THE OHIO STATE UNIVERSITY
Department of Physics

Big Data Analytics in Physics

FALL, 2019

R. HUGHES

Goals of the course

Primary goal: prepare students to run their own machine learning projects.

By the end of this class, students should be able to:

- Examine, clean, and graphically explore datasets
- Gain some experience with common machine learning algorithms
- Design a machine learning approach to solve big data problems

Administrative

Grading

- 70%: A mixture of in-class and out-of-class jupyter notebooks
  - I expect that every class will have an in-class component that will be due (at least partially) at the end of class
  - You are expected to attend every class
- 30%: Final project

Class format

- Meet Tuesdays and Thursdays in Smith 1094 from 12:40-2:45pm
- Start with (very) Short lectures based on the Jupyter Notebook for that day
- Most of our class time: Hands on computer work guided by existing jupyter notebooks, followed by notebooks that you are expected to complete.

Course Expectations

My expectations of the Students

- I expect you to come to every class with a hunger to learn the material.
- I expect you to be self-motivated to solve problems that you come across. You will come across many problems!
- I expect you to do your best to finish the assignments.
- I expect you to work hard on the final project.
- I expect that you will have a lot of fun learning this material!

My expectations of myself:

- I expect to work hard to provide an intellectually stimulating atmosphere.
- I expect to do my best to structure the course so that you all learn as much as you are capable of throughout the semester.
- I expect to have a lot of fun teaching (and learning) this material!

What this course is NOT

1. A programming course! We will use python throughout the course, and if you don't already know how to program in python you should be reasonably competent in using python by the end of this course. But I am not teaching python! This course uses programming as a tool to solve problems.

2. A theory course in machine learning! We might touch on some underlying theory concepts, but I am not an expert in the theory and I am not competent to teach it. This is a course in the practical application of machine learning to a variety of different problems, and in this I have a fair amount of practical experience. However, to learn the material **you** will have to do the work.

3. I expect you to figure out the vast majority bugs in your code, and I encourage you to get help from friends and the internet (especially stackoverflow!!!) .

My Background

1. Education: Undergrad degree in Elec Eng; PhD Physics .
2. Part of the team that discovered the Top Quark in 1995.
3. Co-led group that designed and built trigger electronics for colliding beam experiment
4. Wrote on-board science software to find gamma ray bursts for Fermi Gamma Ray Space Telescope
5. Starting using artificial neural networks in 1997.
6. Have programmed in pascal, fortran, assembler, java, C++, AHDL, Matlab, Torch7/Lua, Python (among others…)
7. Have co-founded 2 startups in data science. One successful and active, the other not dead (yet).

Textbooks

1. Hands-On Machine Learning With Scikit-Learn and TensorFlow, by Aurelien Geron.

2. Deep Learning with Python, by Francois Chollet

Textbooks are not required but would be useful.

What about the "…In Physics" Part?

We will use a variety of datasets in order to master as much as possible a variety of machine learning techniques.

Some of these datasets may be physics oriented but many (and probably most) won't be. There is a fair amount of effort required to get even simple examples to work, and we want to focus on mastery of these methods. It should be obvious as we go through these examples that the application of these methods to physics research problems is in many cases straightforward.

Course Flow… not a schedule:

- Week 1: Introduction to tools (jupyter/colab/git). Chapter 1 of Hands on Machine Learning By Geron (HML)
- Weeks 2-n: Some fraction of chapters 2-7 of HML.
- Weeks n-14: Some fraction of chapters 1-6 of Deep Learning with Python by Chollet
- Around week 9: Final projects proposed and/or assigned. Some fraction of remaining class time will be assigned to project work.
- Note: After week 9 we will still be doing non-project work in class.

Tools

1. Python: the programming language we will use
2. Git: A version control system for our code.
3. Jupyter Notebooks: A browser based environment for programming in a variety of languages including of course python. Can be run locally on your own machine or as a online hosted service.
4. stackoverflow: THE premier place on the web to find ANY answer to ANY problem. If you don't find the answer here, you are not asking the right question!

