Lecture 2

April 24, 2019

1 Lecture Notes 2: Numpy, Timing, Plotting

2 Numpy

2.1 Basics

```
In [5]: # Import the module such that we can use the built-in functionality
        import numpy
```

2.1.1 Numpy arrays

```
In [6]: X = numpy.array([1, 2, 3, 4])
        Y = numpy.array([5, 6, 7, 8])
```

2.1.2 Operations between arrays

```
In [7]: A = X + Y             # element-wise addition
        M = X * Y             # element-wise multiplication
        D = numpy.dot(X, Y)   # dot product
        T = X.T               # transposing
        X_tail = X[2:]        # indexing (similar to lists)

        A, M, D, T, X_tail

Out[7]: (array([ 6,  8, 10, 12]),
         array([ 5, 12, 21, 32]),
         70,
         array([1, 2, 3, 4]),
         array([3, 4]))

In [8]: # Compare this to operations on lists
        X_list = [1, 2, 3, 4]
        Y_list = [5, 6, 7, 8]
        print(X_list + Y_list)
        print(X_list * Y_list) #  -> raises Exception

[1, 2, 3, 4, 5, 6, 7, 8]
```

```
---------------------------------------------------------------------------

TypeError                                 Traceback (most recent call last)

<ipython-input-8-669fbdcb86b4> in <module>()
      3 Y_list = [5, 6, 7, 8]
      4 print(X_list + Y_list)
----> 5 print(X_list * Y_list) #  -> raises Exception


TypeError: can't multiply sequence by non-int of type 'list'
```

2.1.3 Equivalent operations with lists

```
In [9]: A_list = [x + y for x, y in zip(X, Y)]       # element-wise addition
        M_list = [x * y for x, y in zip(X, Y)]       # element-wise multiplication
        D_list = sum([x * y for x, y in zip(X, Y)]) # dot product

        A_list, M_list, D_list
```

```
Out[9]: ([6, 8, 10, 12], [5, 12, 21, 32], 70)
```

Observation: Results are the same, but the Numpy syntax is much more readable (i.e. more compact) than the Python syntax for the same vector operations.

2.1.4 Shapes of arrays

```
In [10]: print(A.shape, M.shape, D.shape)
```

```
(4,) (4,) ()
```

2.2 Matrices

```
In [11]: A = numpy.array(
                [
                    [1, 2, 3],
                    [4, 5, 6]
                ]
          )
```

```
In [12]: print(A)
         print(10 * "-")
         print(A.shape)
```

```
[[1 2 3]
 [4 5 6]]
```

```
----------
(2, 3)
```

```
In [13]: # Elementwise multiplication
         A * A

Out[13]: array([[ 1,  4,  9],
                [16, 25, 36]])

In [14]: # Matrix-matrix multiplication
         numpy.dot(A, A.T)

Out[14]: array([[14, 32],
                [32, 77]])
```

Observation: Unlike Matlab, "*" denotes an element-wise multiplication. Matrix multiplication is instead implemented by the function "dot".

```
In [15]: numpy.dot(A, A)  # -> raises Exception


        ---------------------------------------------------------------------------

        ValueError                                Traceback (most recent call last)

        <ipython-input-15-6491280b970b> in <module>()
    ----> 1 numpy.dot(A, A)  # -> raises Exception


        ValueError: shapes (2,3) and (2,3) not aligned: 3 (dim 1) != 2 (dim 0)
```

2.3 Performance evaluation

To verify that in addition to the more compact syntax, Numpy also provides a computational benefit over standard Python, we compare the running time of a similar computation performed in pure Python and in Numpy. The module "time" provides a function "clock" to measure the current time.

```
In [16]: import time
         time.clock()

Out[16]: 1.38571
```

we now wait a little bit...

```
In [17]: time.clock()

Out[17]: 1.393106
```

and can observed that the value is higher than before (time has passed). We now define two functions to test the speed of matrix multiplication for two $n \times n$ matrices.

```
In [18]: # pure Python implementation

         def benchmark_python(n):

             # initialization
             X = numpy.ones((n, n))
             Y = numpy.ones((n, n))
             Z = numpy.zeros((n, n))

             # actual matrix multiplication
             start = time.clock()
             for i in range(n):
                 for j in range(n):
                     for k in range(n):
                         Z[i,j] += X[i, k] * Y[k, j]
             end = time.clock()

             return end-start
```

```
In [19]: # Numpy implementation

         def benchmark_numpy(n):

             # initialization
             X = numpy.ones((n, n))
             Y = numpy.ones((n, n))
             Z = numpy.zeros((n, n))

             # actual matrix multiplication
             start = time.clock()
             Z = numpy.dot(X, Y)
             end = time.clock()

             return end-start
```

