sorted, and can be compared against start and stop labels, then slicing will still work as expected, by selecting labels which rank between the two:

```
# Obviously 6 is out of the index range, but the data can still be filtered out
# There may be some problems later
# from an unexpected bug on December 13, 2022
s.sort_index().loc[1: 6]
```