专业 • 创新 • 增值

金程教育
GOLDEN FUTURE

Numpy数据分析基础

Numpy Basics in Data Analysis

纪慧诚

金程教育资深培训讲师

CFA FRM RFP

CONTENTS

PROFESSIONAL · LEADING · VALUE-CREATING

PART 1	Numpy简介
PART 2	认识和创建ndarray
PART 3	切片与索引
PART 4	常用通用函数
PART 5	基本数学统计方法
PART 6	条件逻辑的矢量化表达
PART 7	集合逻辑
PART 8	线性代数
PART 9	随机数

专业来自101%的投入!

Numpy简介

Numpy

- Python中科学计算的基础包
- 提供的基本功能：
  - ✓ a powerful **N-dimensional array object**;
  - ✓ sophisticated (broadcasting) functions;
  - ✓ tools for integrating C/C++ and Fortran code;
  - ✓ useful linear algebra, Fourier transform, and random number capabilities;
  - ✓ **vectorization** for fast operations without having to write loops;
  - ✓ tools for reading / writing array data to disk and working with memory-mapped files.

- 参考网站：
  - ✓ 英文官网：http://www.numpy.org/
  - ✓ 《用Python做科学计算》http://old.sebug.net/paper/books/scipydoc/index.html
  - ✓ 《利用Python进行数据分析》

CONTENTS

PROFESSIONAL · LEADING · VALUE-CREATING

PART 1 Numpy简介

PART 2 **认识和创建ndarray**

PART 2 切片与索引

PART 3 常用通用函数

PART 4 基本数学统计方法

PART 5 条件逻辑的矢量化表达

PART 6 集合逻辑

PART 7 线性代数

PART 8 随机数

专业来自101%的投入!

认识和创建ndarray

➢ What is ‘array’ ?

- Numpy中的主要对象：N维数组对象（ N dimentional array, ndarray ）
- 所有元素必须是相同类型的。

```
data = [[1,2,3,4],[5,6,7,8]]
arr = np.array(data)
```

- 基本属性：
  - ✓ ndim: 一个衡量数组维度的对象
  - ✓ shape: 一个衡量各维度大小的元组
  - ✓ dtype: 一个用于说明数组数据类型的对象

```
print(arr.ndim)
print(arr.shape)
print(arr.dtype)
```

```
2
(2, 4)
int32
```

认识和创建ndarray

- **How to create an ‘array’ ?**

Function	Description
array	Convert input data (list, tuple, array, or other sequence type) to an ndarray either by inferring a dtype or explicitly specifying a dtype. Copies the input data by default.
asarray	Convert input to ndarray, but do not copy if the input is already an ndarray
arange	Like the built-in range but returns an ndarray instead of a list.
ones, ones_like	Produce an array of all 1’s with the given shape and dtype. ones_like takes another array and produces a ones array of the same shape and dtype.
zeros, zeros_like	Like ones and ones_like but producing arrays of 0’s instead
empty, empty_like	Create new arrays by allocating new memory, but do not populate with any values like ones and zeros
eye, **identity**	Create a square N x N identity matrix (1’s on the diagonal and 0’s elsewhere)

认识和创建ndarray

ndarray数组的数据类型

- Numpy中的数据类型有int8、uint8、int16、unit16 、int32、unit32、int64、unit64、float16, float32, float64, float128, complex64, complex128, complex256, bool, object, string, unicode.
- 数据类型的转换
  - ✓ astype

```
data = np.array(['1.23','5.25','7.41'])
print(data)
print(data.astype(float))
```

```
['1.23' '5.25' '7.41']
[ 1.23  5.25  7.41]
```

  - ✓ dtype

```
data = np.array([1,2,3],dtype=np.float64) #默认是int32
print(data.dtype)
```

```
float64
```

CONTENTS

PROFESSIONAL · LEADING · VALUE-CREATING

PART 1	Numpy简介
PART 2	认识和创建ndarray
PART 2	**切片与索引**
PART 3	常用通用函数
PART 4	基本数学统计方法
PART 5	条件逻辑的矢量化表达
PART 6	集合逻辑
PART 7	线性代数
PART 8	随机数

专业来自101%的投入!