Evaluating this function for $n = 100$ iterations, we can observe that Numpy is much faster than pure Python.

```
In [20]: num_iterations = 100
         benchmark_python(num_iterations), benchmark_numpy(num_iterations)

Out[20]: (0.556764, 0.0013250000000000206)
```

```
In [21]: ### Common alternative way of importing Numpy: alias "np"
         import numpy as np
         print(np.ones((3, 3)))
```

```
[[1. 1. 1.]
 [1. 1. 1.]
 [1. 1. 1.]]
```

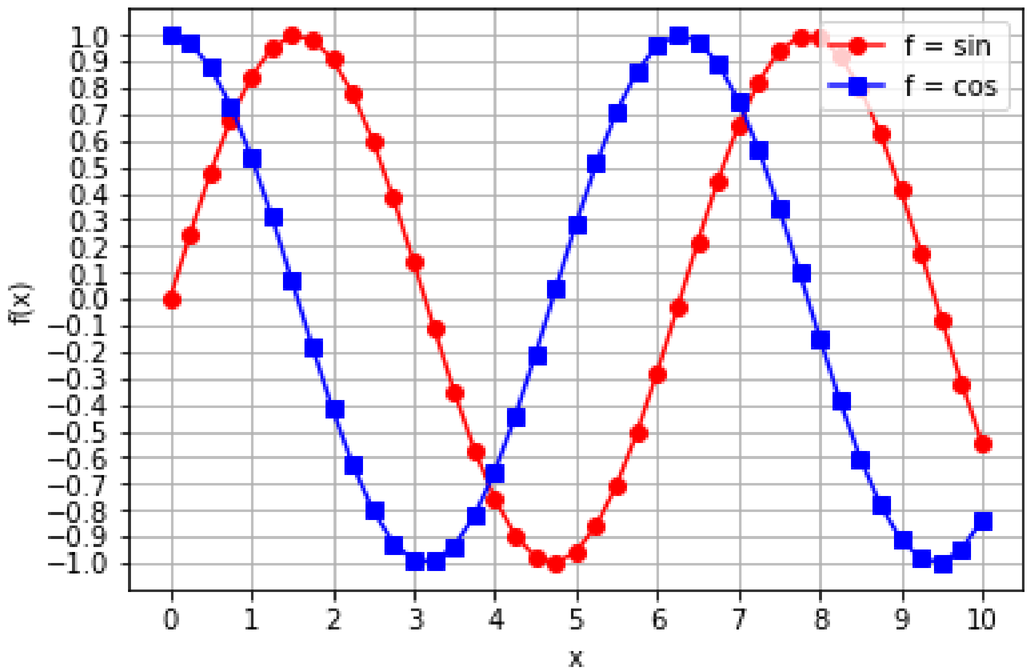
3 Plotting

In machine learning, it is often necessary to visualize the data, or to plot properties of algorithms such as their accuracy or their speed. For this, we can make use of the matplotlib library, which we load with the following sequence of commands.

```
In [22]: import matplotlib
         import matplotlib.pyplot as plt
         # Needed in Jupyter Notebook
         %matplotlib inline
```

3.1 Basic plot

```
In [23]: x = numpy.arange(0, 10.001, 0.25)
         y = numpy.sin(x)
         z = numpy.cos(x)

         plt.plot(x, y, 'o-', color='red', label='f = sin')
         plt.plot(x, z, 's-', color='blue', label='f = cos')

         plt.legend(loc = 'upper right')

         xtks = np.arange(0, 10.01, 1)
         ytks = np.arange(-1,1.01,0.1)
         plt.xticks(xtks)
         plt.yticks(ytks)

         plt.xlabel('x')
         plt.ylabel('f(x)')
         plt.grid(True)
```

