

Python for Data Analysis, 3rd edition

Data Wrangling with pandas, NumPy, and Jupyter

Stephen CUI¹

January 4, 2022

¹cuixuanStephen@gmail.com

Contents

| | | |
|----------|--|-----------|
| 1 | NumPy Basics: Arrays and Vectorized Computation | 3 |
| 1.1 | The NumPy ndarray: A Multidimensional Array Object | 4 |
| 1.1.1 | Creating ndarrays | 5 |
| 1.1.2 | Data Types for ndarrays | 6 |
| 1.1.3 | Arithmetic with NumPy Arrays | 9 |
| 1.1.4 | Basic Indexing and Slicing | 9 |
| 1.1.5 | Boolean Indexing | 13 |
| 1.1.6 | Fancy Indexing | 14 |
| 1.1.7 | Transposing Arrays and Swapping Axes | 15 |
| 1.2 | Pseudorandom Number Generation(伪随机数生成) | 16 |
| 1.3 | Universal Functions: Fast Element-Wise Array Functions | 16 |
| 1.4 | Array-Oriented Programming with Arrays | 16 |
| 1.5 | Linear Algebra | 16 |
| 1.6 | Example: Random Walks | 16 |
| 1.7 | Conclusion | 16 |
| 2 | Appendix A | 17 |

Chapter 1

NumPy Basics: Arrays and Vectorized Computation

NumPy, short for Numerical Python, is one of the most important foundational packages for numerical computing in Python.

Here are some of the things you'll find in NumPy:

- `ndarray`, an efficient multidimensional array providing fast array-oriented arithmetic operations and flexible broadcasting capabilities
- Mathematical functions for fast operations on entire arrays of data without having to write loops
- Tools for reading/writing array data to disk and working with memory-mapped files
- Linear algebra, random number generation, and Fourier transform capabilities
- A C API for connecting NumPy with libraries written in C, C++, or FORTRAN

For most data analysis applications, the main areas of functionality I'll focus on are:

- Fast array-based operations for data munging and cleaning, subsetting and filtering, transformation, and any other kind of computation
- Common array algorithms like sorting, unique, and set operations
- Efficient descriptive statistics and aggregating/summarizing data
- Data alignment and relational data manipulations for merging and joining heterogeneous datasets
- Expressing conditional logic as array expressions instead of loops with `if-elif-else` branches
- Group-wise data manipulations (aggregation, transformation, and function application)

One of the reasons NumPy is so important for numerical computations in Python is because it is designed for efficiency on large arrays of data. There are a number of reasons for this:

- NumPy internally stores data in a contiguous block of memory, independent of other built-in Python objects. NumPy's library of algorithms written in the C language can operate on this memory without any type checking or other overhead. NumPy arrays also use much less memory than built-in Python sequences.
- NumPy operations perform complex computations on entire arrays without the need for Python for loops, which can be slow for large sequences. NumPy is faster than regular Python code because its C-based algorithms avoid overhead present with regular interpreted Python code.

```
1 import numpy as np
2 my_arr = np.arange(1_000_000)
3 my_list = list(range(1_000_000))
4
5 %timeit my_arr2 = my_arr * 2
6 %timeit my_list2 = [x * 2 for x in my_list]
```

1.1 The NumPy ndarray: A Multidimensional Array Object

One of the key features of NumPy is its N-dimensional array object, or ndarray, which is a fast, flexible container for large datasets in Python. Arrays enable you to perform mathematical operations on whole blocks of data using similar syntax to the equivalent operations between scalar elements.

```
1 import numpy as np
2
3 data = np.array([[1.5, -.1, 3], [0, -3, 6.5]])
4 data * 10
5 data + data
```

Notes

In this chapter and throughout the book, I use the standard NumPy convention of always using `import numpy as np`. It would be possible to put `from numpy import *` in your code to avoid having to write `np.`, but I advise against making a habit of this. The `numpy` namespace is large and contains a number of functions whose names conflict with built-in Python functions (like `min` and `max`). Following standard conventions like these is almost always a good idea.

An ndarray is a generic multidimensional container for homogeneous data; that is, all of the elements must be the same type. Every array has a `shape`, a tuple indicating the size of each dimension, and a `dtype`, an object describing the *data type* of the array:

Notes

Whenever you see “array”, “NumPy array,” or “ndarray”, in most cases they all refer to the ndarray object.

