Machine Learning for Earth and Environmental Sciences

Master in Environmental Sciences, FGSE, University of Lausanne

Syllabus (for Fall 2022, last updated on August 28, 2022)

Main Instructor: Tom Beucler (Assistant Professor at IDYST, Lab Website, tom.beucler@unil.ch)
Teaching Assistants: Milton Gomez (PhD Student at IDYST, milton.gomez@unil.ch, Frederick Iat-Hin Tam (PhD Student at IDYST, iathin.tam@unil.ch), Jingyan Yu (Postdoctoral Fellow at IGD, jingyan.yu@surrey.ac.uk).

1 Summary

Our ever-improving ability to observe and model the environment produces Petabytes of data every day, which overwhelm traditional data analysis methods. Machine learning (ML) algorithms, broadly defined as algorithms that can automatically learn to perform a task from data without requiring explicit programming to do so, have recently emerged as efficient tools to extract knowledge from large geoscientific datasets. Once trained, these algorithms are inexpensive to use, making them ideal shortcuts when time or resources prevent running a full-complexity model. In addition to providing computational shortcuts, the ability of ML algorithms to summarize large amounts of data makes them promising tools for environmental scientific discovery.

In this 10-week hands-on course, we will introduce common ML algorithms in the context of their application in Earth and environmental sciences. By the end of this course, you should be able to:

1. Name common ML algorithms (listed in Sec2) and summarize their advantages and limitations, especially in the context of environmental science,
2. Implement them in Python (mostly using the Numpy/Scikit-Learn/Keras/Tensorflow libraries in Google Colab notebooks),
3. Know from experience which algorithms are most appropriate for environmental applications you are passionate about (e.g., your Masters thesis).

To achieve these three objectives, the course will combine:

- **Lectures** (≈2 hours/week, 15% of grade): Typical structure = 15-min answering your questions about readings, 15-min live quiz based on readings (taking quiz is 10% of the grade, correct answers are 5%), 15-min interactive lecture diving into main algorithm & environmental application of the week, 15-min reviewing quiz, 30-min overview lecture on additional algorithms covered in readings. Note that during lecture, we will favor ML applications over mathematical foundations; if you are interested in the latter, we encourage you to take the appropriate ML courses at EPFL.
- **Readings** (≈3 hours/week, 15% of grade): Reading1 is a textbook chapter covering next week's algorithms, while Reading2 is (usually) an extract from a recently published article that successfully applied ML algorithm to tackle key environmental science issues. Both readings are posted on Moodle with guiding questions: Even if your answers aren't correct, you'll get the full 15% of the grade as long as you write thoughtful answers no more than 24 hours before the lecture.
- **Computer labs** (≈3 hours/week, 20% of grade): ≈1.5hr applying ML covered in Reading1 on standard ML datasets + ≈1.5hr applying ML on environmental dataset covered in Reading2. To get full credits (20%), simply push the completed Colab notebook to your fork of the course's GitHub repo no more than 24 hours before the following lecture.
- **Final project** (≈2 hours/week, 50% of grade): The final project's goal is to answer a well-defined scientific question by applying one of the ML algorithms introduced in class on an environmental dataset of your choice (e.g., related to your Masters thesis). We will give more specific instructions during the first lecture and upload them on Moodle. You may collaborate with peers on the same dataset and present your projects together in class (20% of grade), but you must choose distinct scientific questions and write separate 4-page final reports (30% of grade) using this template on Overleaf (LaTeX tutorial at this link). Please submit the final report in PDF format via Moodle before 8PM CET on May 27th.
- There will be **no final exam**inations or homework other than the readings and the final project.

For the ML components of the course, we will mainly use Géron's 2019 textbook "Hands-on Machine Learning with Scikit-Learn, Keras & Tensorflow (2nd ed)" (code/pdf) and Chollet's 2021 textbook "Deep Learning with Python (2nd ed)" (code/pdf), but we encourage you to use the wealth of online resources on machine learning (link to get started).

