cloudera

Python 数据分析

Cloudera 大数据培训基地

重庆翰海睿智大数据科技有限公司

高级数据结构和操作类库 Pandas 基础

课程章节

高级数据结构和操作类库 Pandas

基础

高级数据结构和操作类库 Pandas 基础

在本章中，您将了解到：

1. Pandas 的数据结构
2. Pandas 的基本操作功能
3. Pandas 的约简与汇总统计
4. Pandas 缺失数据处理
5. Pandas 层次化索引功能

本章主题

高级数据结构和操作类库 Pandas 基础

- Pandas 的数据结构
- Pandas 的基本操作功能
- Pandas 的约简与汇总统计
- Pandas 缺失数据处理
- Pandas 层次化索引功能
- 结论

Pandas 简介

Pandas 类库含有使数据分析工作变得更快更简单的高级数据结构和操作工具。Pandas 是基于 NumPy 构建的，让以 NumPy 为中心的应用变得更加简单。

Pandas 特点

- 具备按轴自动或显式数据对齐功能的数据结构
- 集成时间序列功能
- 既能处理事件序列数据也能处理非事件序列数据的数据结构
- 数据运算和约简可以根据不同的元数据执行
- 灵活处理缺失数据
- 合并及其他出现在常见数据库中的关系型运算

Pandas 数据结构介绍

Pandas 引入方式

```
1 form pandas import Series,DataFrame
2 import pandas as pd
```

使用 Pandas 之前需要熟悉它的两个主要数据结构

- Series：类似一维数组的对象。
- DataFrame：一个表格型的数据结构。

Pandas 的 Series 对象 (1)

Series 对象是由一组数据以及一组与之相关的数据标签 (即索引) 组成。

Series 对象相关运用

- 一维数据的生成
- 获取值和索引
- 自定义数据点标记的索引
- 获取单个或一组值

Pandas 的 Series 对象 (2)

Series 对象相关运用

- 保留索引和值之间的链接
- 查看定长有序字典
- Python 字典创建 Series 对象
- 匹配索引找不到对应值时的结果
- isnull 和 notnull 函数检测缺失数据
- 自动对齐不同索引的数据
- 设置 name 属性
- 索引修改

Pandas 的 DataFrame 对象 (1)

DataFrame 对象是含有一组有序的列，每列可以是不同的值类型 (数值、字符串、布尔值等)。其实，DataFrame 对象中的数据是以一个或多个二维块存放的 (而不是列表、字典或别的一维数据结构)。

DataFrame 对象相关运用

- 创建 DataFreame 对象
- 指定排列顺序
- 传入的列无数据时的结果
- 获取列的数据
- 获取行的数据
- 修改列的数据
- 新增列的数据
- 删除列的数据

Pandas 的 DataFrame 对象 (2)

DataFrame 对象相关运用

- 嵌套字典创建 DataFreame 对象
- 设置行列 name 属性
- 获取所有数据

DataFrame 构造函数所能接受的数据

- 二维 ndarray
- 由数组、列表或元组组成的字典
- NumPy 的结构化/记录数组
- 由 Series 组成的字典
- 由字典组成的字典
- 字典或 Series 的列表
- 由列表或元组组成的列表

索引对象 (1)

pandas 的索引对象负责管理轴标签和其它元数据 (比如轴名称等)。

Pandas 中主要的 Index 对象

类	说明
Index	最泛化的Index对象，将轴标签表示为一个由Python对象组成的NumPy数组
Int64Index	针对整数的特殊Index
MultiIndex	“层次化”索引对象，表示单个轴上的多层索引。可以看做由元组组成的数组
DatetimeIndex	存储纳秒级时间戳（用NumPy的datetime64类型表示）
PeriodIndex	针对Period数据（时间间隔）的特殊Index

图 1: ‘Index 对象’

索引对象 (2)

Index 的方法和属性

方法	说明
append	连接另一个Index对象，产生一个新的Index
diff	计算差集，并得到一个Index
intersection	计算交集
union	计算并集
isin	计算一个指示各值是否都包含在参数集合中的布尔型数组
delete	删除索引i处的元素，并得到新的Index
drop	删除传入的值，并得到新的Index
insert	将元素插入到索引i处，并得到新的Index
is_monotonic	当各元素均大于等于前一个元素时，返回True
is_unique	当Index没有重复值时，返回True
unique	计算Index中唯一值的数组

图 2: Index 的方法和属性

本章主题

高级数据结构和操作类库 Pandas 基础

- Pandas 的数据结构
- Pandas 的基本操作功能
- Pandas 的约简与汇总统计
- Pandas 缺失数据处理
- Pandas 层次化索引功能
- 结论

重新索引 (1)

Pandas 对象重要方法 reindex，用于创建一个适应新索引的新对象。

reindex 的 (插值) method 选项

参数	说明
ffill或pad	前向填充（或搬运）值
bfill或backfill	后向填充（或搬运）值

图 3: ‘method 选项’