Topics I expect/hope to cover this semester:

1. Data preparation and visualization
2. Scikit-learn:
   a) regression
   b) classification: binary and multi-class
   c) details of linear and logistic regression
   d) decision trees
3. Keras:
   a) binary classification with images
   b) multi-class classification with text data
   c) autoencoders
   d) convolutional neural networks
   e) siamese networks
   f) recursive neural networks

OSC
- Program memory and storage; CPU/GPU
- Account which is backed up
- Provides remote machine running jupyter notebook
- FREE!

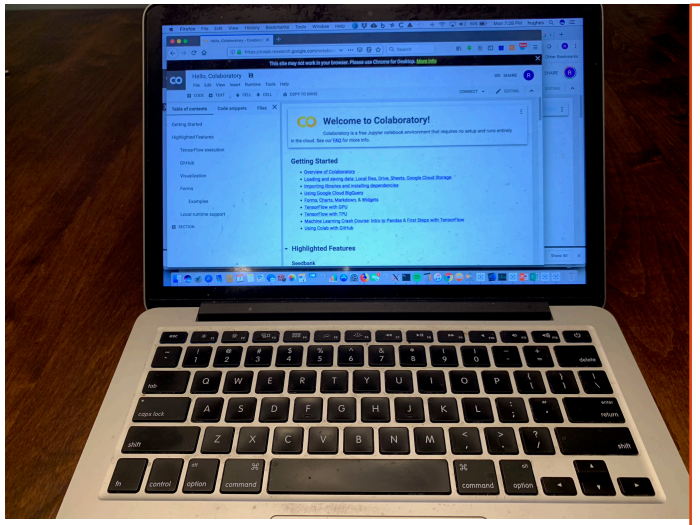
Jupyter commands and data

Workflow

Pull new assignments

Push finished assignments

Github
- Remote code repo
- Where your code must end up in order to be graded!
- FREE!



Your laptop
- Where you run a browser which renders the jupyter notebook (the notebook is actually running on a machine at OSC)
- Also provides browser-based access to your account at OSC

OSC
- Program memory and storage; CPU/GPU
- Account which is backed up
- Provides remote machine running jupyter notebook
- FREE!

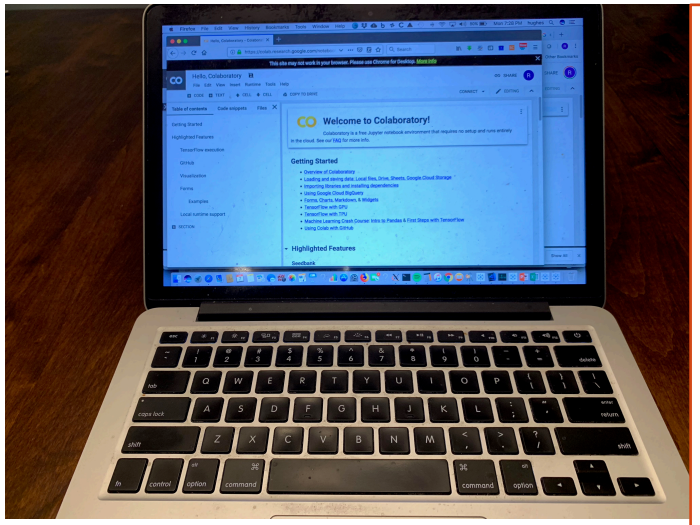
Pull new assignments

Push finished assignments

Github
- Remote code repo
- Where your code must end up in order to be graded!
- FREE!

Workflow

Jupyter commands and data



Your laptop
- Where you run a browser which renders the jupyter notebook (the notebook is actually running on a machine at OSC)
- Also provides browser-based access to your account at OSC

Goals for today

1. Make sure everyone (with a laptop):
   - git accounts
   - Local repo download
   - Ability to make minor changes to local repo and push those changes back to the remote (github.com)
2. Startup ondemand OSCenvironment and begin to explore how it works
3. Start on python_intro.ipynb workbook
   - This hits all of the high points of python, and even does a bit with numpy package (which allows for multidimensional array creation and fast array operations). It does nothing with pandas (we will do that separately).
4. Reach goal: start on assignment1.ipynb
   - This is a exercise in python programming which will either be fun or frustrating (or a bit of both)!