# PHYS3600ID – Machine Learning for Physics and Astronomy

# D862 - Fall 2018

## Instructors

Dr. Viviana Acquaviva (Physics)
vacquaviva@citytech.cuny.edu
Office hours: Thursday 3.30-4.30pm, Namm 828

Dr. Ashwin Satyanarayana (CST)
asatyanarayana@citytech.cuny.edu
Office hours: Monday and Wednesday, 2:30-3:30pm, Namm 913

## Course philosophy

This course is very hands-on and we assume that you are here because you have a genuine interest in machine learning methods and in understanding how they are applied to real-world research problems. We will guide you and we are happy to help but we also expect you to be independent thinkers and to take initiative – make a Stack Overflow account and don't be afraid to ask questions in class, to your fellow students and online!

## Class/assignment rules

We encourage you to talk to each other in class and beyond but your assignments need to be the result of your own work. Identical or very similar assignments are not acceptable. This is valid also for longer coding assignments and reports. **Using online sources as inspiration for coding assignments is allowed but sources should be cited. Using large chunks of text from outside sources in reports is not allowed and will be considered plagiarism.**

We expect students to attend class; you can't make up worksheets for which you weren't in class, but we'll drop the lowest two grades. If you miss several classes, your grade will suffer. If you have extenuating circumstances, please talk to us and we'll figure out something.

## Class material

We will distribute course material (lecture slides, notebooks, worksheets) on an ongoing basis using Blackboard. You will also be asked to upload homework (worksheets) and reports to Blackboard.

**Resources**

This course uses Open Education Resources. Rather than a proper textbook, since the material is new and ever-changing, we have a collection of useful resources.

**On programming:**

These are good introductory lectures and we will use some of this material in the first few weeks:

https://github.com/jrjohansson/scientific-python-lectures

https://github.com/jakevdp/WhirlwindTourOfPython/

https://developers.google.com/edu/python/

**On Data Science:**

https://github.com/jakevdp/PythonDataScienceHandbook

(all content is available free of charge in the form of Python notebooks, and the book can be purchased on Amazon).

**On practical implementation of Machine Learning algorithms in Python:**

http://scikit-learn.org/stable/user_guide.html

**Software**

We used Jupyter notebooks, which require Python, and we have them installed in the computer lab, but if you would like to install them on your personal machine, you can find easy-to-follow instructions here:

http://jupyter.readthedocs.io/en/latest/install.html

(refer to "Instally Jupyter with Anaconda" section)

**Grading Policy**

The course will be graded according to: weekly worksheets (30%), three special homework sets (reports) (25%), two in-class quizzes (20%), and final project (25%).

**Class Schedule**

| Week | Date | Topic | Tools introduced | HW due<br>**Worksheets due weekly starting from week 2**<br>**(due February 13)** |
|---|---|---|---|---|
| 1<br>PHYS | Jan 30 (T)<br>Feb 1 (Th) | Introduction; course overview; first steps in Python | Basic Python commands and tutorials; Jupyter notebooks | |
| 2<br>PHYS | Feb 6 (T)<br>Feb 8 (Th) | Basic statistics and programming recap | Numpy, Scipy | |
| 3<br>PHYS | Feb 13 (T)<br>Feb 15 (Th) | Fitting data and building models I | | |
| 4<br>PHYS | Feb 22 (Th)<br>Feb 27 (T) | Fitting data and building models II<br><br>Types of charts, graphs and tables. Composite charts. Visualization of multidimensional data. | Package: Matplotlib. | |
| 5<br>CST | Mar 1 (Th)<br>Mar 6 (T) | Introduction to machine learning terminology and concepts; linear regression, logistic regression | | HW due: Special report essay on the accelerated expansion of the Universe |
| 6<br>CST | Mar 8 (Th)<br>Mar 13 (T) | Evaluating performance; Diagnostics; Troubleshooting; Learning curves; Accuracy paradox | | |

| | | | | |
|---|---|---|---|---|
| 7<br>PHYS | Mar 15 (Th)<br>Mar 20 (T) | Classification problems and algorithms I( Naïve Bayes; Gradient Boosting; SVMs)<br><br>Train/Test/Cross-validation. | Packages: Pandas, Scikit-learn | |
| 8<br>PHYS | Mar 22 (Th)<br>Mar 27 (T) | Classification problems and algorithms II (Decision Trees; Random Forests) Hyperparameter fitting | | |
| QUIZ | Mar 29 (Th) | | | |
| 10<br>CST | Apr 10 (T)<br>Apr 12 (Th) | Unsupervised learning algorithms; Clustering; Dimensionality reduction (PCA) | | HW due: Worksheet/report on particle physics data challenge |
| 11<br>CST/PHYS | Apr 17 (T)<br>Apr 19 (Th) | Unsupervised learning algorithms II; Feature selection | | |
| 12<br>PHYS | Apr 24 (T)<br>Apr 26 (Th) | Data cleaning/ real world data: how to deal with formatting, outliers, variable selection | | Final project topics due April 26 |
| 13<br>PHYS | May 1 (T)<br>May 3 (Th) | Algorithm comparison and ensembling uncertainty estimation | | HW due: Worksheet/report on galaxies and unsupervised learning |
| 14<br>PHYS | May 8 (T)<br>May 10 (Th) | Chosen by students | | Final projects outlines due May 8 |
| QUIZ | May 15 (T) | Review/Exam | | |
| 15<br>PHYS/CST | May 17 (Th)<br>May 22 (T) | Final project presentations | | |