Intro to EN5422/EV4238

Fall 2023
intro.pdf
(Week 1 – ½)

Contents

1 Course Website

Main Course Webpage: https://www.hydroai.net/teaching/2023-en5422-ev4238

2 About us

2.1 About the Instructor

- Faculty Webpage https://www.hydroai.net/people
- GitHub https://github.com/hyunglok-kim
- Google Scholar https://scholar.google.com/citations?user=ZJx_f8gAAAAJ&hl=en&authuser=1

2.2 About you

Fill out a notecard with the following information:

1. Your name (with pronunciation hints)
2. Hometown (include country/region if you think I won't know)
3. Previous and Current Degrees
4. What type of job to hope to receive upon graduation (title & industry)
5. 2 things you hope to learn in this course
6. 2 interesting things about you (to help me remember you)

3 The course

3.1 Topics

- See website: https://www.hydroai.net/teaching/2023-en5422-ev4238
- Course contains aspects of: data analysis, modeling, stats, shallow machine learning (ML), deep learning (DL), coding, algorithms, land surface model (LSM), and probability with Earth sciences and environmental data.

3.2 Examples

- Predict how much temperature will increase in the future (**Figure 1**)
- Estimate the water resources over ungagged areas with LSM data (**Figure 2**)
- Build drought prediction model with satellite data sets (**Figure 3**)
- Predict dust storm/flood events with satellite data sets (**Figure 4**)
- and many more

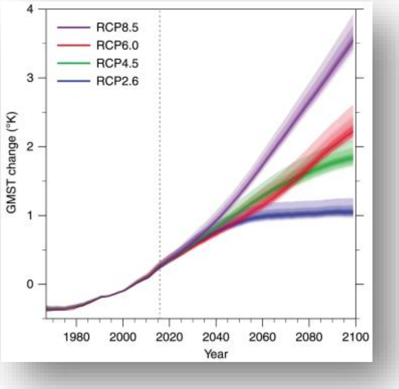

Figure 1 NOAA Global Historical Climatology Network (GHCN) and Extended Reconstructed Sea Surface Temperature (ERSST)

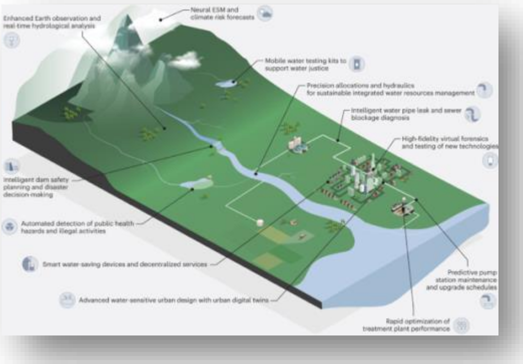
Enhanced Earth observation and real-time hydrological analysis
Neural ESM and climate risk forecasts
Mobile water testing kits to support water justice
Precision allocations and hydraulics for sustainable integrated water resources management
Intelligent water pipe leak and sewer blockage diagnosis
High-fidelity virtual forensics and testing of new technologies
Intelligent dam safety planning and disaster decision-making
Automated detection of public health hazards and illegal activities
Smart water-saving devices and decentralized services
Advanced water-sensitive urban design with urban digital twins
Predictive pump station maintenance and upgrade schedules
Rapid optimization of treatment plant performance

Figure 2 AI has the potential to yield system-wide benefits ranging from enhanced catchment insights to optimized network efficiency to improved service for end-users (Richards et al., 2023)

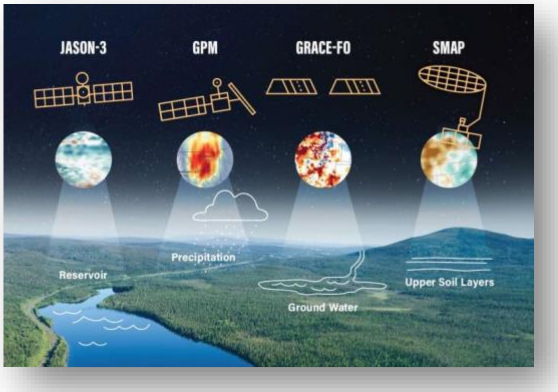

Figure 3 The collaboration of various satellites, including GPM, SMAP, MODIS, Landsat, Jason-3, and GRACE-FO, offers comprehensive global water insights by measuring precipitation, soil moisture, vegetation impact, river height, reservoir content, and groundwater distribution.

