PHYS4038/MLiS and ASI/MPAGS

Scientific Programming in python

mpags-python.github.io

Steven Bamford
An introduction to scientific programming with Python

Session 4: Numerical Python
In this session:

• Creating and using numerical arrays
• Plotting with matplotlib
• Exercises

• Talk to me:
  • During teaching sessions (preferred)
    • Specific questions, clarifications – just ask
    • Bigger issues – wait until end of lecture / start of examples class
  • Remote students connect via skype group:
    • https://join.skype.com/KpW5oCLNNiJt
    • text during lecture, video during examples class
  • Via email: steven.bamford@nottingham.ac.uk
  • Arrange a meeting
• So far…
  • core Python language and libraries

• Extra features required:
  • fast, multidimensional arrays
  • plotting tools
  • libraries of fast, reliable scientific functions (*to come…*)
• Lists ok for storing small amounts of one-dimensional data

```python
>>> a = [1, 3, 5, 7, 9]
>>> print(a[2:4])
[5, 7]
>>> b = [[1, 3, 5, 7, 9], [2, 4, 6, 8, 10]]
>>> print(b[0])
[1, 3, 5, 7, 9]
>>> print(b[1][2:4])
[6, 8]
```
Numpy – array creation and use

```python
>>> import numpy
>>> l = [[1, 2, 3], [3, 6, 9], [2, 4, 6]]  # create a list
>>> a = np.array(l)  # convert a list to an array
>>> print(a)
[[1 2 3]
 [3 6 9]
 [2 4 6]]
>>> a.shape
(3, 3)
>>> print(a.dtype)  # get type of an array
int32
>>> print(a[0])  # this is just like a list of lists
[1 2 3]
>>> print(a[1, 2])  # arrays can be given comma separated indices
9
>>> print(a[1, 1:3])  # and slices
[6 9]
>>> print(a[:,1])
[2 6 4]
```
Numpy – array creation and use

```python
>>> a[1, 2] = 7
>>> print(a)
[[1 2 3]
 [3 6 7]
 [2 4 6]]

>>> a[:, 0] = [0, 9, 8]
>>> print(a)
[[0 2 3]
 [9 6 7]
 [8 4 6]]

>>> b = np.zeros(5)
>>> print(b)
[ 0.  0.  0.  0.  0.]
>>> b.dtype
dtype('float64')

>>> n = 1000
>>> my_int_array = np.zeros(n, dtype=np.int)
>>> my_int_array.dtype
dtype('int32')
```
Numpy – array creation and use

```python
>>> c = np.ones(4)
>>> print(c)
[ 1.  1.  1.  1. ]

>>> d = np.arange(5)  # just like range()
>>> print(d)
[0 1 2 3 4]

>>> d[1] = 9.7
>>> print(d)  # arrays keep their type even if elements changed
[0 9 2 3 4]
>>> print(d*0.4)  # operations create a new array, with new type
[ 0.  3.6  0.8  1.2  1.6]

>>> d = np.arange(5, dtype=np.float)
>>> print(d)
[ 0.  1.  2.  3.  4.]

>>> np.arange(3, 7, 0.5)  # arbitrary start, stop and step
array([ 3.,  3.5,  4.,  4.5,  5.,  5.5,  6.,  6.5])
```
Numpy – array creation and use

```python
>>> a = np.arange(4.0)
>>> b = a * 23.4
>>> c = b/(a+1)
>>> c += 10
>>> print c
[ 10.  21.7  25.6  27.55]

>>> arr = np.arange(100, 200)
>>> select = [5, 25, 50, 75, -5]
>>> print(arr[select])  # can use integer lists as indices
[105, 125, 150, 175, 195]

>>> arr = np.arange(10, 20)
>>> div_by_3 = arr%3 == 0  # comparison produces boolean array
>>> print(div_by_3)
[False False True False False True False False True False]
>>> print(arr[div_by_3])  # can use boolean lists as indices
[12 15 18]
```
>>> b = arr[1:].reshape((3,3))  # now 2d 3x3 array
>>> print b
[[11 12 13]
 [14 15 16]
 [17 18 19]]

>>> b_2 = b%2 == 0
>>> b_3 = b%3 == 0
>>> b_2_3 = b_2 & b_3  # boolean operators
>>> print b_2_3  # also logical_and(b_2, b_3)
[[False True False]
 [False False False]
 [False True False]]

>>> print b[b_2_3]  # select array elements
[12 18]

>>> i_2_3 = b_2_3.nonzero()  # with boolean arrays
>>> print i_2_3
(array([0, 2]), array([1, 1]))

