python_tools_ds

May 12, 2020

1 Python tools for data science

1.1 The PyData Stack

The Python Data Stack comprises a set of packages that makes Python a powerful data science language. These include

- Numpy: provides arrays and matrix algebra
- Scipy: provides scientific computing capabilities
- matplotlib: provides graphing capabilities

These were the original stack that was meant to replace Matlab. However, these were meant to tackle purely numerical data, and the kinds of heterogeneous data we regularly face needed more tools. These were added more recently.

- Pandas: provides data analytic structures like the data frame, as well as basic descriptive statistical capabilities
- statsmodels: provides a fairly comprehensive set of statistical functions
- scikit-learn: provides machine learning capabilities

This is the basic stack of packages we will be using in this workshop. Additionally we will use a few packages that add some functionality to the data science process. These include

- seaborn: Better statistical graphs
- plotly: Interactive graphics
- biopython: Python for bioinformatics

We may also introduce the package `rpy2` which allows one to run R from within Python. This can be useful since many bioinformatic pipelines are already implemented in R.

> The PyData stack also includes `sympy`, a symbolic mathematics package emulating Maple

1.2 Numpy (numerical and scientific computing)

We start by importing the Numpy package into Python using the alias `np`.

[ ]:
```
import numpy as np
```

Numpy provides both arrays (vectors) and vectorized functions which are very fast. Let's see how this works.

[ ]:
```
z = [1,2,3,4,5,6,7,8,9.3,10.6] # This is a list
z_array = numpy.array(z)
z_array
```

Now, we have already seen functions in Python earlier. In Numpy, there are functions that are optimized for arrays, that can be accessed directly from the array objects. This is an example of *object-oriented programming* in Python, where functions are provided for particular *classes* of objects, and which can be directly accessed from the objects. We will use several such functions over the course of this workshop, but we won't actually talk about how to do this program development here.

We apply the functions `sum`, `min` (minimum value) and `max` (maximum value) to `z_array`.

[ ]:
```
z_array.sum()
```

[ ]:
```
z_array.min()
```

[ ]:
```
z_array.max()
```

The versions of these functions in Numpy are optimized for arrays and are quite a bit faster than the corresponding functions available in base Python. When doing data work, these are the preferred functions.

These functions can also be used in the usual function manner:

[ ]:
```
np.max(z_array)
```

Calling `np.max` ensures that we are using the `max` function from numpy, and not the one in base Python.

1.2.1 Numpy data types

Numpy arrays are homogeneous in type

[ ]:
```
y = [1,3,'a']
np.array(y)
```

So what's going on here? Upon conversion from a heterogeneous list, numpy converted the numbers into strings. This is necessary since, by definition, numpy arrays can hold data of a single type. When one of the elements is a string, numpy casts all the other entities into strings as well. Think about what would happen if the opposite rule was used. The string 'a' doesn't have a corresponding number, while both numbers 1 and 3 have corresponding string representations, so going from string to numeric would create all sorts of problems.

1.2.2 Generating data in numpy

We had seen earlier how we could generate a sequence of numbers in a list using `range`. In numpy, you can generate a sequence of numbers in an array using `arange`

[ ]:
```
np.arange(10)
```

You can also generate regularly spaced sequences of numbers between particular values

[ ]:
```
np.linspace(start=0, stop=1, num=11) # or np.linspace(0, 1, 11)
```

You can generate an array of 0's

[ ]:
```
np.zeros(10)
```

This can easily be extended to a two-dimensional array (a matrix), by specifying the dimension of the matrix as a tuple.

[ ]:
```
np.zeros((10,10))
```

You can also generate a matrix of 1s in a similar manner.

[ ]:
```
np.ones((3,4))
```

In matrix algebra, the identity matrix is important. It is a square matrix with 1's on the diagonal and 0's everywhere else.

[ ]:
```
np.eye(4)
```

Random numbers Generating random numbers is quite useful in many areas of data science. All computers don't produce truly random numbers but generate *pseudo-random* sequences. These are completely deterministic sequences defined algorithmically that emulate the properties of random numbers. Since these are deterministic, we can set a *seed* or starting value for the sequence, so that we can exactly reproduce this sequence to help debug our code. To actually see how things behave in simulations we will often run several sequences of random numbers starting at different seed values.

The seed is set by the `RandomState` function within the `random` submodule of numpy. Note that all Python names are case-sensitive.

[ ]:
```
rng = np.random.RandomState(35) # set seed
rng.randint(0, 10, (3,4))
```

We have created a 3x4 matrix of random integers between 0 and 10 (in line with slicing rules, this includes 0 but not 10).

We can also create a random sample of numbers between 0 and 1.

[ ]:
```
rng.random_sample((5,2))
```

We'll see later how to generate random numbers from particular probability distributions.

1.2.3 Vectors and matrices

Numpy generates arrays, which can be of arbitrary dimension. However the most useful are vectors (1-d arrays) and matrices (2-d arrays).

In these examples, we will generate samples from the Normal (Gaussian) distribution, with mean 0 and variance 1.

[ ]:
```
A = rng.normal(0,1,(4,5))
```

We can compute some characteristics of this matrix's dimensions. The number of rows and columns are given by `shape`.

[ ]:
```
A.shape
```

The total number of elements are given by `size`.

[ ]:
```
A.size
```

If we want to create a matrix of 0's with the same dimensions as `A`, we don't actually have to compute its dimensions. We can use the `zeros_like` function to figure it out.

[ ]:
```
np.zeros_like(A)
```

We can also create vectors by only providing the number of rows to the random sampling function. The number of columns will be assumed to be 1.

