Numpy

Numpy is the fundamental package for numeric computing with Python. It provides powerful ways to create, store, and/or manipulate data, which makes it able to seamlessly and speedily integrate with a wide variety of databases. This is also the foundation that Pandas is built on, which is a high-performance data-centric package that we will learn later in the course.

In this lecture, we will talk about creating array with certain data types, manipulating array, selecting elements from arrays, and loading dataset into array. Such functions are useful for manipulating data and understanding the functionalities of other common Python data packages.

In [ ]:
```
# You'll recall that we import a library using the `import` keyword as numpy's comm
import numpy as np
import math
```

Array Creation

In [ ]:
```
# Arrays are displayed as a list or list of lists and can be created through list c
# array, we pass in a list as an argument in numpy array
a = np.array([1, 2, 3])
print(a)
# We can print the number of dimensions of a list using the ndim attribute
print(a.ndim)
```

In [ ]:
```
# If we pass in a list of lists in numpy array, we create a multi-dimensional array
b = np.array([[1,2,3],[4,5,6]])
b
```

In [ ]:
```
# We can print out the length of each dimension by calling the shape attribute, wh
b.shape
```

In [ ]:
```
# We can also check the type of items in the array
a.dtype
```

In [ ]:
```
# Besides integers, floats are also accepted in numpy arrays
c = np.array([2.2, 5, 1.1])
c.dtype.name
```

In [ ]:
```
# Let's look at the data in our array
c
```

In [ ]:
```
# Note that numpy automatically converts integers, like 5, up to floats, since ther
# Numpy will try and give you the best data type format possible to keep your data
# means all the same, in the array
```

In [ ]:
```
# Sometimes we know the shape of an array that we want to create, but not what we w
# offers several functions to create arrays with initial placeholders, such as zero
# Lets create two arrays, both the same shape but with different filler values
d = np.zeros((2,3))
```

```
print(d)

e = np.ones((2,3))
print(e)
```

In [ ]:
```
# We can also generate an array with random numbers
np.random.rand(2,3)
```

In [ ]:
```
# You'll see zeros, ones, and rand used quite often to create example arrays, espec
# posts and other forums.
```

In [ ]:
```
# We can also create a sequence of numbers in an array with the arrange() function
# starting bound and the second argument is the ending bound, and the third argumer
# each consecutive numbers

# Let's create an array of every even number from ten (inclusive) to fifty (exclus
f = np.arange(10, 50, 2)
f
```

In [ ]:
```
# if we want to generate a sequence of floats, we can use the linspace() function.
# argument isn't the difference between two numbers, but the total number of items
np.linspace( 0, 2, 15 ) # 15 numbers from 0 (inclusive) to 2 (inclusive)
```

Array Operations

In [ ]:
```
# We can do many things on arrays, such as mathematical manipulation (addition, sub
# exponents) as well as use boolean arrays, which are binary values. We can also do
# as product, transpose, inverse, and so forth.
```

In [ ]:
```
# Arithmetic operators on array apply elementwise.

# Let's create a couple of arrays
a = np.array([10,20,30,40])
b = np.array([1, 2, 3,4])

# Now let's look at a minus b
c = a-b
print(c)

# And let's look at a times b
d = a*b
print(d)
```

In [ ]:
```
# With arithmetic manipulation, we can convert current data to the way we want it t
# problem I face - I moved down to the United States about 6 years ago from Canada
# for temperatures, and my wife still hasn't converted to the US system which uses
# could easily convert a number of farenheit values, say the weather forecase, to c

# Let's create an array of typical Ann Arbor winter farenheit values
farenheit = np.array([0,-10,-5,-15,0])

# And the formula for conversion is ((°F - 32) × 5/9 = °C)
celcius = (farenheit - 31) * (5/9)
celcius
```

In [ ]:
```
# Great, so now she knows it's a little chilly outside but not so bad.
```

In [ ]:
```
# Another useful and important manipulation is the boolean array. We can apply an c
# boolean array will be returned for any element in the original, with True being e
# For instance, if we want to get a boolean array to check celcius degrees that are
celcius > -20
```

In [ ]:
```
# Here's another example, we could use the modulus operator to check numbers in an
celcius%2 == 0
```

In [ ]:
```
# Besides elementwise manipulation, it is important to know that numpy supports mat
# look at matrix product. if we want to do elementwise product, we use the "*" sign
A = np.array([[1,1],[0,1]])
B = np.array([[2,0],[3,4]])
print(A*B)

# if we want to do matrix product, we use the "@" sign or use the dot function
print(A@B)
```

