Basic Machine Learning Tools:

Python, NumPy, Pandas & Jupyter Notebooks

Christian Seberino, Ph.D.

Outline

- Python
- NumPy
- Pandas
- Jupyter Notebooks

Python

- Programming is hard, expensive and slow.

Python

- How would you design a language that was as easy and productive as possible if you didn't care about execution speed and resource requirements?

Python

- **compactness**

- **C++ and Perl have multiple ways of doing things.**

Python

- **few great high level data structures**

- **C only has primitive data structures.**

Python

- **Lists are collections with elements accessible by positions.**

  a = [2, “hello”, True, [5, “world”]]

  Can easily insert, remove, nest, etc.

Python

- **Dictionaries are collections with elements accessible by names.**

  a = {“TX” : “Austin”, “CA” : “Sacramento”}

  Can easily insert, remove, nest, etc.

Python

- **Python also includes tuples, sets, classes and not much more.**

Python

- **Avoid (explicit) compiling.**

- **C, C++ and Java require compilation.**

Python

- **Avoid memory management.**

- **C, C++ and Rust require memory management.**

Python

- **Avoid crashes.**

- **C programs crash rather than raising exceptions.**

Python

- **Avoid porting.**

- **Python has a virtual machine.**

Python

- **Avoid declaring types.**

- **Python automatically handles types!**

Python

- **Leads to more flexible and compact code:**

```
def add(x, y):
    return x + y
```

Python

- **Detects errors at run time instead of compile time.**
- **Can always add unit tests if desired.**
- **Large groups have successfully developed massive enterprise applications in Python.**

Python

- **Avoid confusing code.**
- **Python requires clear indentation!**
- **Python blocks avoids the semicolons and braces!**

Python

- **if a + b >= 4:**

  a *= 2

  b = 4

  print(“Python is like executable pseudocode!”)

Python

- **Avoid duplication of effort.**
- **Python has huge powerful libraries included.**
- **Python can use foreign language extensions.**

Python

- **BUT PYTHON IS TOO SLOW!**

Python

- Are you sure?
- Are you interacting with humans or networks?
- Will your prototype be discarded quickly?
- Can you just use a faster CPU or more RAM?

Python

- **Profiling often reveals tiny bottlenecks can easily be replaced with foreign language extensions.**

Python

- **Even faster and better to work out design, get feedback and win funding before porting to a faster language.**

- **Many things can happen before forced to port!**

NumPy

- **homogeneous array library**

  All elements must of the same type.

  vectors, matrices, tensors, etc.

  (Some support for heterogeneous arrays too.)

NumPy

- **NumPy is fast.**

  Written in C.

  Merger of work of Numeric and Numarray projects.

  Transparently utilizing fast foreign language code!

NumPy

- ndarrays

```
import numpy
numpy.array([1.1, 2.2, 3.3])
numpy.array([[1, 2], [3, 4]])
numpy.array(["quick", "brown", "fox"])
```

NumPy

- **comprehensive set of methods**

  resize, flatten, round, max, min, mean, etc.

  matrix multiplications, inverses, transposes, etc.

NumPy

- extended indexing and slicing options

```
A[1, :]
B[:, 3:6]
```

NumPy

- To leverage NumPy describe your code in ndarrays rather than numbers.
- Avoid for loops when possible!
- No high level description can be catastrophic!

NumPy

- **Is a de facto standard API!**

  People have written compatible libraries supporting massive datasets, multiple CPUs, GPUs and TPUs that are compatible with NumPy!

Pandas

- **heterogeneous array library**

  data analytics

  Built on top of NumPy!

Pandas

- **Series and DataFrames**

```
s = pandas.Series([11, 22, 33])

df = pandas.DataFrame([["35", "a"], [91, "d"]])
```

Pandas

- **comprehensive set of methods**

  Lots of string and mathematical operations.

  Lots of data analytics support.

Jupyter Notebooks

- **browser based interactive language shells**

- **Developed by Dr. Fernando Perez, creator of IPython.**

Jupyter Notebooks

- **Great for running small fast experiments.**

- **Useful for presenting results, graphics, videos, etc. alongside code!**

Jupyter Notebooks

- **kernels**

  language specific background processes

Jupyter Notebooks

- Not as important as Python, NumPy and Pandas.
- Issues with state.
- Restarting the kernel is helpful like restarting a laptop.