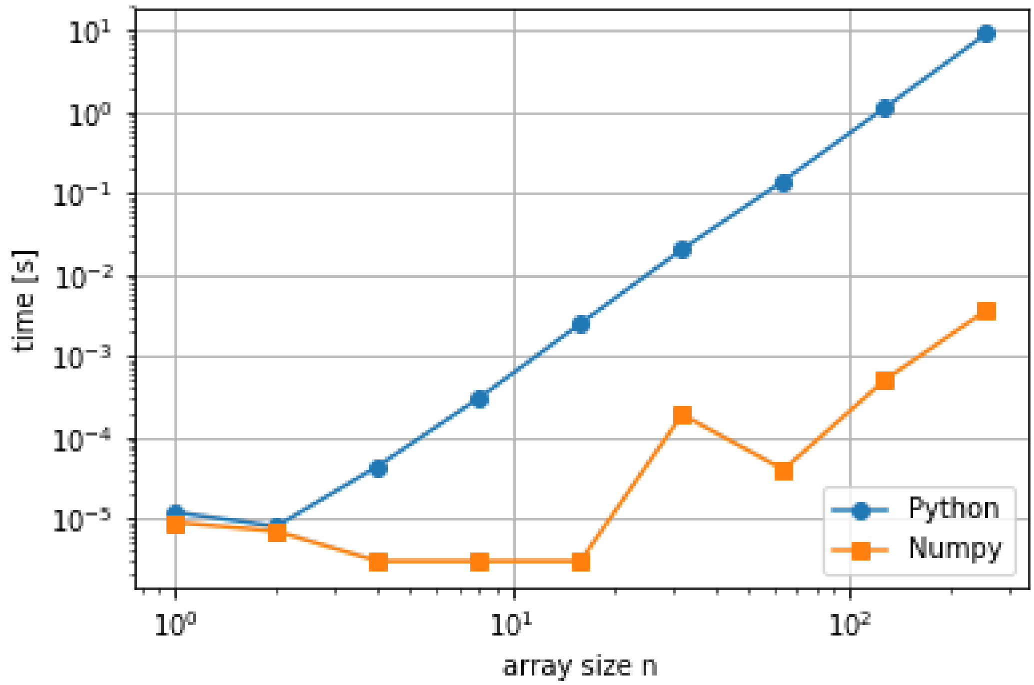


Plotting a performance curve for matrix multiplication
We run the computation with different parameters (e.g. size of input arrays)

```
In [24]: N = [1, 2, 4, 8, 16, 32, 64, 128, 256]

         py_t = [benchmark_python(n) for n in N]
         np_t = [benchmark_numpy(n) for n in N]
```

Then, we render the plot

```
In [25]: plt.plot(N, py_t, 'o-', label='Python')
         plt.plot(N, np_t, 's-', label='Numpy')
         plt.grid(True)
         plt.xscale('log')
         plt.yscale('log')
         plt.xlabel('array size n')
         plt.ylabel('time [s]')
         plt.legend(loc='lower right')

Out[25]: <matplotlib.legend.Legend at 0x7fcc67afaa20>
```



3.2 Advanced Numpy

Special Array Initializations

Numpy arrays can be initialized to specific values (`numpy.zeros`, `numpy.ones`, …). Special numpy arrays (e.g. diagonal, identity) can be created easily.

```
In [26]: A = numpy.zeros((3, 3))          # array of size 2x2 filled with zeros
         B = numpy.ones((3, 3))           # same, but filled with ones
         C = numpy.diag((1.0, 2.0, 3.0)) # diagonal matrix
         D = numpy.eye(3)                 # identity matrix
         E = numpy.random.rand(3, 3)      # random numbers
         F = numpy.triu(B)                # upper triagonal matrix

         print(A)
         print(B)
         print(C)
         print(D)
         print(E)
```

```
[[0. 0. 0.]
 [0. 0. 0.]
 [0. 0. 0.]]
[[1. 1. 1.]
 [1. 1. 1.]
```

```
 [1. 1. 1.]]
[[1. 0. 0.]
 [0. 2. 0.]
 [0. 0. 3.]]
[[1. 0. 0.]
 [0. 1. 0.]
 [0. 0. 1.]]
[[0.96354617 0.37010909 0.10944295]
 [0.45992084 0.70310845 0.80665066]
 [0.8516851  0.87890791 0.59880082]]
```

Array type

```
In [27]: A = numpy.ones((2, 2))
         type(A), A.shape, A.size, A.ndim, A.dtype

Out[27]: (numpy.ndarray, (2, 2), 4, 2, dtype('float64'))

In [28]: A = numpy.ones((3,3,3))
         type(A), A.shape, A.size, A.ndim, A.dtype

Out[28]: (numpy.ndarray, (3, 3, 3), 27, 3, dtype('float64'))
```

