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

数组和矢量计算类库 NumPy

课程章节

数组和矢量计算类库 NumPy

数组和矢量计算类库 NumPy

在本章中，您将了解到：

1. NumPy 简介
2. NumPy 数组对象
3. 数组与标量之间的运算
4. 基本的索引与切片
5. 数组对象的相关操作
6. NumPy 通用函数与方法

本章主题

数组和矢量计算类库 NumPy

- NumPy 简介
- Numpy 数组对象
- 数组与标量之间的运算
- 基本的索引与切片
- 数组对象的相关操作
- NumPy 通用函数与方法
- 结论

NumPy 概述

Numpy(Numerical Python 的简称) 是高性能科学计算和数据分析的基础包。

部分功能

- ndarray，一个具有矢量算术运算和复杂广播能力的快速且节省空间的多维数组。
- 用于对整组数据进行快速运算的标准数学函数 (无需编写循环)。
- 用于读写磁盘数据的工具一级用于操作内存映射文件的工具。
- 线性代数、随机数生成以及傅里叶变换功能。
- 用于集成由 C，C++、Fortran 等语言编写的代码的工具。

本章主题

数组和矢量计算类库 NumPy

- NumPy 简介
- Numpy 数组对象
- 数组与标量之间的运算
- 基本的索引与切片
- 数组对象的相关操作
- NumPy 通用函数与方法
- 结论

NumPy 数组对象 ndarray

NumPy 最重要的一个特点就是其 N 维数组对象`ndarray`，该对象是一个快速而灵活的大数据集容器。可以利用它对整块数据执行一些数学运算。

`ndarray`是一个多维数组对象，该对象由两部分组成：

- 实际的数据
- 描述这些数据的元数据

示例 1：Python 与 NumPy 向量相加对比

示例 2：创建多维数组及属性

数组创建函数

函数	说明
array	将输入数据（列表、元组、数组或其他序列类型）转换为ndarray。要么推断出dtype，要么显式指定dtype。默认直接复制输入数据
asarray	将输入转换为ndarray，如果输入本身就是一个ndarray就不进行复制
arange	类似于内置的range，但返回的是一个ndarray而不是列表
ones、ones_like	根据指定的形状和dtype创建一个全1数组。ones_like以另一个数组为参数，并根据其形状和dtype创建一个全1数组
zeros、zeros_like	类似于ones和ones_like，只不过产生的是全0数组而已
empty、empty_like	创建新数组，只分配内存空间但不填充任何值
eye、identity	创建一个正方的N×N单位矩阵（对角线为1，其余为0）

图 1: ‘数组创建函数’

NumPy 数据类型 (1)

`dtype`(数据类型) 是一个特殊的对象，它包含有`ndarray`将一块内存解释为特定数据类型所需的信息。

```
1 arr1 = np.array([10,12,14,16],dtype=np.float64)
2 arr2 = np.array([11,13,15,17],dtype=np.int32)
```

NumPy 数据类型 (2)

类型	类型代码	说明
int8、uint8	i1、u1	有符号和无符号的8位（1个字节）整型
int16、uint16	i2、u2	有符号和无符号的16位（2个字节）整型
int32、uint32	i4、u4	有符号和无符号的32位（4个字节）整型
int64、uint64	i8、u8	有符号和无符号的64位（8个字节）整型
float16	f2	半精度浮点数
float32	f4或f	标准的单精度浮点数。与C的float兼容
float64	f8或d	标准的双精度浮点数。与C的double和Python的float对象兼容
float128	f16或g	扩展精度浮点数
complex64、complex128、complex256	c8、c16、c32	分别用两个32位、64位或128位浮点数表示的复数
bool	?	存储True和False值的布尔类型

图 2: ‘NumPy 数据类型’

NumPy 数据类型 (3)

类型	类型代码	说明
object	O	Python对象类型
string_	S	固定长度的字符串类型（每个字符1个字节）。例如，要创建一个长度为10的字符串，应使用S10
unicode_	U	固定长度的unicode类型（字节数由平台决定）。跟字符串的定义方式一样（如U10）