重新索引 (2)

reindex 函数的参数

参数	说明
index	用作索引的新序列。既可以是Index实例，也可以是其他序列型的Python数据结构。Index会被完全使用，就像没有任何复制一样
method	插值（填充）方式
fill_value	在重新索引的过程中，需要引入缺失值时使用的替代值
limit	前向或后向填充时的最大填充量
level	在MultiIndex的指定级别上匹配简单索引，否则选取其子集
copy	默认为True，无论如何都复制；如果为False，则新旧相等就不复制

图 4: ‘reindex 函数参数’

丢弃指定轴上的项

dorp 方法可以丢弃某个轴上的一个或多个项，只要有一个索引数组或列表即可。它返回一个指定轴上删除了指定值的新对象。

索引、选取和过滤 (1)

Series 索引的操作方式类似于 NumPy 数组的索引，只是索引值不只是整数。

DataFrame 进行索引就是获取一个或多个列。

DataFrame 的行上进行标签索引，可以通过 NumPy 的标记法以及轴标签选取行和列的子集。

索引、选取和过滤 (2)

DataFrame 索引的选项

类型	说明
obj[val]	选取DataFrame的单个列或一组列。在一些特殊情况下会比较便利：布尔型数组（过滤行）、切片（行切片）、布尔型DataFrame（根据条件设置值）
obj.ix[val]	选取DataFrame的单个行或一组行
obj.ix[:, val]	选取单个列或列子集
obj.ix[val1, val2]	同时选取行和列
reindex方法	将一个或多个轴匹配到新索引
xs方法	根据标签选取单行或单列，并返回一个Series
icol、irow方法	根据整数位置选取单列或单行，并返回一个Series
get_value、set_value方法	根据行标签和列标签选取单个值。[译注2]

图 5: ‘DataFrame 索引选项’

算术运算和数据对齐

Pandas 最重要的功能可以对不同索引的对象进行算术运算。

```
1 s1 = Series([10.3,-3.6,4.3,2.7],index=['a','b','d','e'])
2 s2 =
      Series([-3.9,7.4,-2.2,1.2,5.3],index=['a','d','e','f','g'])
3 s1 + s2
4
5 d1 =
      DataFrame(np.arange(1,10).reshape(3,3),columns=list('abc'))
6 d2 =
      DataFrame(np.arange(1,13).reshape(4,3),columns=list('bcd'))
7 d1 + d2
```

在算术方法中充值

在对不同索引的对象进行算术运算时，可能希望当一个对象某个轴标签在另一个对象中找不到是填充一个特殊值。

灵活的算术方法

方法	说明
add	用于加法（+）的方法
sub	用于减法（-）的方法
div	用于除法（/）的方法
mul	用于乘法（*）的方法

图 6: '算术方法'

排序和排名 (1)

根据条件对数据集排序也是一种重要的内置运算。要对行或列索引进行排序 (按字典排序)。

- 按索引进行排序 (默认升序)
- 按降序排序
- 按值排序
- 列值排序

排序和排名 (2)

排名跟排序关系密切，且它会增设一个排名值。

排名是用于破坏平级关系的 method 选项

method	说明
'average'	默认：在相等分组中，为各个值分配平均排名
'min'	使用整个分组的最小排名
'max'	使用整个分组的最大排名
'first'	按值在原始数据中的出现顺序分配排名

图 7: ‘method 选项’

带有重复值的轴索引

前面所讲解范例都有着唯一的索引值 (轴标签)，虽然许多 Pandas 函数都要求标签唯一，但这并不是强制性的。

- 重复索引 Series 对象
- 判断唯一值的 is_unique 属性
- 重复索引与唯一索引
- 重复索引 DataFrame 对象

本章主题

高级数据结构和操作类库 Pandas 基础

- Pandas 的数据结构
- Pandas 的基本操作功能
- Pandas 的约简与汇总统计
- Pandas 缺失数据处理
- Pandas 层次化索引功能
- 结论

汇总和计算描述统计 (1)

Pandas 对象拥有一组常用的数学和统计方法，它们大部分都属于简约和汇总统计，用于从 Series 中提取单个值或从 DataFrame 的行或列中提取一个 Series。

简约方法的选项

选项	说明
axis	约简的轴。DataFrame的行用0，列用1
skipna	排除缺失值，默认值为True
level	如果轴是层次化索引的（即MultiIndex），则根据level分组约简

图 8: '简约方法'

汇总和计算描述统计 (2)

描述和汇总统计

方法	说明
count	非NA值的数量
describe	针对Series或各DataFrame列计算汇总统计
min、max	计算最小值和最大值
argmin、argmax	计算能够获取到最小值和最大值的索引位置（整数）
idxmin、idxmax	计算能够获取到最小值和最大值的索引值
quantile	计算样本的分位数（0到1）
sum	值的总和
mean	值的平均数
median	值的算术中位数（50%分位数）
mad	根据平均值计算平均绝对离差
var	样本值的方差
std	样本值的标准差