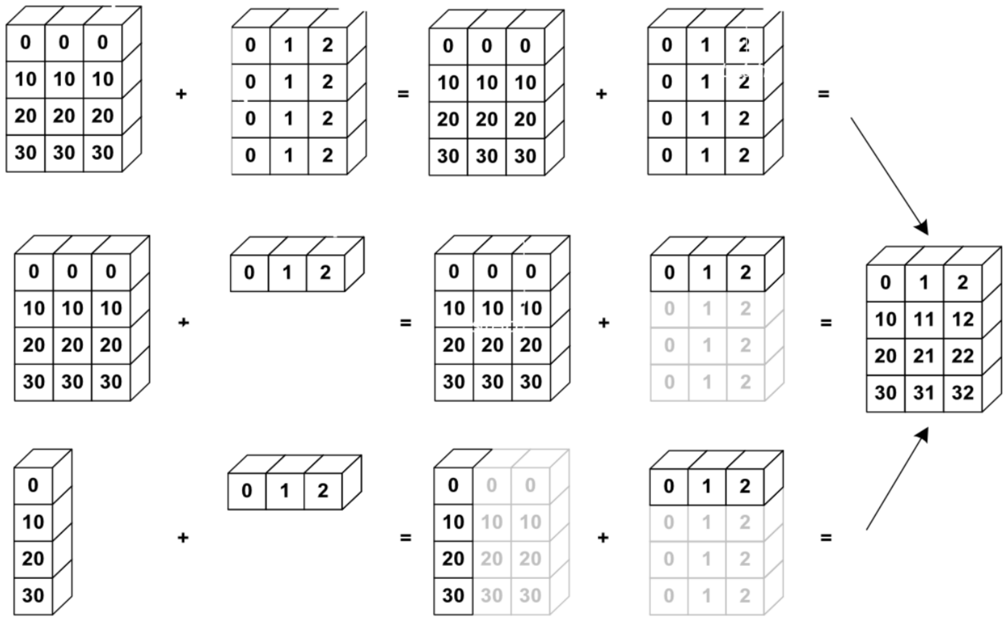
Casting

An array can be explicitly forced to have elements of a certain type (e.g. half-precision). When applying an operator to two arrays of different types, the returned array retains the type of the highest-precision input array (here, float64).

```
In [29]: E = A.astype('float32')
         A.dtype, E.dtype, (A + E).dtype

Out[29]: (dtype('float64'), dtype('float32'), dtype('float64'))
```

Reshaping and transposing

```
In [30]: A = numpy.array([[1, 2, 3], [4, 5, 6]])

         print(A)
         print(A.reshape((3,2)))
         print(A.ravel())
         print(A.T)

[[1 2 3]
 [4 5 6]]
[[1 2]
 [3 4]
 [5 6]]
[1 2 3 4 5 6]
[[1 4]
 [2 5]
 [3 6]]
```



Numpy broadcasting

Numpy broadcasting

Broadcasting

See also https://docs.scipy.org/doc/numpy/user/basics.broadcasting.html

```
In [31]: numpy.ones((3, 2)) + 1

Out[31]: array([[2., 2.],
                [2., 2.],
                [2., 2.]])

In [32]: numpy.ones((3, 2)) + numpy.ones((3, 2))

Out[32]: array([[2., 2.],
                [2., 2.],
                [2., 2.]])

In [33]: numpy.ones((3, 1)) + numpy.ones((1, 2))

Out[33]: array([[2., 2.],
                [2., 2.],
                [2., 2.]])

In [34]: numpy.ones((3, 1)) + numpy.ones((2))

Out[34]: array([[2., 2.],
                [2., 2.],
                [2., 2.]])
```

Indexing
See also https://docs.scipy.org/doc/numpy/reference/arrays.indexing.html

```
In [35]: A = numpy.arange(30).reshape(6, 5)
         print(A)
```

```
[[ 0  1  2  3  4]
 [ 5  6  7  8  9]
 [10 11 12 13 14]
 [15 16 17 18 19]
 [20 21 22 23 24]
 [25 26 27 28 29]]
```

Select rows/columns

```
In [36]: print(A[3, :])
         print(A[:, 3])
```

```
[15 16 17 18 19]
[ 3  8 13 18 23 28]
```

Select window

```
In [37]: print(A[1:5, 1:4])
```

```
[[ 6  7  8]
 [11 12 13]
 [16 17 18]
 [21 22 23]]
```

