用Pandas加工数据

北京理工大学计算机学院　高玉金

2019年3月

缺失值NaN

- 数据缺失在很多数据中存在，是首先要解决的常见问题
- NaN（Not a Number）在NumPy中是浮点值，易检测
- Python的None关键字在数组中也被作为NAN处理
- 补全空值：fillna()

```
s= pd.Series(list("Hello"))
s[0]=None
s[1]=np.NaN
```

```
In [3]: type(np.NaN)
Out[3]: float
```

```
0    None
1     NaN
2       l
3       l
4       o
dtype: object
```

判断是否存在空值

- 判断是否为空值：isnull()或notnull()方法
- Series和DataFrame均可使用该方法

```
s= pd.Series(list("Hello"))
s[0]=None
s[1]=np.NaN
print(s.isnull())
```

```
0    False
1    False
2     True
3     True
4     True
dtype: bool
```

```
0     True
1     True
2    False
3    False
4    False
dtype: bool
```

过滤空值dropna()

- Series过滤空值

```
0    None
1     NaN
2       1
3       1
4       o
dtype: object
```

dropna()

```
2    1
3    1
4    o
dtype: object
```

DataFrame过滤空值时，dropna()默认会删除包含缺失值的行

传入how=‘all’参数时，删除所有值均为NaN的行

传入axis=1，可以按照同样的方式删除列

df = pd.DataFrame([[1,None,3],[4,5,6]])

	0	1	2
0	1	NaN	3
1	4	5.0	6

df.dropna()

	0	1	2
1	4	5.0	6

df.dropna(axis=1)

	0	2
0	1	3
1	4	6

补全空值

- Dropna()简单粗暴，抛弃了其他非空数据
- 可以采取补全数据的方式，如fillna()
  - 常数补全，如fillna(0)
  - 字典补全，为不同的列设置不同的填充值，字典的键就是列的索引名，如fillna({1:0.5, 2:0})
- fillna()返回一个新的对象，或设定inplace=True修改原对象，如df.fillna(0, inplace=True)
- 空值补全可能会污染数据
  - 和reindex的插值方法一样，如前向填充，如fillna(method=“ffile”, limit=2)
  - 将Series的平均值或中位数填充NaN，如data.fillna(data.mean())

删除重复值

- 由于各种原因，DataFrame中会出现重复的行
- duplicated()用于判断某行是否重复，返回一个Series对象
- drop_duplicates()返回不重复的值
- 可以指定特定列，如：

drop_cuplicates([‘K1’])

df.duplicated()

```
0    False
1    False
2     True
dtype: bool
```

	K1	K2
0	a	4
1	b	5
2	b	5

df.drop_duplicates()

	K1	K2
0	a	4
1	b	5

df = pd.DataFrame({"K1":['a','b','b'],'K2':[4,5,5]})

检测和过滤异常值

- 异常值（outlier）的过滤或变换运算在很大程度上其实就是数组运算
- df = pd.DataFrame(np.random.randn(100,2))
- 找出一列中绝对值大于2的值

df[1][np.abs(df[1])>2]

- 使用any方法选出全部满足条件的行
- any(1)一行中有1个数据满足即可

df[(np.abs(df)>2).any(1)]

df.describe()

	0	1
count	100.000000	100.000000
mean	0.045470	0.042882
std	1.075840	0.975414
min	-2.394205	-2.901784
25%	-0.698770	-0.654541
50%	0.048204	0.132028
75%	0.713365	0.702426
max	3.606555	2.527802

数据排序

- Sort_index()在指定轴上对索引排序，默认升序，默认0轴（行）
- Sort_values()在指定轴上根据数值进行排序，默认升序
- NaN排序末尾（最大或最小）
- 取前n行最大值Nlargest()

```
df= pd.DataFrame(np.random.randn(10,2),columns=['A','B'])
```

```
df.sort_values('B',ascending=False)[:5]
```

```
df.nlargest(5,'B')
```

	A	B
9	1.342106	1.767738
7	-1.111334	1.369328
8	-0.274723	0.568214
4	0.558773	0.543989
6	0.408689	0.525098

数据替换

- 使用fillna填充空值是通用值替换的特殊例子
- 不同的数据源，对空值的标识不同，如-999标识空值
- 单值：Df.replace(-999,np.NaN, inplace = True)或
- 列表：Df.replace([-999, -1000], np.NaN)或
- 列表对：Df.replace([-999, -1000], [np.NaN,0])或
- 字典：Df.replace({-999: np.NaN, -1000:0})

其他数据处理功能

- Pandas的数据处理功能非常强大
  - 离散化和分箱
  - 计算指标和虚拟变量
  - 置换和随机抽样
  - 分层索引
  - 连接、重塑与透视
- 限于篇幅，本节内容仅是冰山一角
- 索引在Series和DataFrame对象中作用很大，请大家充分练习，加深理解