图 3: ‘NumPy 数据类型’

示例 3: NumPy 模拟商品数据类型

本章主题

数组和矢量计算类库 NumPy

- NumPy 简介
- Numpy 数组对象
- 数组与标量之间的运算
- 基本的索引与切片
- 数组对象的相关操作
- NumPy 通用函数与方法
- 结论

数组与标量之间的运算介绍

数组很重要，因为它使你不用编写循环即可对数据执行批量运算。这通常就叫做矢量化。

```
1 import numpy as np
2 arr = np.array([[1,2,3],[4,5,6]])
3 arr
4 arr * arr
5 arr - arr
6 1 / arr
7 arr ** 0.05
```

不同大小的数组之间的运算称为广播。

本章主题

数组和矢量计算类库 NumPy

- NumPy 简介
- Numpy 数组对象
- 数组与标量之间的运算
- 基本的索引与切片
- 数组对象的相关操作
- NumPy 通用函数与方法
- 结论

一维数组的索引

NumPy 数组的索引是一个内容丰富的功能，它选取数据子集或单个元素的方式很多。

```
1 import numpy as np
2 arr1 = np.arange(20)
3 arr1
4 arr1[7]
5 arr1[7:13]
6 arr1[7:13] = 21
7 arr1
```

切片自动传播与复制副本

将一个标量值赋值给一个切片时，该值会自动传播到整个选区。

自动传播

```
arr_section = arr1[7:13]
arr_section
arr_section[:] = 30
arr1
```

复制副本

```
arr_section1 = arr1[7:13].copy()
arr_section1[:] = 22
arr_section1
arr1
```

多维数组的索引

在高维度数组中，能做更多的事情。

```
1  arr2d = np.arange(1,24,2).reshape(3,4)
2  arr2d.ndim
3  arr2d
4  arr2d[1]
5  arr2d[1,2]
6  arr3d = np.arange(1,25).reshape(2,3,4)
7  arr3d.ndim
8  arr3d
9  arr3d[1]
10 arr3d[1,2]
11 arr3d[1,2,1]
```

改变数组维度

- ravel() 函数完成展平, 返回数组的一个视图。
- flatten() 函数完成平展，请求分配内存来保存结果。
- shape() 函数用元组设置维度。
- transpose() 函数是转置矩阵的操作。
- resize() 函数对原始多维数组进行修改。

更多的索引

- 切片索引

ndarray 的切片与 Python 列表一维对象一样。而高维度对象的花样更多，可以在一个或多个轴上进行切片，也可以跟整数索引混合使用。

- 布尔型索引

布尔值索引指的是一个由布尔值组成的数组可以作为一个数组的索引，返回的数据为 True 值对应位置的值。

- 花式索引

花式索引是指用整数数组进行索引。

示例 4：二维数组切片索引

示例 5：布尔型索引模拟超标 PM2.5

本章主题

数组和矢量计算类库 NumPy

- NumPy 简介
- Numpy 数组对象
- 数组与标量之间的运算
- 基本的索引与切片
- 数组对象的相关操作
- NumPy 通用函数与方法
- 结论

数组组合

NumPy 数组有水平组合、垂直组合和深度组合等多种组合方式。

- vstackd() 函数
- stack() 函数
- hstack() 函数
- column_stack() 函数
- row_stack() 函数
- concatenate() 函数

数组分割

NumPy 数组可以进行水平、垂直或深度分割，相关的函数分别有有：

- hsplit() 函数
- vsplit() 函数
- dsplit() 函数
- split() 函数

可以将数组分割成相同大小的子数组，也可以指定原数组中需要分割的位置。

数组属性

除了 shape 和 dtype 属性以外，ndarray 对象还有很多其他的属性。

- size：获取数组元素的总个数。
- itemsize：获取数组中的元素在内存中所占的字节数。
- nbytes：获取整个数组所占的存储空间。
- T：与 transpose() 函数一样。
- real：获取复数数组的实部。
- imag：获取复数数组的虚部。