Select even rows and odd columns

```
In [38]: print(A[::2, 1::2])
```

```
[[ 1  3]
 [11 13]
 [21 23]]
```

Select last two columns

```
In [39]: print(A[:, -2:])
```

```
[[ 3  4]
 [ 8  9]
 [13 14]
 [18 19]
 [23 24]
 [28 29]]
```

Select column 1 and 4

```
In [40]: print(A[:, [1, 4]])
```

```
[[ 1  4]
 [ 6  9]
 [11 14]
 [16 19]
 [21 24]
 [26 29]]
```

3.3 Boolean Arrays

```
In [41]: a = numpy.random.rand(4, 4)
         print(a)
         b = a > 0.5
         print(b)
         print(b.astype(int))
         print(a[b])
```

```
[[0.6384684  0.31712013 0.82203648 0.33102264]
 [0.05813107 0.93787492 0.00378267 0.15911665]
 [0.97148463 0.53602779 0.06128071 0.04622343]
 [0.29966374 0.03329309 0.53367259 0.43153876]]
[[ True False  True False]
 [False  True False False]
 [ True  True False False]
 [False False  True False]]
[[1 0 1 0]
 [0 1 0 0]
 [1 1 0 0]
 [0 0 1 0]]
[0.6384684  0.82203648 0.93787492 0.97148463 0.53602779 0.53367259]
```

```
In [42]: # Is any/all of the elements True?
         numpy.any(b), numpy.all(b)
```

```
Out[42]: (True, False)
```

```
In [43]: # Apply to specific axes only
         numpy.any(b, axis=1), numpy.all(b, axis=0)
```

```
Out[43]: (array([ True,  True,  True,  True]), array([False, False, False, False]))
```

4 Analyzing a Dataset

Let's load the Boston dataset (506 examples composed of 13 features each).

```
In [44]: # extract two interesting features of the data
         from sklearn.datasets import load_boston
         boston = load_boston()
         print(boston.keys())

         X = boston['data']
         F = boston['feature_names']

         print(F)
```

```
dict_keys(['data', 'target', 'feature_names', 'DESCR'])
['CRIM' 'ZN' 'INDUS' 'CHAS' 'NOX' 'RM' 'AGE' 'DIS' 'RAD' 'TAX' 'PTRATIO'
 'B' 'LSTAT']
```

Reduce-type operations

```
In [45]: print(X.mean())                                   # Global dataset mean feature value
         print(X[:, 0].mean())                             # Mean of first feature (CRIM)
         print(X.mean(axis=0), X.mean(axis=0).shape)       # Mean of all features
         print(X.std(axis=0), X.std(axis=0).shape)         # Standard deviation of all features
         print(X.min(), X.max())                           # Extreme values
         print(X.shape, X.sum(axis=1).shape, X.sum(axis=1, keepdims=True).shape)
```

```
70.0724468257829
3.5937607114624512
[3.59376071e+00 1.13636364e+01 1.11367787e+01 6.91699605e-02
 5.54695059e-01 6.28463439e+00 6.85749012e+01 3.79504269e+00
 9.54940711e+00 4.08237154e+02 1.84555336e+01 3.56674032e+02
 1.26530632e+01] (13,)
[8.58828355e+00 2.32993957e+01 6.85357058e+00 2.53742935e-01
 1.15763115e-01 7.01922514e-01 2.81210326e+01 2.10362836e+00
 8.69865112e+00 1.68370495e+02 2.16280519e+00 9.12046075e+01
 7.13400164e+00] (13,)
0.0 711.0
(506, 13) (506,) (506, 1)
```

```
In [46]: # Show the feature name along with the mean and standard deviation
         list(zip(F, X.mean(axis=0), X.std(axis=0)))
```

```
Out[46]: [('CRIM', 3.593760711462451, 8.588283547653553),
          ('ZN', 11.363636363636363, 23.299395694766027),
          ('INDUS', 11.136778656126504, 6.853570583390873),
          ('CHAS', 0.0691699604743083, 0.25374293496034855),
          ('NOX', 0.5546950592885372, 0.11576311540656153),
          ('RM', 6.284634387351787, 0.7019225143345692),
          ('AGE', 68.57490118577078, 28.121032570236885),
          ('DIS', 3.795042687747034, 2.103628356344459),
```

```
('RAD', 9.549407114624506, 8.698651117790645),
('TAX', 408.2371541501976, 168.3704950393814),
('PTRATIO', 18.455533596837967, 2.162805191482142),
('B', 356.67403162055257, 91.20460745217272),
('LSTAT', 12.653063241106723, 7.134001636650485)]
```