1.1.1 Creating ndarrays

The easiest way to create an array is to use the array function. This accepts any sequence-like object (including other arrays) and produces a new NumPy array containing the passed data.

```
1 data1 = [6, 7.5, 8, 0, 1]
2 arr1 = np.array(data1)
3 arr1
```

Nested sequences, like a list of equal-length lists, will be converted into a multidimensional array:

```
1 data2 = [[1, 2, 3, 4], [5, 6, 7, 8]]
2 arr2 = np.array(data2)
3 arr2
```

Unless explicitly specified (discussed in [Subsection 1.1.2](#)), `numpy.array` tries to infer a good data type for the array that it creates. The data type is stored in a special dtype metadata object.

```
1 print(arr1.dtype) # float64
2 print(arr2.dtype) # int32
```

In addition to `numpy.array`, there are a number of other functions for creating new arrays. To create a higher dimensional array with these methods, pass a tuple for the shape:

```
1 np.zeros(10)
2 np.zeros((3, 6))
3 np.empty((1, 2, 3))
```

Warnings

It’s not safe to assume that `numpy.empty` will return an array of all zeros. This function returns uninitialized memory and thus may contain nonzero “garbage” values. You should use this function only if you intend to populate(填充) the new array with data.

`numpy.arange` is an array-valued version of the built-in Python range function:

```
1 np.arange(15)
2 # array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11, 12, 13, 14])
```

Table 1.1: Some important NumPy array creation functions

| Function | Description |
|--------------------------------|--|
| <code>array</code> | Convert input data (list, tuple, array, or other sequence type) to an ndarray either by inferring a data type or explicitly specifying a data type; copies the input data by default |
| <code>asarray</code> | Convert input to ndarray, but do not copy if the input is already an ndarray |
| <code>arange</code> | Like the built-in range but returns an ndarray instead of a list |
| <code>ones, ones_like</code> | Produce an array of all 1s with the given shape and data type; <code>ones_like</code> takes another array and produces a ones array of the same shape and data type |
| <code>zeros, zeros_like</code> | Like <code>ones</code> and <code>ones_like</code> but producing arrays of 0s instead |
| <code>empty, empty_like</code> | Create new arrays by allocating new memory, but do not populate with any values like <code>ones</code> and <code>zeros</code> |
| <code>full, full_like</code> | Produce an array of the given shape and data type with all values set to the indicated “fill value” ; <code>full_like</code> takes another array and produces a filled array of the same shape and data type |
| <code>eye, identity</code> | Create a square $N \times N$ identity matrix (1s on the diagonal and 0s elsewhere) |

See [Table 1.1](#) for a short list of standard array creation functions. Since NumPy is focused on numerical computing, the data type, if not specified, will in many cases be `float64` (floating point).

1.1.2 Data Types for ndarrays

The data type or `dtype` is a special object containing the information (or metadata, data about data) the ndarray needs to interpret a chunk of memory as a particular type of data:

```

1 arr1 = np.array([1, 2, 3], dtype=np.float64)
2 arr2 = np.array([1, 2, 3], dtype=np.int32)
3 print(arr1.dtype) # float64
4 print(arr2.dtype) # int32

```

Notes

Don't worry about memorizing the NumPy data types, especially if you're a new user. It's often only necessary to care about the general kind of data you're dealing with, whether floating point, complex, integer, Boolean, string, or general Python object. When you need more control over how data is stored in memory and on disk, especially large datasets, it is good to know that you have control over the storage type.

See [Table 1.2](#) for a full listing of NumPy's supported data types.

Table 1.2: NumPy data types

| Type | Type code | Description |
|-----------------------------------|--------------|---|
| int8, uint8 | i1, u1 | Signed and unsigned 8-bit (1 byte) integer types |
| int16, uint16 | i2, u2 | Signed and unsigned 16-bit integer types |
| int32, uint32 | i4, u4 | Signed and unsigned 32-bit integer types |
| int64, uint64 | i8, u8 | Signed and unsigned 64-bit integer types |
| float16 | f2 | Half-precision floating point |
| float32 | f4 or f | Standard single-precision floating point; compatible with C float |
| float64 | f8 or d | Standard double-precision floating point; compatible with C double and Python float object |
| float128 | f16 or g | Extended-precision floating point |
| complex64, complex128, complex256 | c8, c16, c32 | Complex numbers represented by two 32, 64, or 128 floats, respectively |
| bool | ? | Boolean type storing True and False values |
| object | O | Python object type; a value can be any Python object |
| string_ | S | Fixed-length ASCII string type (1 byte per character); for example, to create astring data type with length 10, use 'S10' |
| unicode_ | U | Fixed-length Unicode type (number of bytes platform specific); same specification semantics as string_ (e.g., 'U10') |

Notes

There are both signed and unsigned integer types, and many readers will not be familiar with this terminology. A signed integer can represent both positive and negative integers, while an unsigned integer can only represent nonnegative integers. For example, `int8` (signed 8-bit integer) can represent integers from -128 to 127 (inclusive), while `uint8` (unsigned 8-bit integer) can represent 0 through 255.