图 9: '描述和汇总统计'

汇总和计算描述统计 (3)

描述和汇总统计 (续)

方法	说明
skew	样本值的偏度（三阶矩）
kurt	样本值的峰度（四阶矩）
cumsum	样本值的累计和
cummin、cummax	样本值的累计最大值和累计最小值
cumprod	样本值的累计积
diff	计算一阶差分（对时间序列很有用）
pct_change	计算百分数变化

图 10: '描述和汇总统计 (续)'

唯一值、值计数以及成员资格

还有一类方法可以从一维数组 Series 的值中抽取信息。

- unique 函数：用于得到 Series 中的唯一值数组。
- values_counts 函数：用于计算一个 Series 中各值出现的次数。
- isin 函数：用于判断矢量化集合的成员资格。

方法	说明
isin	计算一个表示“Series各值是否包含于传入的值序列中”的布尔型数组
unique	计算Series中的唯一值数组，按发现的顺序返回
value_counts	返回一个Series，其索引为唯一值，其值为频率，按计数值降序排列

图 11: ‘唯一值、值计数以及成员资格方法’

本章主题

高级数据结构和操作类库 Pandas 基础

- Pandas 的数据结构
- Pandas 的基本操作功能
- Pandas 的约简与汇总统计
- Pandas 缺失数据处理
- Pandas 层次化索引功能
- 结论

缺失数据介绍

缺失数据在大部分数据分析应用中都很常见。Pandas 的设计目标之一就是让缺失数据的处理任务尽量轻松。

Pandas 使用浮点值 NaN(Not a Number) 表示浮点和非浮点数组中的缺失数据。它只是一个便于被检查处理的标记而已。

滤除缺失数据

NA 处理的方法

方法	说明
dropna	根据各标签的值中是否存在缺失数据对轴标签进行过滤，可通过阈值调节对缺失值的容忍度
fillna	用指定值或插值方法（如ffill或bfill）填充缺失数据
isnull	返回一个含有布尔值的对象，这些布尔值表示哪些值是缺失值/NA，该对象的类型与源类型一样
notnull	isnull的否定式

图 12: ‘NA 处理方法’

填充缺失数据

如果不信过滤掉缺失数据，而希望通过其他方式填补哪些“空洞”。大多数情况下可以使用 fillna 方法将缺失值替换为某个常数值。

fillna 方法的参数

参数	说明
value	用于填充缺失值的标量值或字典对象
method	插值方式。如果函数调用时未指定其他参数的话，默认为“ffill”
axis	待填充的轴，默认axis=0
inplace	修改调用者对象而不产生副本
limit	（对于前向和后向填充）可以连续填充的最大数量

图 13: ‘fullna 的参数’

本章主题

高级数据结构和操作类库 Pandas 基础

- Pandas 的数据结构
- Pandas 的基本操作功能
- Pandas 的约简与汇总统计
- Pandas 缺失数据处理
- Pandas 层次化索引功能
- 结论

层次化索引介绍

层次化索引是 Pandas 的一项重要功能，它使你能在一个轴上拥有多个索引级别。

- Series 对象用列表或数组组成的列表作为索引。
- Series 对象层次化索引对象选取数据子集。
- Series 对象层次化索引在数据重塑和基于分组的操作。
- DataFrame 对象每条轴分层索引。
- DataFrame 对象各层指定名称。
- DataFrame 选取列分组。

重排分级顺序

有时候我们需要重新调整某条轴上各级别的顺序，或根据指定级别上的值对数据进行排序。

- swaplevel 方法

它接受两个级别编号或名称，并返回一个互换了级别的新对象 (但数据不会发生变化)。

- sort_index 方法

根据单个级别中的值对数据进行排序 (稳定的)，在交换级别时常常会用到。

根据级别汇总统计

许多对 DataFrame 和 Series 的描述和汇总统计都有一个 level 选项，它用于指定在某条轴上求和的级别。

本章主题

高级数据结构和操作类库 Pandas 基础

- Pandas 的数据结构
- Pandas 的基本操作功能
- Pandas 的约简与汇总统计
- Pandas 缺失数据处理
- Pandas 层次化索引功能
- 结论

基本要点

- Pandas 的数据结构
- Pandas 的 Series 和 DataFrame 对象
- 索引对象
- Pandas 的基本操作功能
- 重新索引、丢弃指定轴上的项
- 索引、选取和过滤
- 算术运算和数据对齐、算术方法中充值、排序和排名
- Pandas 的约简与汇总统计
- 汇总和计算描述统计、唯一值、值计数以及成员资格
- Pandas 缺失数据处理
- 滤除缺失数据、填充缺失数据
- Pandas 层次化索引功能
- 重新分级顺序、根据级别汇总统计