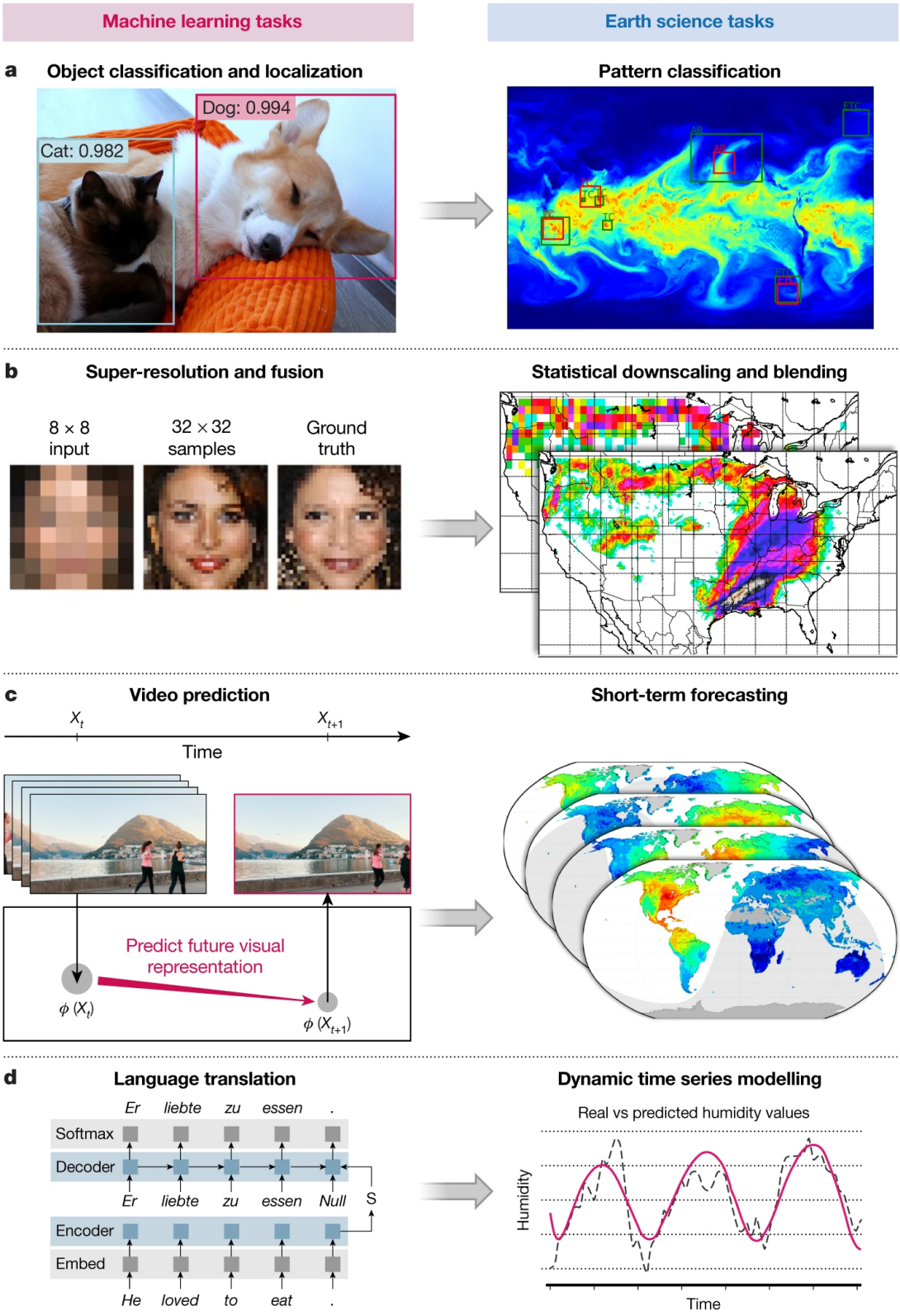

Figure 1: a) Object recognition in images links to classification of extreme weather patterns using a unified convolutional neural network on climate simulation data41. b) Super-resolution applications relate to statistical downscaling of climate model output72. c) Video prediction is similar to short-term forecasting of Earth system variables. d) Language translation links to modeling of dynamic time series.
Source: Figure 2 in https://www.nature.com/articles/s41586-019-0912-1.