In [ ]:
```
# You don't have to worry about complex matrix operations for this course, but it's
# numpy is the underpinning of scientific computing libraries in python, and that t
# element-wise operations (the asterix) as well as matrix-level operations (the @ s
# in a subsequent course.
```

In [ ]:
```
# A few more linear algebra concepts are worth layering in here. You might recall t
# matrices is only plausible when the inner dimensions of the two matrices are the
# to the number of elements both horizontally and vertically in the rendered matric
# can use numpy to quickly see the shape of a matrix:
A.shape
```

In [ ]:
```
# When manipulating arrays of different types, the type of the resulting array will
# the more general of the two types. This is called upcasting.

# Let's create an array of integers
array1 = np.array([[1, 2, 3], [4, 5, 6]])
print(array1.dtype)

# Now let's create an array of floats
array2 = np.array([[7.1, 8.2, 9.1], [10.4, 11.2, 12.3]])
print(array2.dtype)
```

In [ ]:
```
# Integers (int) are whole numbers only, and Floating point numbers (float) can hav
# and a decimal portion. The 64 in this example refers to the number of bits that t
# reserving to represent the number, which determines the size (or precision) of th
# represented.
```

In [ ]:
```
# Let's do an addition for the two arrays
array3=array1+array2
print(array3)
print(array3.dtype)
```

In [ ]:
```
# Notice how the items in the resulting array have been upcast into floating point
```

In [ ]:
```
# Numpy arrays have many interesting aggregation functions on them, such as  sum()
print(array3.sum())
print(array3.max())
print(array3.min())
print(array3.mean())
```

In [ ]:
```
# For two dimensional arrays, we can do the same thing for each row or column
# let's create an array with 15 elements, ranging from 1 to 15,
```

```
# with a dimension of 3X5
b = np.arange(1,16,1).reshape(3,5)
print(b)
```

In [ ]:
```
# Now, we often think about two dimensional arrays being made up of rows and columns
# of these arrays as just a giant ordered list of numbers, and the *shape* of the a
# and columns, is just an abstraction that we have for a particular purpose. Actual
# basic images are stored in computer environments.

# Let's take a look at an example and see how numpy comes into play.
```

In [1]:
```
# For this demonstration I'll use the python imaging library (PIL) and a function t
# Jupyter notebook
from PIL import Image
from IPython.display import display

# And let's just look at the image I'm talking about
im = Image.open('../chris.tiff')
display(im)
```

In [ ]:
```
# Now, we can conver this PIL image to a numpy array
array=np.array(im)
print(array.shape)
array
```

In [ ]:
```
# Here we see that we have a 200x200 array and that the values are all uint8. The u
# unsigned integers (so no negative numbers) and the 8 means 8 bits per byte. This
# be up to 2*2*2*2*2*2*2*2=256 in size (well, actually 255, because we start at zer
# images black is stored as 0 and white is stored as 255. So if we just wanted to t
# use the numpy array to do so

# Let's create an array the same shape
mask=np.full(array.shape,255)
mask
```

In [ ]:
```
# Now let's subtract that from the modified array
modified_array=array-mask

# And lets convert all of the negative values to positive values
modified_array=modified_array*-1

# And as a last step, let's tell numpy to set the value of the datatype correctly
modified_array=modified_array.astype(np.uint8)
modified_array
```

In [ ]:
```
# And lastly, lets display this new array. We do this by using the fromarray() func
# imaging library to convert the numpy array into an object jupyter can render
display(Image.fromarray(modified_array))
```

In [ ]:
```
# Cool. Ok, remember how I started this by talking about how we could just think o
# of bytes, and that the shape was an abstraction? Well, we could just decide to r
# try and render it. PIL is interpreting the individual rows as lines, so we can c
# and columns if we want to. What do you think that would look like?
reshaped=np.reshape(modified_array,(100,400))
print(reshaped.shape)
display(Image.fromarray(reshaped))
```

In [ ]:
```
# Can't say I find that particularly flattering. By reshaping the array to be only
# columns we've essentially doubled the image by taking every other line and stack
# makes the image look more stretched out too.

# This isn't an image manipulation course, but the point was to show you that these
# just abstractions on top of data, and that data has an underlying format (in this
# we can build abstractions on top of that, such as computer code which renders a b
# white, which has meaning to people. In some ways, this whole degree is about data
# we can build on top of that data, from individual byte representations through to
# functions or interactive visualizations. Your role as a data scientist is to unde
# (it's context an collection), and transform it into a different representation to
```

In [ ]:
```
# Ok, back to the mechanics of numpy.
```

Indexing, Slicing and Iterating

Indexing, slicing and iterating are extremely important for data manipulation and analysis because these techinques allow us to select data based on conditions, and copy or update data.