切片与索引

➢ 索引（ Indexing ）

0	1	2	3	4	5	…

一维数组的索引方式

		axis 1		
		0	**1**	**2**
	0	0,0	0,1	0,2
axis 0	**1**	1,0	1,1	1,2
	2	2,0	2,1	2,2

二维数组的索引方式

切片与索引

➢ 切片（Slicing）

- ndarray的切片是原始数组的视图，做修改时，数据不会被复制，而是直接反映到源数据上。如果想要得到切片的副本，则需要使用copy()，例如 arr[2:3].copy()。

➢ 丰富的索引和切片方式

- 基本索引和切片方式
  - ✓ 分别对如下的一维、二维、三维数组实现如下形式的切片方式，观察输出结果

```
import numpy as np
arr1d = np.arange(10)
arr2d = np.array([[1,2,3],[4,5,6]])
arr3d = np.array([[[1,2,3],[4,5,6]],[[7,8,9],[10,11,12]]])
```

[:]、 [x] 、[x : y]、 [x,y]、 [x][y] 、[x:]、 [:y] 、[:y, x:]、 [x,:y]、 [:,:y]（比如x=1,y=2 ）

切片与索引

➢ 丰富的索引和切片方式（续）

- 布尔型索引
  - ✓ 布尔型索引可以帮助我们筛选出符合条件的数据（类似Excel中的Vlookup函数）

```
GDP_Percent = np.array([7.90,7.80,7.30, 6.90,6.70])
Year = np.array([2012,2013,2014,2015,2016])
print(Year[GDP_Percent>7])
```

```
[2012 2013 2014]
```

- 花式索引（Fancy Indexing）
  - ✓ 利用整数数组进行索引，index为默认的以0开始的整数形式
    - ◆ 观察以下代码的输出结果

```
data= np.random.randn(8,4)
print(data)
print(data[[2,4,0,6]])
print(data[[-6,-4,-8,-2]])
```

CONTENTS

PROFESSIONAL · LEADING · VALUE-CREATING

PART 1	Numpy简介
PART 2	认识和创建ndarray
PART 2	切片与索引
PART 3	**常用通用函数**
PART 4	基本数学统计方法
PART 5	条件逻辑的矢量化表达
PART 6	集合逻辑
PART 7	线性代数
PART 8	随机数

专业来自101%的投入!

常用通用函数

什么是通用函数？

- 通用函数（universal function, 简称ufunc）是指Numpy中对ndarray执行元素级运算的函数。
  - ✓ ufunc支持array broadcasting, type casting等数组的标准特征

常见的ufunc

- 目前Numpy中有超过60种通用函数。其中有一些函数是内部自动调用的，比如a+b就会自动调用add(a,b)
  - ✓ ufunc可划分为数学运算符、三角函数、位操作函数、比较函数和浮点函数五大类。

常用通用函数

➢ 所有通用函数ufunc汇总表ufunc

Math operations
add
subtract
multiply
divide
logaddexp
logaddexp2
true_divide
floor_divide
negative
positive
power
remainder
mod
fmod
divmod
absolute
fabs
rint
sign
heaviside
conj
exp
exp2
log
log2
log10
expm1
log1p
sqrt
square
cbrt
reciprocal

Trigonometric functions
sin
cos
tan
arcsin
arccos
arctan
arctan2
hypot
sinh
cosh
tanh
arcsinh
arccosh
arctanh
deg2rad
rad2deg

Bit-twiddling functions
bitwise_and
bitwise_or
bitwise_xor
invert
left_shift
right_shift

Comparison functions
greater
greater_equal
less
less_equal
not_equal
equal
logical_and
logical_or
logical_xor
logical_not
maximum
minimum
fmax
fmin

Floating functions
isfinite
isinf
isnan
fabs
signbit
copysign
nextafter
spacing
modf
ldexp
frexp
fmod
floor
ceil
trunc

CONTENTS

PROFESSIONAL · LEADING · VALUE-CREATING

PART 1	Numpy简介
PART 2	认识和创建ndarray
PART 2	切片与索引
PART 3	常用通用函数
PART 4	**基本数学统计方法**
PART 5	条件逻辑的矢量化表达
PART 6	集合逻辑
PART 7	线性代数
PART 8	随机数

专业来自101%的投入!