Retain two interesting features (5 and 12)

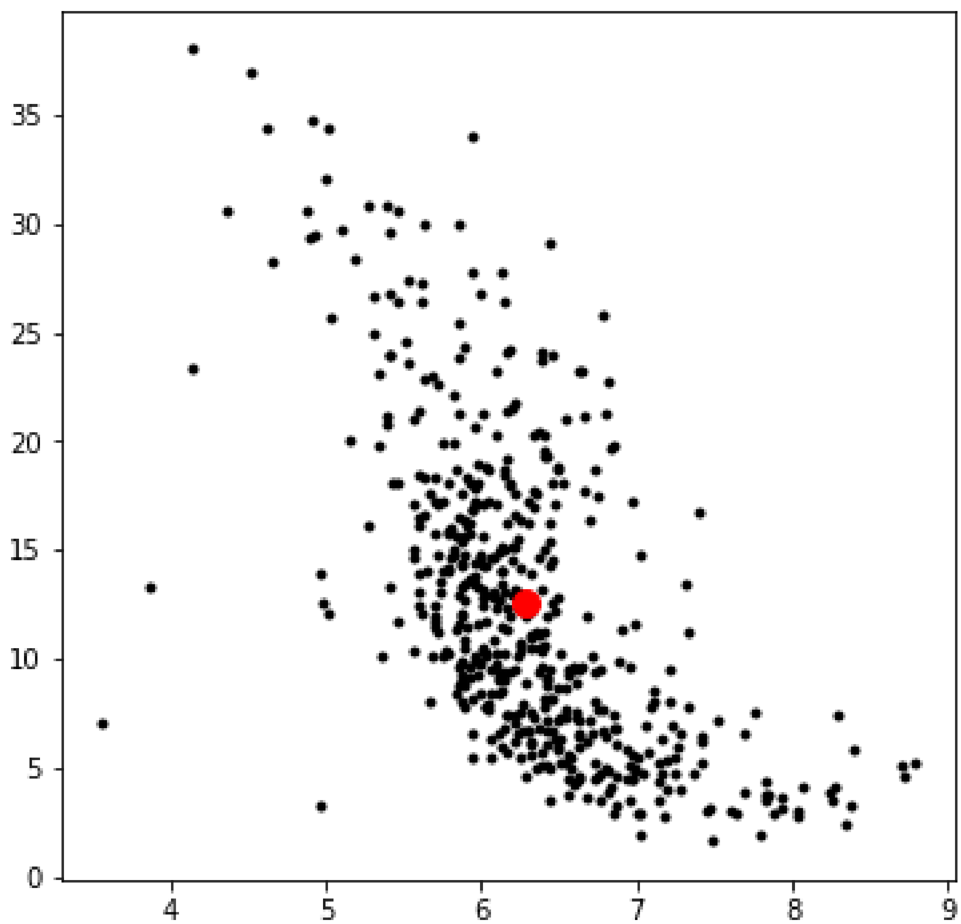
```
In [47]: X = X[:, [5, 12]]
         print(X.shape)
```

```
(506, 2)
```

Scatter-plot the first two dimensions

```
In [48]: plt.figure(figsize=(6, 6))
         plt.plot(X[:, 0], X[:, 1], 'o', color='black', ms=3)
         plt.plot(X[:, 0].mean(), X[:, 1].mean(), 'o', color='red', ms=10)
```

```
Out[48]: [<matplotlib.lines.Line2D at 0x7fcc4a178a20>]
```