Indexing

In [ ]:
```
# First we are going to look at integer indexing. A one-dimensional array, works in
# To get an element in a one-dimensional array, we simply use the offset index.
a = np.array([1,3,5,7])
a[2]
```

In [ ]:
```
# For multidimensional array, we need to use integer array indexing, let's create a
a = np.array([[1,2], [3, 4], [5, 6]])
a
```

In [ ]:
```
# if we want to select one certain element, we can do so by entering the index, wh
# integers the first being the row, and the second the column
a[1,1] # remember in python we start at 0!
```

In [ ]:
```
# if we want to get multiple elements
# for example, 1, 4, and 6 and put them into a one-dimensional array
# we can enter the indices directly into an array function
np.array([a[0, 0], a[1, 1], a[2, 1]])
```

In [ ]:
```
# we can also do that by using another form of array indexing, which essentiall "z
# second list up
print(a[[0, 1, 2], [0, 1, 1]])
```

Boolean Indexing

In [ ]:
```
# Boolean indexing allows us to select arbitrary elements based on conditions. For
# just talked about we want to find elements that are greater than 5 so we set up
print(a >5)
# This returns a boolean array showing that if the value at the corresponding index
```

In [ ]:
```
# We can then place this array of booleans like a mask over the original array to
# array relating to the true values.
print(a[a>5])
```

In [ ]:
```
# As we will see, this functionality is essential in the pandas toolkit which is th
```

Slicing

In [ ]:
```
# Slicing is a way to create a sub-array based on the original array. For one-dimen
# works in similar ways to a list. To slice, we use the : sign. For instance, if we
# brackets, we get elements from index 0 to index 3 (excluding index 3)
a = np.array([0,1,2,3,4,5])
print(a[:3])
```

In [ ]:
```
# By putting 2:4 in the bracket, we get elements from index 2 to index 4 (excluding
print(a[2:4])
```

In [ ]:
```
# For multi-dimensional arrays, it works similarly, lets see an example
a = np.array([[1,2,3,4], [5,6,7,8], [9,10,11,12]])
a
```

In [ ]:
```
# First, if we put one argument in the array, for example a[:2] then we would get
# first (0th) and second row (1th)
a[:2]
```

In [ ]:
```
# If we add another argument to the array, for example a[:2, 1:3], we get the first
# second and third column values only
a[:2, 1:3]
```

In [ ]:
```
# So, in multidimensional arrays, the first argument is for selecting rows, and the
# selecting columns
```

In [ ]:
```
# It is important to realize that a slice of an array is a view into the same data
# reference. So modifying the sub array will consequently modify the original array

# Here I'll change the element at position [0, 0], which is 2, to 50, then we can s
# original array is changed to 50 as well

sub_array = a[:2, 1:3]
print("sub array index [0,0] value before change:", sub_array[0,0])
sub_array[0,0] = 50
print("sub array index [0,0] value after change:", sub_array[0,0])
print("original array index [0,1] value after change:", a[0,1])
```

Trying Numpy with Datasets

In [ ]:
```
# Now that we have learned the essentials of Numpy let's use it on a couple of data
```

In [ ]:
```
# Here we have a very popular dataset on wine quality, and we are going to only loo
# fields include: fixed acidity, volatile aciditycitric acid, residual sugar, chlor
```

```
# total sulfur dioxidedensity, pH, sulphates, alcohol, quality
```

In [ ]:
```
# To load a dataset in Numpy, we can use the genfromtxt() function. We can specify
# (which is optional but often used), and number of rows to skip if we have a heade

# The genfromtxt() function has a parameter called dtype for specifying data types
# parameter is optional. Without specifying the types, all types will be casted the
# general/precise type

wines = np.genfromtxt("datasets/winequality-red.csv", delimiter=";", skip_header=1
wines
```