>>> print b[i_2_3]  # or index arrays
[12 18]
Numpy – array methods

>>> arr.sum()
145
>>> arr.mean()
14.5
>>> arr.std()
2.8722813232690143
>>> arr.max()
19
>>> arr.min()
10
>>> div_by_3.all()
False
>>> div_by_3.any()
True
>>> div_by_3.sum()
3
>>> div_by_3.nonzero()  # note singleton tuple returned
(array([2, 5, 8]),)  # for consistency with N-dim case
>>> arr = np.array([4.5, 2.3, 6.7, 1.2, 1.8, 5.5])
>>> arr.sort()  # acts on array itself
>>> print(arr)
[ 1.2  1.8  2.3  4.5  5.5  6.7]

>>> x = np.array([4.5, 2.3, 6.7, 1.2, 1.8, 5.5])
>>> y = np.array([1.5, 2.3, 4.7, 6.2, 7.8, 8.5])
>>> np.sort(x)
array([ 1.2,  1.8,  2.3,  4.5,  5.5,  6.7])
>>> print(x)
[ 4.5  2.3  6.7  1.2  1.8  5.5]
>>> s = x.argsort()
>>> s
array([3, 4, 1, 0, 5, 2])
>>> x[s]
array([ 1.2,  1.8,  2.3,  4.5,  5.5,  6.7])
>>> y[s]
array([ 6.2,  7.8,  2.3,  1.5,  8.5,  4.7])
Numpy – array functions

Many element-by-element math, trig., etc. operations
  • e.g., add(x1, x2), absolute(x), log10(x), sin(x), logical_and(x1, x2)

```python
>>> np.log10(x)
array([0.65321251, 0.36172784, 0.8260748 , 0.07918125, 0.25527251,
      0.74036269])

>>> np.sin(x)
array([-0.97753012,  0.74570521,  0.40484992,  0.93203909,
       0.97384763, -0.70554033])

>>> np.add(x, y)  # but would normally use x + y
array([ 6. ,  4.6, 11.4,  7.4,  9.6, 14. ])
```
Use regular numpy arrays as matrices and vectors

```python
>>> a = np.array([[1, 0],
...                [0, 1]])
>>> b = np.array([[4, 1],
...                [2, 2]])

>>> np.matmul(a, b)
array([[4, 1],
       [2, 2]])

>>> a @ b
array([[4, 1],
       [2, 2]])

>>> c = np.array([1, 2])
>>> a @ c
array([1, 2])

>>> b.T
array([[4, 2],
       [1, 2]])

>>> np.linalg.det(b)
6.0

>>> np.linalg.eig(b)
(array([4.73205081, 1.26794919]),
 array([[ 0.80689822, -0.34372377],
        [ 0.59069049, 0.9390708 ]]))
```
Numpy – array functions

Most array methods have equivalent functions

```python
>>> arr.sum()  # array method
45
>>> np.sum(arr)  # array function
45
```

Array functions often return a result, leaving original array unchanged

Array methods often perform the operation in-place

```python
>>> a = np.array([23, 7, 80])
>>> s = np.sort(a)  # returns sorted array
>>> print a, s
[23 7 80] [7 23 80]
>>> a.sort()  # nothing returned
>>> print a  # operation applied in-place
[7 23 80]
```
Many array functions (and methods) can take an *axis*, with the operation only applied along that one direction in the array.

<table>
<thead>
<tr>
<th>Expression</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>a.sum()</code></td>
<td>135</td>
</tr>
<tr>
<td><code>a.sum(axis=0)</code></td>
<td><code>array([48, 45, 42])</code></td>
</tr>
<tr>
<td><code>a.sum(axis=1)</code></td>
<td><code>array([54, 45, 36])</code></td>
</tr>
</tbody>
</table>
| `np.sort(a)` | `array([[17, 18, 19],
          [14, 15, 16],
          [11, 12, 13]])` |
| `np.sort(a, axis=0)` | `array([[13, 12, 11],
         [16, 15, 14],
         [19, 18, 17]])` |
| `np.sort(a, axis=1)` | `array([[17, 18, 19],
         [14, 15, 16],
         [11, 12, 13]])` |

Defaults are to operate on the whole array (axis=None) for accumulative operations and on the highest dimension (axis=-1) otherwise.
Numpy – random numbers

