AT 780A7: Machine Learning for the Atmospheric Sciences Course Syllabus for Fall 2023

Class: 09:00AM - 09:50AM MW
Course Webpage: `https://colostate.instructure.com/courses/172622`

An overview of machine learning methods used to interpret data sets in the atmospheric sciences with an emphasis on implementation and interpretation of the results, including learning new science.

1 Instructor

Prof. Elizabeth A. Barnes

email:	eabarnes@colostate.edu
Office Hours:	W 10:00-11:00 and other times by appointment

2 Teaching Assistant

Charlotte Connolly

email:	cconn@rams.colostate.edu
Office Hours:	TBD

3 Course Focus

An overview of machine learning methods used to interpret data sets in the atmospheric sciences. Lectures will include foundational concepts common across machine learning methods, and then dive deeper into atmospheric science applications of machine learning methods, including neural networks and a variety of unsupervised methods. Focus will be on the application to atmospheric science tasks, with an emphasis on implementation and interpretation of the results, including learning new science.

Upon successful completion of this course students will be able to:

- describe basic machine learning concepts and their use in atmospheric science research
- compare and contrast machine learning approaches to more standard atmospheric science approaches in the literature
- explain how they chose the specific parameters in their machine learning methodologies, and how it impacts their scientific conclusions
- design and implement machine learning algorithms with their own atmospheric/oceanic data
- communicate final results in a clear and professional manner

4 Course Expectations

The following list presents the minimum requirements for passing this course:

- participate in live lectures,
- keep up with the reading,
- submit all assignments on time and at an acceptable level of quality.

5 Course Prerequisites

Students are expected to have a familiarity with the data science methods and tools used in atmospheric science research (prerequisite ATS 655), which will not be covered in this class. This includes basic high-school and college-level mathematical concepts as well as more advanced data science methods, for example:

- algebra (e.g. equations for lines, solving basic algebraic equations)
- basic calculus (e.g. how to take a derivative and an integral)
- basic matrix algebra (e.g. addition, subtraction, multiplication)
- convolution
- principal component analysis

If you are concerned about your background in these areas, please speak with me. While the concepts, tools, and techniques explored in this course will be taught within the context of atmospheric science, there are no atmospheric dynamics prerequisites.

6 Course Web Page

The course web site will be used for posting resources and homework assignments. The course web site is through CSU Canvas and is listed at the top of this syllabus. All students enrolled in the course should have access to the Canvas material.

You will submit your assignments via Canvas by uploading a *.pdf* file of your final report for each assignment by the due date. Canvas will also allow you to keep track of your grade/points in the course.

7 Grading

7.1 Grade Break-down

Your course grade will be made up of homeworks only. That is, homeworks and class participation will together cover 100% of your grade.

7.2 Homework

There will be 3 homeworks/projects throughout this course (although I maintain the right to increase or decrease this number). The homeworks will be based on your own data, and so, will be very open-ended. You do not need to turn-in code for the assignments. I will only be interested in the final write-up which should contain: motivation, experimental design, results and conclusions.

You are *strongly encouraged* to interact with your classmates by sharing ideas and discussing the specifics of the material and homeworks at any point in the course. You are, however, expected to hand-in your own homework assignment, and it should not be a direct copy of your classmate's.

Your homework assignments must be typed-up and clearly written. Figures should be of publication quality - no low-resolution figures. I repeat, *no low-resolution figures.* By doing this, you are not just being nice to the me and the TA, who have to read your work, but you will gain practice in presenting your results clearly and professionally as required for your careers as scientists.

7.3 Midterm & Final Exams

There will not be a midterm or a final exam in this course.

8 Textbooks & Resources

There is one required resource in this course - **the internet**. Google is amazing - use it. In addition, there are thousands (millions?) of blog posts written about machine learning concepts, and you should feel free to use them. In addition, you will find it is useful to refer to the software documentation regularly. One of the most important things to learn in graduate school is “how to look it up.” In my own research, I use most of the techniques we will discuss, but I have very few of them memorized. By the end of this course, you should aim to be self-sufficient in finding the analysis techniques you need. You should not care whether you have a specific function memorized, but whether you know how to find it.

Additional resources are specified on Canvas and may be made available through the CSU Library or the instructor’s personal collection, including the following textbook: *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, A. Geron, 3rd edition.* This is one of the best books out there I have found, so if you want a physical book, I highly recommend it. I will be assigning reading from this book, so if you do not want to purchase it let me know and I will provide you with a copy to borrow for the semester.