In [ ]:
```
# Recall that we can use integer indexing to get a certain column or a row. For exa
# the fixed acidity column, which is the first coluumn, we can do so by entering th
# Also remember that for multidimensional arrays, the first argument refers to the
# argument refers to the column, and if we just give one argument then we'll get a
# back.

# So all rows combined but only the first column from them would be
print("one integer 0 for slicing: ", wines[:, 0])
# But if we wanted the same values but wanted to preserve that they sit in their ow
print("0 to 1 for slicing: \n", wines[:, 0:1])
```

In [ ]:
```
# This is another great example of how the shape of the data is an abstraction whic
# intentionally on top of the data we are working with.
```

In [ ]:
```
# If we want a range of columns in order, say columns 0 through 3 (recall, this mea
# third, since we start at zero and don't include the training index value), we can
wines[:, 0:3]
```

In [ ]:
```
# What if we want several non-consecutive columns? We can place the indices of the
# an array and pass the array as the second argument. Here's an example
wines[:, [0,2,4]]
```

In [ ]:
```
# We can also do some basic summarization of this dataset. For example, if we want
# quality of red wine, we can select the quality column. We could do this in a coup
# appropriate is to use the -1 value for the index, as negative numbers mean slicin
# list. We can then call the aggregation functions on this data.
wines[:,-1].mean()
```

In [ ]:
```
# Let's take a look at another dataset, this time on graduate school admissions. It
# score, TOEFL score, university rating, GPA, having research experience or not, an
# With this dataset, we can do data manipulation and basic analysis to infer what c
# with higher chance of admission. Let's take a look.
```

In [ ]:
```
# We can specify data field names when using genfromtxt() to loads CSV data. Also,
# infer the type of a column by setting the dtype parameter to None
graduate_admission = np.genfromtxt('datasets/Admission_Predict.csv', dtype=None, de
                                   names=('Serial No','GRE Score', 'TOEFL Score',
                                          'LOR','CGPA','Research', 'Chance of Admit
graduate_admission
```

In [ ]:
```
# Notice that the resulting array is actually a one-dimensional array with 400 tupl
graduate_admission.shape
```

In [ ]:
```
# We can retrieve a column from the array using the column's name for example, let
# only the first five values.
graduate_admission['CGPA'][0:5]
```

In [ ]:
```
# Since the GPA in the dataset range from 1 to 10, and in the US it's more common t
# a common task might be to convert the GPA by dividing by 10 and then multiplying
graduate_admission['CGPA'] = graduate_admission['CGPA'] /10 *4
graduate_admission['CGPA'][0:20] #let's get 20 values
```

In [ ]:
```
# Recall boolean masking. We can use this to find out how many students have had re
# creating a boolean mask and passing it to the array indexing operator
len(graduate_admission[graduate_admission['Research'] == 1])
```

In [ ]:
```
# Since we have the data field chance of admission, which ranges from 0 to 1, we co
# with high chance of admission (>0.8) on average have higher GRE score than those
# admission (<0.4)

# So first we use boolean masking to pull out only those students we are interested
# of admission, then we pull out only their GPA scores, then we print the mean valu
print(graduate_admission[graduate_admission['Chance_of_Admit'] > 0.8]['GRE_Score']
print(graduate_admission[graduate_admission['Chance_of_Admit'] < 0.4]['GRE_Score']
```

In [ ]:
```
# Take a moment to reflect here, do you understand what is happening in these calls

# When we do the boolean masking we are left with an array with tuples in it still,
# this a list of the columns we specified and their name and indexes
graduate_admission[graduate_admission['Chance_of_Admit'] > 0.8]
```

In [ ]:
```
# Let's also do this with GPA
print(graduate_admission[graduate_admission['Chance_of_Admit'] > 0.8]['CGPA'].mean
print(graduate_admission[graduate_admission['Chance_of_Admit'] < 0.4]['CGPA'].mean
```

In [ ]:
```
# Hrm, well, I guess one could have expected this. The GPA and GRE for students who
# being admitted, at least based on our cursory look here, seems to be higher.
```

So that's a bit of a whirlwing tour of numpy, the core scientific computing library in python. Now, you're going to see a lot more of this kind of discussion, as the library we'll be focusing on in this course is pandas, which is built on top of numpy. Don't worry if it didn't all sink in the first time, we're going to dig in to most of these topics again with pandas. However, it's useful to know that many of the functions and capabilities of numpy are available to you within pandas.