# PHYS 3359: Data Analysis for the Natural Sciences II: Machine Learning

**Spring 2023 – TR 1:45 - 3:15pm (Lectures),**

**T 4-5pm, W 1-2pm (Office hours)**

### Instructor

Prof. Bhuvnesh Jain and Dr. Mike Jarvis
Office : DRL 4N12A
Email: bjain@physics.upenn.edu, mjarvis@physics.upenn.edu

### Teaching Assistants

TBD

### Course Description

This is a course on data analysis and statistical inference for the natural sciences focused on machine learning techniques. The main topics are: classification, training/validation samples, cross-validation, supervised vs. unsupervised learning, regularization and resampling methods, tree-based methods, support vector machines, neural networks, deep learning and image analysis with convolutional neural networks. Students will obtain both the theoretical background in data analysis and get hands-on experience analyzing real scientific data. This course forms a two-course sequence with Phys 3358.

Class time will be used partially for lectures and partially for discussions and exercises.

There will be biweekly assignments, averaging to about 5 hours/week of work. In addition students will have the opportunity for group work on zoom for debugging or other exercises.

### Prerequisites and Requirements

Prerequisite: Calculus and linear algebra. Prior programming experience in python. Phys 3358 or equivalent.

### Primary Textbook

- *Introduction to Statistical Learning (ISLR)*, Gareth James, Daniela Witten, Trevor Hastie, & Robert Tibshirani (Springer, 2013)
  - This is a now standard textbook on introduction to machine learning for advanced undergraduates. The book assumes a working knowledge of statistics. There are worked out examples in R (even though we will be using Python). The book is free from the author's website (http://www-bcf.usc.edu/~gareth/ISL/)! We will supplement this textbook with several other statistics and data analysis books listed below.

- *The Elements of Statistical Learning*, Trevor Hastie, Robert Robert Tibshirani & Jerome Friedman (Second Edition, Springer, 2009)
  - This book is more mathematical than ISLR. It has a more detailed treatment of several ML topics, including a chapter on neural networks. We will use it for some mathematical derivations.
- *Neural Networks and Deep Learning*, Michael Nielsen (http://neuralnetworksanddeeplearning.com/)
  - Online book, excellent pedagogical introduction to neural networks and deep learning. Python code available on author's GitHub site. Download book as pdf here: https://static.latexstudio.net/article/2018/0912/neuralnetworksanddeeplearning.pdf

**Other Useful Textbooks**

Python for the Sciences

- *A Student's Guide to Python for Physical Modeling*, Updated Edition, Jesse M. Kinder & Philip Nelson (Princeton, 2018)
  - This is co-authored by our colleague Phil Nelson. It is an excellent place to start for beginners and those who already know another programming language. It has a good introduction to many of the essential tools for the physical sciences (data I/O, plotting, numerical methods, and even image processing) all in one place.
- *Python Data Science Handbook*, Jake VanderPlas (O'Reilly )
  - A nice introduction to data analysis with python. The ML parts use the Scikit-Learn package.
  - The book is available for free: https://jakevdp.github.io/PythonDataScienceHandbook/index.html

Statistics and Data Analysis

- *Data Reduction and Error Analysis for the Physical Sciences*, 3rd Ed., Philip Bevington & D. Keith Robinson (McGraw-Hill, 2002)
  - A great introductory book on basic data analysis with special emphasis on linear regression and curve fitting. Also available as pdf, e.g. here: *hosting.astro.cornell.edu/academics/courses/astro3310/Books/Bevington_opt.pdf*

**Grading**

- Homework - 70%
  - There are ~8 homework assignments consist primarily of coding and data analysis exercises. Each assignment will take approximately 4 - 6 hours to complete.
- Final Project - 30%

- The final coding project will involve an analysis of a large dataset (e.g., stock market, DNA, sports games, etc). Students will make predictions using machine learning techniques covered throughout the semester.

**Detailed Syllabus (we will devote 1 week to most topics listed below)**

- Fundamentals of Machine Learning
- Supervised classification methods — K nearest-neighbors, logistic regression, discriminant analysis
- Resampling Methods — cross validation, bootstrap
- Model selection and regularization — subset selection, ridge regression, lasso, PCA
- Bayes' theorem, parameter estimation, regression — a review
- Non-linear models — polynomial regression, cubic splines, additive models
- Tree-based Methods — decision trees, random forests, boosting and bagging
- Support Vector Machines — maximal margin and support vector classifiers
- Neural networks — basic neuron architecture, sigmoid activation, back propagation
- Deep learning and image analysis with Convolutional Neural Networks
- Advanced topics in deep learning