基本数学统计方法

➢ 以下是常见的统计量

Method	Description
sum	Sum of all the elements in the array or along an axis. Zero-length arrays have sum 0.
mean	Arithmetic mean. Zero-length arrays have NaN mean.
std, var	Standard deviation and variance, respectively, with optional degrees of freedom adjustment (default denominator n).
min, max	Minimum and maximum.
argmin, argmax	Indices of minimum and maximum elements, respectively.
cumsum	Cumulative sum of elements starting from 0
cumprod	Cumulative product of elements starting from 1

基本数学统计方法

特别地，针对bool类型数据

- sum()真值计数：满足条件的对象一共有多少

```
data= np.random.normal(size=4)
print(data)
print((data>0).sum())
```

- any()或真：数组中元素是否至少有一个真
- all()与真：数组中元素是否都为真

CONTENTS

PROFESSIONAL · LEADING · VALUE-CREATING

PART 1 Numpy简介

PART 2 认识和创建ndarray

PART 2 切片与索引

PART 3 常用通用函数

PART 4 基本数学统计方法

PART 5 **条件逻辑的矢量化表达**

PART 6 集合逻辑

PART 7 线性代数

PART 8 随机数

专业来自101%的投入!

条件逻辑的矢量化表达

- **简单条件逻辑where**

```
where(condition, x, y)
```

- If condition is true :

    yield x

  else

    yield y

类似简单函数定义匿名函数
lambda

CONTENTS

PROFESSIONAL · LEADING · VALUE-CREATING

PART 1 Numpy简介

PART 2 认识和创建ndarray

PART 2 切片与索引

PART 3 常用通用函数

PART 4 基本数学统计方法

PART 5 条件逻辑的矢量化表达

PART 6 集合逻辑

PART 7 线性代数

PART 8 随机数

专业来自101%的投入!

ndarray的集合运算

Method	Description
unique(x)	Compute the sorted, unique elements in x
intersect1d(x, y)	Compute the sorted, common elements in x and y
union1d(x, y)	Compute the sorted union of elements
in1d(x, y)	Compute a boolean array indicating whether each element of x is contained in y
setdiff1d(x, y)	Set difference, elements in x that are not in y
setxor1d(x, y)	Set symmetric differences; elements that are in either of the arrays, but not both

CONTENTS

PROFESSIONAL · LEADING · VALUE-CREATING

PART 1 Numpy简介

PART 2 认识和创建ndarray

PART 2 切片与索引

PART 3 常用通用函数

PART 4 基本数学统计方法

PART 5 条件逻辑的矢量化表达

PART 6 集合逻辑

PART 7 **线性代数**

PART 8 随机数

专业来自101%的投入!

线性代数

基本计算

- 数组与标量的计算
  - ✓ 广播（broadcasting）：不同形状的数组之间的算术运算的执行方式
    - ◆ 标量与数组的计算会广播到每一个元素中
- 转置（transpose）
  - ✓ .T
- 点乘运算
  - ✓ 值得注意的是，数组之间的“*”运算不是点乘运算，点乘运算为dot()

```
arr = np.arange(5)
print(arr*arr)
print(arr.T.dot(arr))
```

```
[ 0  1  4  9 16]
30
```

➢ 常用的numpy下的线性代数函数

Function	Description
diag	Return the diagonal (or off-diagonal) elements of a square matrix as a 1D array, or convert a 1D array into a square matrix with zeros on the off-diagonal
dot	Matrix multiplication
trace	Compute the sum of the diagonal elements
det	Compute the matrix determinant
eig	Compute the eigenvalues and eigenvectors of a square matrix
inv	Compute the inverse of a square matrix
pinv	Compute the Moore-Penrose pseudo-inverse inverse of a square matrix
qr	Compute the QR decomposition
svd	Compute the singular value decomposition (SVD)
solve	Solve the linear system Ax = b for x, where A is a square matrix
lstsq	Compute the least-squares solution to y = Xb

CONTENTS

PROFESSIONAL · LEADING · VALUE-CREATING

PART 1 Numpy简介

PART 2 认识和创建ndarray

PART 2 切片与索引

PART 3 常用通用函数

PART 4 基本数学统计方法

PART 5 条件逻辑的矢量化表达

PART 6 集合逻辑

PART 7 线性代数

PART 8 **随机数**

专业来自101%的投入!

随机数

Function	Description
seed	Seed the random number generator
permutation	Return a random permutation of a sequence, or return a permuted range
shuffle	Randomly permute a sequence in place
rand	Draw samples from a uniform distribution
randint	Draw random integers from a given low-to-high range
randn	Draw samples from a normal distribution with mean 0 and standard deviation 1 (MATLAB-like interface)
binomial	Draw samples a binomial distribution
normal	Draw samples from a normal (Gaussian) distribution
beta	Draw samples from a beta distribution
chisquare	Draw samples from a chi-square distribution
gamma	Draw samples from a gamma distribution
uniform	Draw samples from a uniform [0, 1) distribution

Thank you！