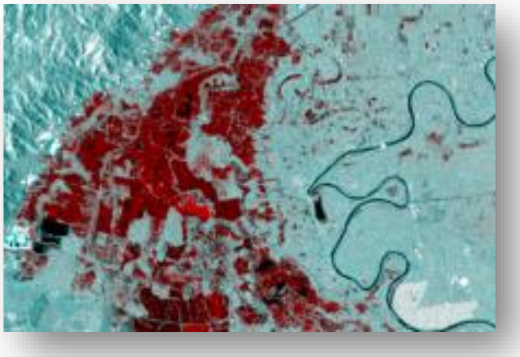

Figure 4 Machine learning approaches provide new possibilities for flood detection as more data becomes available, computing power increases and machine learning algorithms improve (Mosavi et al., 2018)

4 Syllabus

4.1 Course Webpage

- We have a course webpage https://www.hydroai.net/teaching/2023-en5422-ev4238
    - lecture
    - Python scripts
    - data sets
    - homework assignments

- We will use the Google Colab site (https://colab.research.google.com) for homework submission, solutions, etc.
- We will use Slack channel: **#en5422_ev4238_fall_2023** for communication

4.2 Course Prereqs

- Linear Regression
  – Multiple Linear Regression
  – Logistic Regression
  – Categorical Predictors (dummy coding)
  – Implementation in R (lm(), predict(), etc.) – Estimation / Model Fitting
  – Cross-validation
- Probability and Statistics
  – Bayes Theorem
  – CDF/PDF/PMF
  – Maximum Likelihood Estimation
  – Distributions: normal, binomial, hypergeometric, etc. – Expected value, variance, median, quantiles
  – Mean Square Error – Confidence Intervals – Hypothesis Testing
- Math
  – Calculus
  – Matrix Calculations
  – PCA, SVD
- Computing
  – data types: vector, matrix, array, list, etc. – writing simple functions
  – flow control: loops, if/else, etc.
  – data wrangling
  – generating random variables
  – Python Markdown [Note: practice HW will cover Python Markdown]

4.3 Exercise 1

Your Turn #1

Let X_1, X_2, …, X_n be the yearly number of flood events in Gwangju province in Korea (X_i is the number of crashes in year i).

- What is an estimate of the probability that there are 30 flood events in year n+1?

4.4 Exercise 1

- Office Hours
- Textbooks
- Anaconda, Jupyter Lab, Google Colab
- Course Assessment
    - Due dates are posted on the course website
    - Python Markdown (See HW0)
    - No class participation grade, but expect you come prepared with questions. Don't be afraid to ask questions in class. Now is your time to learn.
- Course Management
- Read all of syllabus and ask questions

4.5 Succeeding in this course

- Most topics are separated into two lectures
    - First is introduction of new topics
    - Second is more advanced coverage
- Homework is due weekly
    - Due on Wednesday's morning, but expected to be completed before Tuesday's class.
    - Should start HW after first lecture.

- Assigned Reading before every class
  - First listed reading is intro, second is more advanced
  - Strat with intro, then re-read the advanced
  - Quizzes based on first reading
- Attend office hours!

4.5.1 Python

The free online source Python for Everybody is an undergrad level “Intro to Python” course. It covers most of the Python uses we will employ this semester. This would provide a good overall preparation.

4.5.2 Coding

I find that many students struggle with coding. This really hinders your ability to get your mind about the concepts and slows down your learning. The course will use Python, but with Jupyter lab-like interactive environment for notebook, code, and data (https://jupyter.org). There is no better tool for interactive data analysis and both exploratory and confirmatory modeling. Jupyter lab is a major improvement over base command line Python, but it can look a bit different and take some time becoming familiar with. The free online Python Data Science Handbook and website provide a good introduction and reference. While I encourage Jupyter lab, you are free to use anything for homework.

- Jupyter Lab has tutorials
  - http://justinbois.github.io/bootcamp/2020_fsri/lessons/l01_welcome.html
  - Handy Jupyter Lab cheat sheets

4.5.3 Statistics

I find students understand the least about statistical concepts. This is so fundamental to all of ML and Data Mining; a strong grasp of statistics will enable the connections between topics to pop out. If you already feel comfortable coding, I suggest you go a quick stat review. Here are two introductory resources:

- https://www.openintro.org/book/ims
- https://moderndive.com/index.htm

4.5.4 Math

The students who gain the most from the course will embrace mathematical equations. As they say “an equation is worth a thousand words”. While we won’t do any proofs in this class, we will judiciously use equations to clarify concepts. Spend time to become intimate with math notation – it is worth the investment.

4.5.5 Trustworthy Material

- The assigned readings are trustworthy
- Blogs and videos you find on the web are not
- Please don’t trust: Toward Data Science, Analytics Vidha, Machine Learning Mastery, Medium
    - There is certainly some good content, but how will know to discern good from bad while still learning?