High quality (pseudo-) random number generator with many common distributions

```python
>>> np.random.seed(12345)  # or default seed taken from clock
>>> np.random.uniform()
0.9296160928171479

>>> np.random.uniform(-1, 1, 3)
array([-0.36724889, -0.63216238, -0.59087944])

>>> r = np.random.normal(loc=3.0, scale=1.3, size=100)
>>> r.mean(), r.std()
(3.1664506480570371, 1.2754634208344433)

>>> p = np.random.poisson(123, size=(1024, 1024))
>>> p.shape
(1024, 1024)
>>> p.mean(), p.std()**2
(123.02306461334229, 122.99512022056578)
```
Arrays usually have homogeneous type, but different type arrays can be combined – with a recarray

But better to use Pandas or Astropy Tables (see later…)

```python
>>> x = np.arange(0,100)
>>> y = np.sqrt(x)
>>> z = y.astype(np.int)
>>> r = np.rec.array((x,y,z), names=('x', 'y', 'z'))
>>> r.x
array([ 0,  1,  2, ..., 9997, 9998, 9999])
>>> r.y
array([ 0.        ,   1.        ,   1.41421356, ...,  99.98499887, 99.9899995 ,  99.99499987])
>>> r.z
array([ 0,  1,  1, ..., 99, 99, 99])
```
Numpy – loading and saving data

- Custom binary format:
  - save
  - load

- Text format:
  - savetxt
  - loadtxt
  - genfromtxt
  - recfromcsv

- Not very portable, not self-describing
- Better to use FITS and HDF5 (see later…)

```python
>>> np.savetxt('mydata', r, fmt='%6i %12.6f %6i') # save to file

>>> data = np.genfromtxt('mydata') # reads a 2d array

>>> data = np.recfromtxt('myfile.txt', names=('x', 'y', 'z'))
```

$ head mydata

0     0.000000      0
1     1.000000      1
2     1.414214      1
3     1.732051      1
4     2.000000      2
5     2.236068      2
6     2.449490      2
7     2.645751      2
8     2.828427      2
9     3.000000      3
Numpy – using arrays wisely

- Array operations are implemented in C or Fortran
- Optimised algorithms - i.e. fast!
- Python loops (i.e. for i in a:...) are much slower
- Prefer array operations over loops, especially when speed important
- Also produces shorter code, often much more readable

- If you're working with large datasets, watch out for swapping…
Numpy – saving memory

- Numpy arrays reside entirely in memory

- Save memory by using lower precision where possible

```python
>>> d = np.arange(100000000, dtype=np.int32)  # default int64
>>> d = np.arange(1e8, dtype=np.float32)    # default float64
```

- Save memory by performing operations in place where possible

```python
>>> a = np.arange(100000000)  # 1e8 * 64 / 8 / 1e6 ~ 800Mb
>>> b = np.random.normal(0, 1000, 100000000)  # also ~ 800 Mb
>>> a = a + b  # requires additional 800Mb memory (maybe swap)
>>> a += b    # in-place: no more memory required and faster
>>> a = np.sqrt(a)  # requires extra 800Mb memory
>>> np.sqrt(a, out=a)  # in-place: no more memory required
```

- Use sparse arrays (provided by SciPy, see later…)

- Use a solution which keeps data on disk (np.memmap, PyTables)

- Change your algorithm
Plotting – matplotlib

- User friendly, but powerful, plotting capabilities for python
- [http://matplotlib.org/](http://matplotlib.org/)

Once installed, to use type:

```python
>>> import pylab  # handy for interactive use
>>> import matplotlib.pyplot as plt  # better for in scripts
```

- Settings can be customised by editing `~/.matplotlib/matplotlibrc`
  - Set `backend` and default font, colours, layout, etc.

- Helpful website
  - many examples

```python
>>> plt.ion()  # turn on interactive mode!
```
Plotting – matplotlib

```python
>>> from numpy import sin, cos, pi
>>> x = np.arange(0, 2*pi, pi/100)
>>> y = sin(x)*cos(2*x)
>>> plt.plot(x, y)
>>> plt.plot(x, sin(x), '--r')
>>> plt.plot(x, cos(2*x),
   linestyle='dotted',
   color='green')
>>> thresh = y > 0.75
>>> plt.plot(x[thresh], y[thresh],
   ...   'r', linewidth=5)
>>> zeros = np.abs(y) < pi/200
>>> plt.plot(x=zeros, y=zeros,
   ...   'ok', markersize=10)
>>> plt.xlabel(r'$x$')
>>> plt.ylabel(r'$\sin(x)\cos(2x)$')
>>> plt.axis([0, 2*pi, -1.1, 1.1])
>>> plt.savefig('wiggles.pdf')
```
Plotting – matplotlib