You can explicitly convert or cast an array from one data type to another using `ndarray`'s `astype` method:

```
1 arr = np.array([1, 2, 3, 4, 5])
2 print(arr) # [1 2 3 4 5]
3 print(arr.dtype) # int32
4 float_arr = arr.astype(np.float64)
5 print(float_arr) # [1. 2. 3. 4. 5.]
6 print(float_arr.dtype) # float64
```

If I cast some floating-point numbers to be of integer data type, the decimal part will be truncated:

```
1 arr = np.array([3.7, -1.2, -2.6, 0.5, 12.9, 10.1])
2 print(arr)
3 arr.astype(np.int32)
4 # array([ 3, -1, -2,  0, 12, 10])
```

If you have an array of strings representing numbers, you can use `astype` to convert them to numeric form:

```
1 numeric_strings = np.array(['1.25', '-9.6', '42'], dtype=np.string_)
2 numeric_strings.astype(float)
3 # array([ 1.25, -9.6 , 42.  ])
```

Warnings

Be cautious when using the `numpy.string_` type, as string data in NumPy is fixed size and may truncate input without warning. `pandas` has more intuitive out-of-the-box behavior on non-numeric data.

If casting were to fail for some reason (like a string that cannot be converted to `float64`), a `ValueError` will be raised. Before, I was a bit lazy and wrote `float` instead of `np.float64`; NumPy aliases the Python types to its own equivalent data types.

You can also use another array's `dtype` attribute:

```
1 int_array = np.arange(10)
2
3 calibers = np.array([.22, .270, .357, .44, .50], dtype=np.float64)
```



```
4 int_array.astype(calibers.dtype)
5 # array([0., 1., 2., 3., 4., 5., 6., 7., 8., 9.]
```

There are shorthand type code strings you can also use to refer to a dtype.

Notes

Calling `astype` always creates a new array (a copy of the data), even if the new data type is the same as the old data type.

1.1.3 Arithmetic with NumPy Arrays

Arrays are important because they enable you to express batch operations on data without writing any for loops. NumPy users call this *vectorization*.

- Any arithmetic operations between equal-size arrays apply the operation element-wise
- Arithmetic operations with scalars propagate the scalar argument to each element in the array
- Comparisons between arrays of the same size yield Boolean arrays

```
1 arr = np.array([[1., 2., 3.], [4., 5., 6.]])
2 arr * arr
3 arr - arr
4
5 1 / arr
6 arr ** 2
7
8 arr2 = np.array([[0., 4., 1.], [7., 2., 12.]])
9 arr2 > arr
```

Evaluating operations between differently sized arrays is called broadcasting and will be discussed in more detail in [Appendix A](#).

1.1.4 Basic Indexing and Slicing

NumPy array indexing is a deep topic, as there are many ways you may want to select a subset of your data or individual elements.

```
1 arr = np.arange(10)
2 arr[5]
3 # 5
4
```

```

5  arr[5: 8]
6  # array([5, 6, 7])
7
8  arr[5: 8] = 12
9  arr
10 # array([ 0,  1,  2,  3,  4, 12, 12, 12,  8,  9])

```

Notes

An important first distinction from Python's built-in lists is that **array slices are views on the original array**. This means that the data is not copied, and any modifications to the view will be reflected in the source array.

```

1  arr_slice = arr[5: 8]
2  arr_slice
3  # array([12, 12, 12])
4  arr_slice[1] = 123
5  arr
6  # array([ 0,  1,  2,  3,  4, 12, 123, 12,  8,  9])
7
8  arr_slice[:] = 64
9  arr
10 # array([ 0,  1,  2,  3,  4, 64, 64, 64,  8,  9])