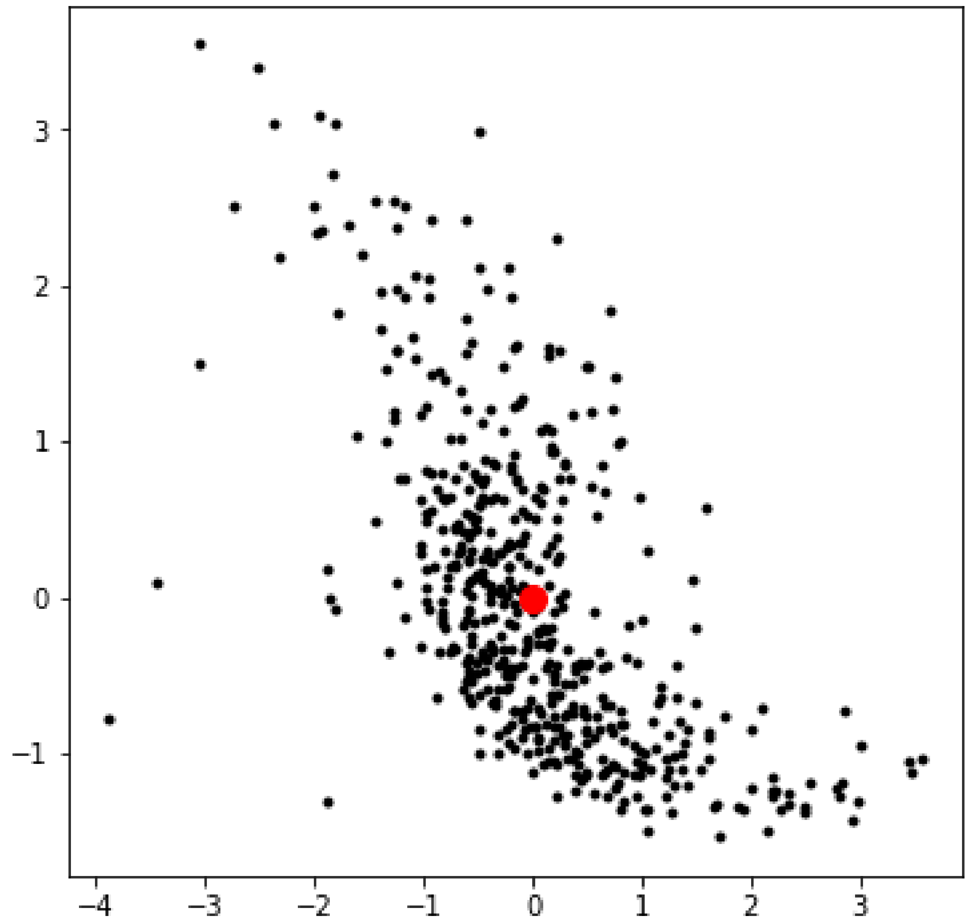


Normalize the data

```
In [49]: X_norm = X - X.mean(axis=0) # center around origin
         X_norm /= X.std(axis=0)  # rescale features so that they have standard deviation 1 in e
```

```
In [50]: plt.figure(figsize=(6, 6))
         plt.plot(X_norm[:, 0], X_norm[:, 1], 'o', color='black', ms=3)
         plt.plot(X_norm[:, 0].mean(), X_norm[:, 1].mean(), 'o', color='red', ms=10)
```

```
Out[50]: [<matplotlib.lines.Line2D at 0x7fcc49d07fd0>]
```



Computing a distance matrix

```
In [51]: import scipy
         import scipy.spatial

         D = scipy.spatial.distance.cdist(X_norm, X_norm)
```