- Plots can be altered in an object oriented manner

For example,

```python
>>> fig = plt.figure(1)
>>> ax = fig.axes[0]
>>> ax.xaxis.labelpad = 10
>>> plt.draw()
>>> l = ax.lines[2]
>>> l.set_linewidth(3)
>>> plt.draw()
>>> ax.xaxis.set_ticks((0, pi/2, pi, 3*pi/2, 2*pi))
>>> plt.draw()
>>> ax.xaxis.set_ticklabels(('0', r'$\frac{1}{2}\pi$', r'$\pi$', r'$\frac{3}{2}\pi$', r'$2\pi$'))
>>> plt.draw()
>>> plt.subplots_adjust(bottom=0.25)
```

Shorthand to get current axes
```python
>>> ax = plt.gca()
```
Some useful functions:
• figure – create a new figure, or get an existing figure object
• plot – add line or points
• hist / hist2d – create a 1D/2D histogram
• imshow / contour – plot an array as an image / contours
• axis – set axis limits
• subplots – create new figure with a grid of subplots
• subplots_adjust – adjust the canvas margins

Some functions update the plot, others don't (for efficiency)
To update the plot display:
• draw() – draw plot and continue
• show() – blocks interpreter until window closed
• close(), close('all') – close figure windows
Matplotlib/Ipython notebook
scatter plot example

[link to online notebook]
Plotting – interactive notebook plots

• There are some tools for producing interactive plots in a web browser (via JavaScript), and hence in IPython notebooks

• matplotlib
  • %matplotlib inline – inserts image of plot in notebook
  • %matplotlib notebook – inserts interactive plot in notebook

• mpld3
  • use matplotlib commands (e.g. some existing plotting code)
  • generate HTML with mpld3 – automatically get pan and zoom
  • optionally add interactivity (tooltips, highlighting, selections, …)

• bokeh, plotly, …
  • similar functionality, but different (non-matplotlib) interface
Plotting – alternative interfaces

• Large variety of different approaches

• seaborn, pandas – easy, sophisticated statistical plots
• plotnine – grammar of graphics (ggplot2) interface
• bokeh, plotly – web targeted
• datashader, yt – for large datasets of points, densities
• altair, vega – declarative visualisation

• Making sense of the deluge:
  • https://www.youtube.com/watch?v=FytuB8nFHPQ
Plotting – astronomical data

- AplPy: plotting library for astronomical images
- Based on matplotlib
- http://aplpy.github.com/

```python
import aplpy
import numpy

gc = aplpy.FITSFigure('fits/2MASS_k.fits', figsize=(10, 9))
gc.show_rgb('graphics/2MASS_arcsinh_color.png')
gc.set_tick_labels_font(size='small')
gc.show_contour('fits/mips_24micron.fits', colors='white')
data = np.loadtxt('data/yso_wcs_only.txt')
ra, dec = data[:, 0], data[:, 1]
gc.show_markers(ra, dec, layer='scatter_set_1', edgecolor='red', facecolor='none', marker='o', s=10, alpha=0.5)
gc.save('tutorial.png')
```

Also WCSAxes from astropy
Coursework submission

- Submission and feedback via your GitHub repository
- Mandatory for MLiS, optional for MPAGS
- Create a branch called sub1
- Should contain a README file including:
  - your full name and university
  - possibly some background (basic explanation, references, …)
  - an overview of the intended functionality of your program
  - ideas of the modules you plan to use
  - ideas of the structure of your code (functions, etc.)
  - possibly snippets or pseudocode
  - any remaining uncertainties or questions
Questions and exercises

Any questions?

- shout and wave
- skype (spbamford)
  - https://join.skype.com/KpW5oCLNNiJt
- email steven.bamford@nottingham.ac.uk

Exercises?

Get started on them now.

I'll be around for support.

Solutions are online.

I will go through them either later in this session or in the next one.
Exercises 4

1) Create an array \( x = [-3.00, -2.99, -2.98, \ldots, 2.98, 2.99, 3.00] \)

2) Create a corresponding array for the Gaussian function

\[
y = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}
\]

3) Check the result is unit normalised:

\[
\sum_i y_i \delta x = 1
\]

4) For convenience, put \( x \) and \( y \) together in a recarray and save to a file

5) Create a sample of one hundred Gaussian random numbers

6) Plot your Gaussian curve, \( x \) versus \( y \), with axis labels

7) Add a histogram of the random numbers to your plot

8) Over-plot another histogram for a different sample and prettify (try histtype='stepfilled' and 'step', and transparency, e.g., alpha=0.5)

Any questions?

- skype (spbamford): [https://join.skype.com/KpW5oCLNNiJt](https://join.skype.com/KpW5oCLNNiJt)
- email steven.bamford@nottingham.ac.uk