本章主题

数组和矢量计算类库 NumPy

- NumPy 简介
- Numpy 数组对象
- 数组与标量之间的运算
- 基本的索引与切片
- 数组对象的相关操作
- NumPy 通用函数与方法
- 结论

元素级数组函数 (1)

函数	说明
abs、fabs	计算整数、浮点数或复数的绝对值。对于非复数值，可以使用更快的fabs
sqrt	计算各元素的平方根。相当于arr ** 0.5
square	计算各元素的平方。相当于arr ** 2
exp	计算各元素的指数e^x
log、log10、log2、log1p	分别为自然对数（底数为e）、底数为10的log、底数为2的log、log(1 + x)
sign	计算各元素的正负号：1（正数）、0（零）、－1（负数）
ceil	计算各元素的ceiling值，即大于等于该值的最小整数
floor	计算各元素的floor值，即小于等于该值的最大整数
rint	将各元素值四舍五入到最接近的整数，保留dtype
modf	将数组的小数和整数部分以两个独立数组的形式返回
isnan	返回一个表示“哪些值是NaN（这不是一个数字）”的布尔型数组
isfinite、isinf	分别返回一个表示“哪些元素是有穷的（非inf，非NaN）”或“哪些元素是无穷的”的布尔型数组
cos、cosh、sin、sinh、tan、tanh	普通型和双曲型三角函数

图 4: ‘一元 ufunc’

元素级数组函数 (2)

函数	说明
arccos、arccosh、arcsin、arcsinh、arctan、arctanh	反三角函数
logical_not	计算各元素not x的真值。相当于-arr

图 5: ‘一元 ufunc’

示例 7 绘制李萨茹曲线

元素级数组函数 (3)

函数	说明
add	将数组中对应的元素相加
subtract	从第一个数组中减去第二个数组中的元素
multiply	数组元素相乘
divide、floor_divide	除法或向下圆整除法（丢弃余数）
power	对第一个数组中的元素A，根据第二个数组中的相应元素B，计算A^B
maximum、fmax	元素级的最大值计算。fmax将忽略NaN
minimum、fmin	元素级的最小值计算。fmin将忽略NaN
mod	元素级的求模计算（除法的余数）
copysign	将第二个数组中的值的符号复制给第一个数组中的值
greater、greater_equal、less、less_equal、equal、not_equal	执行元素级的比较运算，最终产生布尔型数组。相当于中缀运算符>、>=、<、<=、==、!=
logical_and、logical_or、logical_xor	执行元素级的真值逻辑运算。相当于中缀运算符&、|、^

图 6: ‘二元 ufunc’

示例 8 返回两个数组元素级的最大值

运用数组进行数据处理

NumPy 数组可以将许多种数据处理任务表述为简洁的数组表达式(否则需要编写循环)。用数组表达式代替循环的做法，通常被称为矢量化。在 NumPy 中 meshgrid() 函数可以接受两个一维数组，并返回一个二维数组。

示例 9 绘制矢量化扩散图

数组条件逻辑运算

numpy.where() 函数是三元表达式 (x if condittion else y) 的矢量化版本。

```
1 arrone = np.array([1.1,1.2,1.3,1.4,1.5])
2 arrtwo = np.array([2.1,2.2,2.3,2.4,2.5])
3 arrbool = ([True,False,True,True,False])
```

示例 10 Where() 三元表达式选取数组的值

数学与统计方法

基本数组统计方法

方法	说明
sum	对数组中全部或某轴向的元素求和
mean	算术平均数。零长度的数组 mean 为 NaN
sta、var	分别为标准差和方差，自由度可调 (默认为 n)
min、max	最小值和最大值
agrmin、argmax	分别为最小和最大元素的索引
cumsum	所有元素的累计和
cumprod	所有元素的累积积