```

Pythonn内置的列表

```

1  a = list(range(10))
2  print(a)
3  # [0, 1, 2, 3, 4, 5, 6, 7, 8, 9]
4  b = a[2: 5]
5  # [2, 3, 4]
6  print(b)
7  b[1] = 1234
8  print(b)
9  # [2, 1234, 4]
10 print(a)
11 # [0, 1, 2, 3, 4, 5, 6, 7, 8, 9]

```

| | | Axis 1 | | |
|--------|---|--------|-----|-----|
| | | 0 | 1 | 2 |
| Axis 0 | 0 | 0,0 | 0,1 | 0,2 |
| | 1 | 1,0 | 1,1 | 1,2 |
| | 2 | 2,0 | 2,1 | 2,2 |

Figure 1.1: Indexing elements in a NumPy array

As NumPy has been designed to be able to work with very large arrays, you could imagine performance and memory problems if NumPy insisted on always copying data.

Warnings

If you want a copy of a slice of an ndarray instead of a view, you will need to explicitly copy the array—for example, `arr[5:8].copy()`. As you will see, pandas works this way, too.

With higher dimensional arrays, you have many more options. In a two-dimensional array, the elements at each index are no longer scalars but rather one-dimensional arrays. Thus, individual elements can be accessed recursively. But that is a bit too much work, so you can pass a comma-separated list of indices to select individual elements.

```
1 arr2d = np.array([[1, 2, 3], [4, 5, 6], [7, 8, 9]])
2 arr2d[2]
3 # array([7, 8, 9])
4
5 # these are equivalent:
6 arr2d[0][2]
7 arr2d[0, 2]
```

See [Figure 1.1](#) for an illustration of indexing on a two-dimensional array. I find it helpful to think of axis 0 as the “rows” of the array and axis 1 as the “columns.”

In multidimensional arrays, if you omit later indices, the returned object will be a lower dimensional ndarray consisting of all the data along the higher dimensions.

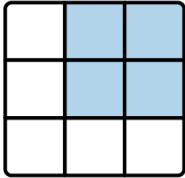
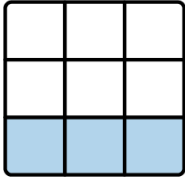
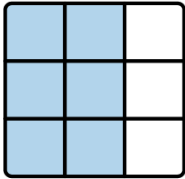
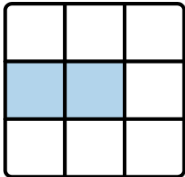
| | Expression | Shape |
|--|---------------------------|--------------------|
|  | <code>arr[:2,1:]</code> | <code>(2,2)</code> |
|  | <code>arr[2]</code> | <code>(3,)</code> |
| | <code>arr[2, :]</code> | <code>(3,)</code> |
| | <code>arr[2:, :]</code> | <code>(1,3)</code> |
|  | <code>arr[:, :2]</code> | <code>(3,2)</code> |
|  | <code>arr[1, :2]</code> | <code>(2,)</code> |
| | <code>arr[1:2, :2]</code> | <code>(1,2)</code> |

Figure 1.2: Two-dimensional array slicing

1

Note that in all of these cases where subsections of the array have been selected, the returned arrays are views.

Warnings

This multidimensional indexing syntax for NumPy arrays will not work with regular Python objects, such as lists of lists.

Indexing with slices

Like one-dimensional objects such as Python lists, ndarrays can be sliced with the familiar syntax.

By mixing integer indexes and slices, you get lower dimensional slices.

```
1 lower_dim_slice = arr2d[1, :2]
```

```
2 lower_dim_slice.shape
```

3

```

4 arr2d[:2, 2]
5 arr2d[:, :1]
6
7 arr2d[:2, 1:] = 0
8 arr2d

```

1.1.5 Boolean Indexing

The Boolean array must be of the same length as the array axis it's indexing. You can even mix and match Boolean arrays with slices or integers (or sequences of integers).

```

1 names = np.array(["Bob", "Joe", "Will", "Bob", "Will", "Joe", "Joe"])
2 print(names)
3 data = np.array([[4, 7], [0, 2], [-5, 6], [0, 0], [1, 2], [-12, 4], [3, 4]])
4 print(data)
5 print(names == "Bob")
6 print(data[names == "Bob"])
7 print(data[names == "Bob", 1:])
8 print(data[names == "Bob", 1])