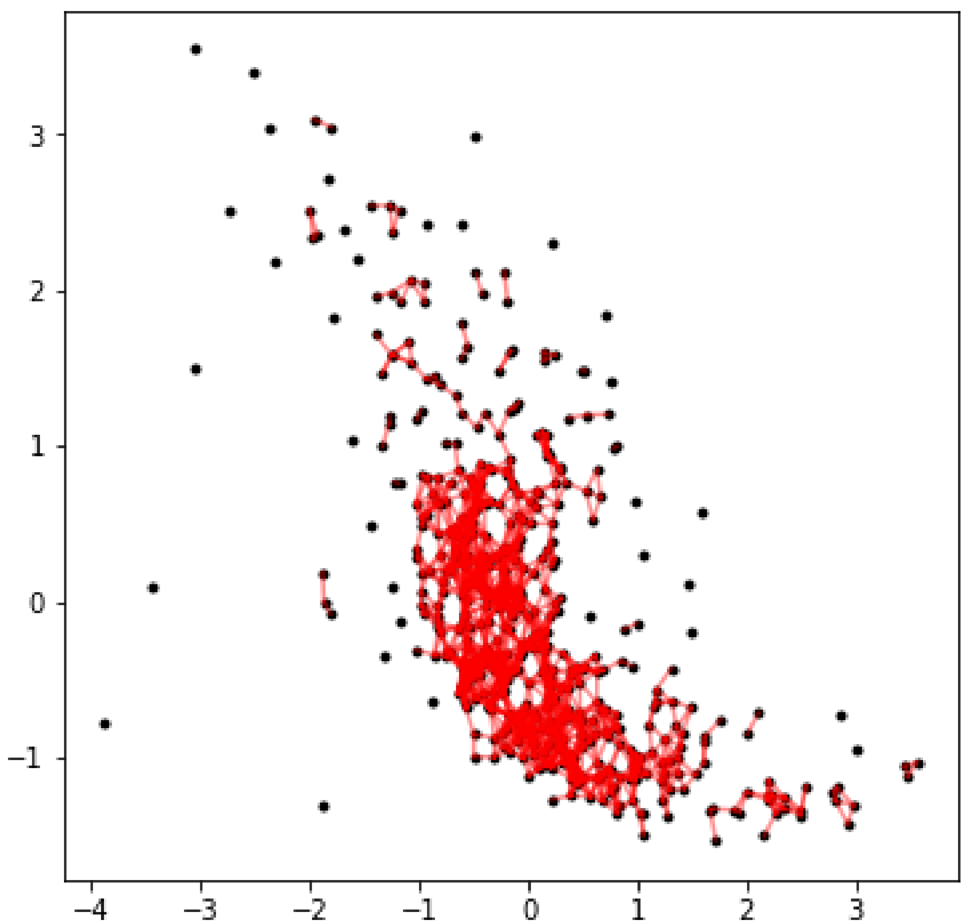
alternative way of computing a distance matrix:

```
In [52]: Dalt = np.sqrt(((X_norm ** 2).sum(axis=1).reshape((1, len(X_norm))) \
                + (X_norm ** 2).sum(axis=1).reshape((len(X_norm), 1)) \
                - 2 * numpy.dot(X_norm, X_norm.T)) + 1e-7)

         print(((Dalt - D) ** 2).mean())
```

```
1.9764174161180842e-10
```

Highlighting nearby data points

```
In [53]: plt.figure(figsize=(6, 6))

         ind = numpy.where(D < 0.2)

         plt.plot(X_norm[:, 0], X_norm[:, 1], 'o', color='black', ms=3)

         for i1,i2 in zip(*ind):
             plt.plot([X_norm[i1, 0], X_norm[i2, 0]], [X_norm[i1, 1], X_norm[i2, 1]], color='red
```



4.1 Getting help

```
In [54]: help(numpy.where)
```

```
Help on built-in function where in module numpy:

where(...)
    where(condition, [x, y])

    Return elements chosen from `x` or `y` depending on `condition`.

    .. note::
        When only `condition` is provided, this function is a shorthand for
        ``np.asarray(condition).nonzero()``. Using `nonzero` directly should be
        preferred, as it behaves correctly for subclasses. The rest of this
        documentation covers only the case where all three arguments are
        provided.

    Parameters
    ----------
    condition : array_like, bool
        Where True, yield `x`, otherwise yield `y`.
    x, y : array_like
        Values from which to choose. `x`, `y` and `condition` need to be
        broadcastable to some shape.

    Returns
    -------
    out : ndarray
        An array with elements from `x` where `condition` is True, and elements
        from `y` elsewhere.

    See Also
    --------
    choose
    nonzero : The function that is called when x and y are omitted

    Notes
    -----
    If all the arrays are 1-D, `where` is equivalent to::

        [xv if c else yv
         for c, xv, yv in zip(condition, x, y)]

    Examples
```

```
--------
>>> a = np.arange(10)
>>> a
array([0, 1, 2, 3, 4, 5, 6, 7, 8, 9])
>>> np.where(a < 5, a, 10*a)
array([ 0,  1,  2,  3,  4, 50, 60, 70, 80, 90])

This can be used on multidimensional arrays too:

>>> np.where([[True, False], [True, True]],
...          [[1, 2], [3, 4]],
...          [[9, 8], [7, 6]])
array([[1, 8],
       [3, 4]])

The shapes of x, y, and the condition are broadcast together:

>>> x, y = np.ogrid[:3, :4]
>>> np.where(x < y, x, 10 + y)  # both x and 10+y are broadcast
array([[10,  0,  0,  0],
       [10, 11,  1,  1],
       [10, 11, 12,  2]])

>>> a = np.array([[0, 1, 2],
...               [0, 2, 4],
...               [0, 3, 6]])
>>> np.where(a < 4, a, -1)  # -1 is broadcast
array([[ 0,  1,  2],
       [ 0,  2, -1],
       [ 0,  3, -1]])
```