9 Software

9.1 Coding Languages

This course will have a substantial coding component. While every student is welcome to use whatever software they desire the instructor and TA will be exclusively using Python 3.x and Tensorflow 2.x. Machine learning has become much more accessible in recent years due to the huge strides made in open source machine learning APIs[1]. For python, the two main choices are Tensorflow and Pytorch. Tensorflow tends to be more user friendly (at least in my experience) and so that is what we will be using throughout the course for in-class discussions and examples. Note that scikit-learn also has some machine learning packages, and we will use this for random forests. However, scikit-learn hides too much under-the-hood for neural networks, and so we will be using Tensorflow instead.

Note that the instructor and TA will *not* be able to assist you with your code if you do not use Tensorflow/Scikit-Learn. You are, however, welcome to use whatever software you wish to perform the assignments.

9.2 Colaboratory

Software package management can be a pain, especially when everyone is using a different operating system. Because of this, you will be in charge of your own packages (e.g. conda environments). To ensure that everyone can still run the python examples from class, we will be using Google’s Colaboratory `https://colab.research.google.com`. Colaboratory, or “Colab”, allows you to write and execute Python in your browser, with zero configuration required, free access to Google’s GPUs and CPUs and allows for easy sharing.

To run the class examples in Colab you will just need a Google account that you can sign-into when you are ready to execute the code.

9.3 *Optional: LaTeX*

LaTeX[2] is a type-set program that takes macro code and formats it into a final (often pdf) document. For example, this syllabus was written with *LaTeX*. The end result is a clean, consistently formatted document. Many scientists use *LaTeX* to write-up their research, and journals are increasingly preferring *LaTeX* files to Mircrosoft Word files for manuscript submission.

A main reason to use *LaTeX* is the ease with which mathematical symbols, equations, etc. are formatted. In addition, including figures is efficient: the user does not “cut and paste” the figure into the text, but rather places the actual document path of the figure in the *LaTeX* code. Thus, whenever the figure is changed, it is automatically updated in the manuscript file. *LaTeX* is free and can be used on all common operating systems

[1] application programming interface

[2] pronounced “LAY-tek” or “LAH-tek”.

(e.g. Linux, Mac, Windows). In addition, `Overleaf.com` is a fantastic site for typesetting and collaborating with *LaTeX* in the could (i.e. does not require you install anything on your computer).

I will not require that you use *LaTeX* for your homeworks, however, I highly encourage you to do so. While the initial learning curve is rather steep, I think that the payoff is worth it. Equation type-setting is easy and always neat, figures will be easily updated, and references are straight-forward to handle with *BibTeX*[3].

10 CSU Honor Pledge

This course will adhere to the CSU Academic Integrity Policy as found in the General Catalog (`http://catalog.colostate.edu/general-catalog/policies/students-responsibilities/#academic-integrity`) and the Student Conduct Code (`http://www.conflictresolution.colostate.edu/conduct-code`). At a minimum, violations will result in a grading penalty in this course and a report to the Office of Conflict Resolution and Student Conduct Services."

[3] *LaTeX*'s bibliography manager.

11 Tentative Outline

The following is a tentative outline for the class. Reality will almost surely deviate from this outline.

Module I **Foundational Concepts**:

- Syllabus, overview of ML in atmospheric science, start of foundational concepts
- Foundational concepts, links to linear regression, overview of pre-processing weather and climate data

Module II **Random Forests**:

- Decision trees: vocabulary, structure, loss functions, pre-processing, parameter choices
- Random forests for weather prediction: classification vs regression, bagging vs boosting, diagnostics during/after training
- Random forests for weather prediction: regularization, interpretability methods

Module III **Neural Networks**:

- Neural networks: feed-forward architecture, backpropagation, gradient descent, software tools, pre-processing
- Neural networks: parameter choices, classification vs regression, overfitting, regularization, diagnostics during/after training
- Neural networks: explainability methods applied to climate predictability studies
- Neural networks: simple uncertainty quantification

Module IV **Advanced Neural Networks**:

- Convolutional Neural Networks: 2D/3D convolution, problem setup, parameter choices
- Convolutional Neural Networks: receptive fields
- Advanced ANN architectures: U-nets, LSTMs, reservoir computing

Module V **Unsupervised Learning**:

- Unsupervised learning: clustering methods, discussion of dimension reduction (e.g. PCA)
- Unsupervised learning: autoencoders, generative adversarial network

Module VI **New Frontiers**:

- Incorporating physics into ML; Transfer learning; future directions of ML for atmospheric science

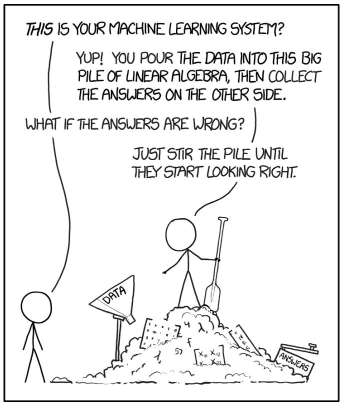
THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

Figure 1: `https://xkcd.com/1838/`