示例 11 绘制锯齿波和三角波

布尔型数组的方法

在统计方法中，布尔值会被强制转换为 1(True) 和 0(False)。sum() 方法经常被用来对布尔值数组中的 True 值计数。

示例 12 统计随机数正值数量

有两个方法对布尔值数组非常有用

- any()：用于测试数组中是否存在一个或多个 True。
- all()：用于测试数组中所有值是否都是 True。

排序

跟 Python 内置的列表一样，NumPy 也可以通过 sort() 方法进行排序。

```
arr6 = np.random.randn((8)
arr6
arr6.sort()
arr6
arr7 = np.random.randn(3,10)
arr7
arr7.sort(1)
arr7
```

唯一化与其他的集合逻辑

NumPy 提供了一些针对一维 ndarray 的基本集合运算，最常用的要数 unique。

方法	说明
unique(x)	计算x中的唯一元素，并返回有序结果
intersect1d(x, y)	计算x和y中的公共元素，并返回有序结果
union1d(x, y)	计算x和y的并集，并返回有序结果
in1d(x, y)	得到一个表示“x的元素是否包含于y”的布尔型数组
setdiff1d(x, y)	集合的差，即元素在x中且不在y中
setxor1d(x, y)	集合的对称差，即存在于一个数组中但不同时存在于两个数组中的元素[译注2]

图 7: ‘数组集合运算’

数组的文件输入输出

NumPy 能够读写磁盘上的文本数据或二进制数据。

二进制数据读写函数

- save()：数组是以未压缩的原始二进制格式保存在扩展名为.npy 文件中
- savez()：将多数组保存到一个压缩文件中，扩展名为.npz 文件中
- load()：读取指定磁盘的.npy 或 npz 文件。

文本数据读写函数

- savetxt()：将数据保存为文本数据
- loadtxt()：读取文本文件数据

示例 13 读取 CSV 文件股票相关数据

线性代数

函数	说明
diag	以一维数组的形式返回方阵的对角线（或非对角线）元素，或将一维数组转换为方阵（非对角线元素为0）
dot	矩阵乘法
trace	计算对角线元素的和
det	计算矩阵行列式
eig	计算方阵的本征值和本征向量
inv	计算方阵的逆
pinv	计算矩阵的Moore-Penrose伪逆
qr	计算QR分解
svd	计算奇异值分解（SVD）
solve	解线性方程组Ax = b，其中A为一个方阵
lstsq	计算Ax = b的最小二乘解

图 8: ‘numpy.linalg 函数’

随机数生成 (1)

函数	说明
seed	确定随机数生成器的种子
permutation	返回一个序列的随机排列或返回一个随机排列的范围
shuffle	对一个序列就地随机排列
rand	产生均匀分布的样本值
randint	从给定的上下限范围内随机选取整数
randn	产生正态分布（平均值为0，标准差为1）的样本值，类似于MATLAB接口
binomial	产生二项分布的样本值
normal	产生正态（高斯）分布的样本值
beta	产生Beta分布的样本值

图 9: ‘numpy.random 函数’

随机数生成 (2)

函数	说明
chisquare	产生卡方分布的样本值
gamma	产生Gamma分布的样本值
uniform	产生在[0, 1)中均匀分布的样本值

图 10: ‘numpy.random 函数’

本章主题

数组和矢量计算类库 NumPy

- NumPy 简介
- Numpy 数组对象
- 数组与标量之间的运算
- 基本的索引与切片
- 数组对象的相关操作
- NumPy 通用函数与方法
- 结论

基本要点

- NumPy 数组对象
- NumPy 数组对象 ndarray
- 数组创建函数、数据类型
- 基本的索引与切片
- 一维数组的索引、切片自动传播、复制副本
- 多维数组的索引、改变数组维度
- 切片索引、布尔值索引、花式索引
- 数组对象的相关操作
- 数组组合、分割、属性
- NumPy 通用函数与方法
- 数学与统计方法、布尔型数组、排序、数组文件输入输出
- 唯一化与其它集合逻辑、线性代数、随机数生成