```

- To select everything but "Bob" you can either use `!=` or negate the condition using `~`
- The `~` operator can be useful when you want to invert a Boolean array referenced by a variable
- To select two of the three names to combine multiple Boolean conditions, use Boolean arithmetic operators like `&` (and) and `|` (or).

```

1 names = np.array(["Bob", "Joe", "Will", "Bob", "Will", "Joe", "Joe"])
2 names
3 data = np.array([[4, 7], [0, 2], [-5, 6], [0, 0], [1, 2], [-12, 4], [3, 4]])
4 data
5 names == "Bob"
6 data[names == "Bob"]
7 data[names == "Bob", 1:]
8 data[names == "Bob", 1]
9
10 names != "Bob"
11 ~(names == "Bob")
12 data[~(names == "Bob")]

```

```
13 cond = names == "Bob"
14 data[~cond]
15
16 mask = (names == "Bob") | (names == "Will")
17 data[mask]
```

Selecting data from an array by Boolean indexing and assigning the result to a new variable always creates a copy of the data, even if the returned array is unchanged.

Warnings

The Python keywords `and` and `or` do not work with Boolean arrays. Use `&` (and) and `|` (or) instead.

- Setting values with Boolean arrays works by substituting the value or values on the righthand side into the locations where the Boolean array's values are True.
- You can also set whole rows or columns using a one-dimensional Boolean array

```
1 data[data < 0] = 0
2 data
3
4 data[names != "Joe"] = 7
5 data
```

1.1.6 Fancy Indexing

Fancy indexing is a term adopted by NumPy to describe indexing using integer arrays.

- To select a subset of the rows in a particular order, you can simply pass a list or ndarray of integers specifying the desired order. Using negative indices selects rows from the end
- Passing multiple index arrays does something slightly different; it selects a one-dimensional array of elements corresponding to each tuple of indices.(返回的是一个一维的数组，这超乎了我的想象)

```
1 arr = np.zeros((8, 4))
2 for i in range(8):
3     arr[i] = i
4 arr
5
6 arr[[4, 3, 0, 6]]
7
```

```

8  arr[[-3, -5, -7]]
9
10 arr = np.arange(32).reshape((8, 4))
11 arr[[1, 5, 7, 2], [0, 3, 1, 2]]

```

To learn more about the reshape method, have a look at [Appendix A](#). Here the elements (1, 0), (5, 3), (7, 1), and (2, 2) were selected. The result of fancy indexing with as many integer arrays as there are axes is always one-dimensional.

如果你想要返回一个二维的数组，你写的索引位置应该也得是一个矩形数组，下面是一种写法：

```

1  arr[[1, 5, 7, 2]][:, [0, 3, 1, 2]]

```

Keep in mind that fancy indexing, unlike slicing, always copies the data into a new array when assigning the result to a new variable. If you assign values with fancy indexing, the indexed values will be modified.

```

1  a = arr[[1, 5, 7, 2], [0, 3, 1, 2]]
2  a[0] = 10000
3  print(a)
4  print(arr)

```

1.1.7 Transposing Arrays and Swapping Axes

Transposing is a special form of reshaping that similarly returns a view on the underlying data without copying anything.

- Arrays have the transpose method and the special T attribute. When doing matrix computations, you may do this very often. The @ infix operator is another way to do matrix multiplication.
- Simple transposing with .T is a special case of swapping axes. ndarray has the method `swapaxes`, which takes a pair of axis numbers and switches the indicated axes to rearrange the data. 只返回源数据的视图，不会复制数据。
- 对于高维数组，`transpose`需要得到一个由轴编号组成的元组才能对这些轴进行转置（费脑子，不要用图像理解）。用代数的方法理解，原本 $x_{103} = 11$ ，轴变换后 $x_{310} = 11$ 或者 $(1,0,3) = x_{103}$ ，变换后 $(3,1,0) = x_{103}$ 。

```

1  arr = np.arange(15).reshape((3, 5))
2  arr.T
3
4  arr = np.array([[0, 1, 0], [1, 2, -2], [6, 3, 2], [-1, 0, -1], [1, 0, 1]])
5  np.dot(arr.T, arr)

```

```
6  arr.T @ arr
7
8  arr.swapaxes(0, 1)
9
10 arr = np.arange(16).reshape((2, 2, 4))
11 arr.transpose((2, 0, 1))
```

1.2 Pseudorandom Number Generation(伪随机数生成)

1.3 Universal Functions: Fast Element-Wise Array Functions

1.4 Array-Oriented Programming with Arrays

1.5 Linear Algebra

1.6 Example: Random Walks

1.7 Conclusion

Chapter 2

Appendix A