[ ]:
```
B = rng.normal(0, 1, (5,))
B
```

1.2.4 Extracting elements from arrays

The syntax for extracting elements from arrays is almost exactly the same as for lists, with the same rules for slices.

Exercise: State what elements of B are extracted by each of the following statements

```
B[:3]
B[:-1]
B[[0,2,4]]
B[[0,2,5]]
```

For matrices, we have two dimensions, so you can slice by rows, or columns or both.

[ ]:
```
A
```

We can extract the first column by specifying `:` (meaning everything) for the rows, and the index for the column (reminder, Python starts counting at 0)

[ ]:
```
A[:,0]
```

Similarly the 4th row can be extracted by putting the row index, and `:` for the column index.

[ ]:
```
A[3,:]
```

All slicing operations work for rows and columns

[ ]:
```
A[:2,:2]
```

1.2.5 Beware of copies

One has to be a bit careful with copying objects in Python. By default, if you just assign one object to a new name, it does a *shallow copy*, which means that both names point to the same memory. So if you change something in the original, it also changes in the new copy.

[ ]:
```
A[0,:]
```

[ ]:
```
A1 = A
A1[0,0] = 4
A[0,0]
```

To actually create a copy that is not linked back to the original, you have to make a *deep copy*, which creates a new space in memory and a new pointer, and copies the original object to the new memory location

[ ]:
```
A1 = A.copy()
A1[0,0] = 6
A[0,0]
```

You can also replace sub-matrices of a matrix with new data, provided that the dimensions are compatible. (Make sure that the sub-matrix we are replacing below truly has 2 rows and 2 columns, which is what `np.eye(2)` will produce)

[ ]:
```
A[:2,:2] = numpy.eye(2)
A
```

1.2.6 Operations on matrices

You can sum all the elements of a matrix using `sum`. You can also sum along rows or along columns by adding an argument to the `sum` function.

[ ]:
```
A = rng.normal(0, 1, (4,2))
A.sum()
```

You can sum along rows with the option `axis = 0`

[ ]:
```
A.sum(axis=0)
```

You can sum along columns with `axis = 1`.

[ ]:
```
A.sum(axis=1)
```

Of course, you can use the usual function calls: `np.sum(A, axis = 1)`

Logical/Boolean operations You can query a matrix to see which elements meet some criterion. In this example, we'll see which elements are negative.

[ ]:
```
A < 0
```

This is called **masking**, and is useful in many contexts.

We can extract all the negative elements of A using

[ ]:
```
A[A<0]
```

This forms a 1-d array. You can also count the number of elements that meet the criterion

[ ]:
```
np.sum(A<0)
```

Since the entity `A<0` is a matrix as well, we can do row-wise and column-wise operations as well.

1.3 Broadcasting in Python

Python deals with arrays in an interesting way, in terms of matching up dimensions of arrays for arithmetic operations. There are 3 rules:

1. If two arrays differ in the number of dimensions, the shape of the smaller array is padded with 1s on its *left* side
2. If the shape doesn't match in any dimension, the array with shape = 1 in that dimension is stretched to match the others' shape
3. If in any dimension the sizes disagree and none of the sizes are 1, then an error is generated

[ ]:
```
A.shape
```

[ ]:
```
B.shape
```

[ ]:
```
A - B
```

B is 1-d, A is 2-d, so B's shape is made into (1,5) (added to the left). Then it is repeated into 4 rows to make it's shape (4,5), then the operation is performed. This means that we subtract the first element of B from the first column of A, the second element of B from the second column of A, and so on.

This can be very useful, since these operations are faster than for loops. For example:

[ ]:
```
d = rng.random_sample((10,2))
d
```

We want to find the Euclidean distance (the sum of squared differences) between the points defined by the rows. This should result in a 10x10 distance matrix

[ ]:
```
d.shape
```

[ ]:
```
d[np.newaxis,:,:]
```

creates a 3-d array with the first dimension being of length 1

[ ]:
```
d[np.newaxis,:,:].shape
```

[ ]:
```
d[:, np.newaxis,:]
```

creates a 3-d array with the 2nd dimension being of length 1

[ ]:
```
d[:,np.newaxis,:].shape
```

Now for the trick, using broadcasting of arrays. These two arrays are incompatible without broadcasting, but with broadcasting, the right things get repeated to make things compatible

[ ]:
```
dist_sq = np.sum((d[:,np.newaxis,:] - d[np.newaxis,:,:]) ** 2)
```

[ ]:
```
dist_sq.shape
```

[ ]:
```
dist_sq
```

Whoops! we wanted a 10x10 matrix, not a scalar.

[ ]:
```
(d[:,np.newaxis,:] - d[np.newaxis,:,:]).shape
```

What we really want is the 10x10 distance matrix.

[ ]:
```
dist_sq = np.sum((d[:,np.newaxis,:] - d[np.newaxis,:,:]) ** 2, axis=2)
```

You can verify what is happening by creating `D = d[:,np.newaxis,:]-d[np.newaxis,:,:]` and then looking at `D[:,:,0]` and `D[:,:,1]`. These are the difference between each combination in the first and second columns of d, respectively. Squaring and summing along the 3rd axis then gives the sum of squared differences.

[ ]:
```
dist_sq
```

[ ]:
```
dist_sq.shape
```

[ ]:
```
dist_sq.diagonal()
```

1.3.1 Conclusions moving forward

It's important to understand numpy and arrays, since most data sets we encounter are rectangular. The notations and operations we saw in numpy will translate to data, except for the fact that data is typically heterogeneous, i.e., of different types. The problem with using numpy for modern data analysis is that if you have mixed data types, it will all be coerced to strings, and then you can't actually do any data analysis.

The solution to this issue (which is also present in Matlab) came about with the `pandas` package, which is the main workhorse of data science in Python