2 Tentative Schedule

To stay up-to-date, consider adding the course's Google Calendar to your own calendar: Google Cal Link. The dates and times indicated in parentheses were planned in Summer 2021 and may have changed since then; please refer to the Google Calendar for the most up-to-date information on dates, times, and location. We highly encourage you to complete the assigned readings and their guiding questions on Moodle no more than 24 hours before lecture so that you get the full 15% of the "Readings" grade.

Week 1: Introduction to the Course, Linear/Logistic Regression for Classification/Regression
Python/Google Colab Notebooks/Git basics, Training/Validation/Test split, Best practices for training and benchmarking.
Reading = Ch 3+4 of Géron
Recommended Reading = Ch 1+2 of Géron

Week 2: Decision Trees/Random Forests/SVMs & Environmental Risk Analysis
Ensemble Learning, RVM, Gaussian Processes.
Reading1 = Ch 5+6+7 of Géron
Reading2 = A ML-Based Approach for Wildfire Susceptibility Mapping. The Case Study of the Liguria Region in Italy (paper)

Week 3: Unsupervised Learning for Clustering/Dimensionality Reduction & Environmental Complexity
K-Means, DBSCAN, Hierarchical clustering, t-SNE, Gaussian Mixtures, Variational Inference.
Reading1 = Ch 8+9 of Géron
Reading2 = Revealing the Impact of Global Heating on North Atlantic Circulation Using Transparent Machine Learning (article/code)

Week 4: Artificial Neural Networks & Surrogate Modeling
Reading1 = Ch 10+11 of Géron
Reading2 = Could Machine Learning Break the Convection Parameterization Deadlock? (article/code)

Week 5: Convolutional Neural Networks & Remote Sensing
Fully Convolutional Networks, ResNets, U-Nets.
Reading1 = Ch 14 of Géron
Reading2 = Review on Convolutional Neural Networks in vegetation remote sensing (article/code)

Week 6: Recurrent Neural Networks & Hydrological Modeling
Attention, Transformers.
Reading1 = Ch 15+16 of Géron
Reading2 = Towards learning universal, regional, and local hydrological behaviors via ML [...] (article/code)

Week 7: Explainable Artificial Intelligence & Understanding/Communicating Predictions
Permutation tests, Partial-dependence plots, Saliency maps, Feature visualization.
Reading1 = Extracts from "Interpretable ML" by Christoph Molnar (book)
Reading2 = Interpretable Deep Learning for Spatial Analysis of Severe Hailstorms (article/code)

Week 8: Office Hours for Final Projects and In-Class Presentations
1) Possible overview of student-chosen related topics not covered in class that may be relevant to final projects:
Bayesian inference, Causal discovery/inference, Data ethics, Generative Modeling, Graph neural networks, Knowledge-guided ML, Reinforcement Learning, Symbolic Regression.
2) In-class peer review: Each class member submits the draft of their final project for review and reviews 3 drafts from peers.

3 Final Project Guidelines

The final project's goal is to answer a well-defined scientific question by applying one of the machine learning (ML) algorithms introduced in class on an environmental dataset of your choice (e.g., related to your Masters thesis or your PhD research). You may collaborate with peers on the same dataset and present your projects together in class (40% of final project's grade), but you must choose distinct scientific questions and write separate 4-page final reports (40% of final project's grade) using this template on Overleaf (LaTeX tutorial at this link). To ensure a smooth peer-review process, you are encouraged to submit a draft of your written report (it does not have to be final) one day before the in-class peer review (20% of final project's grade). If you write constructive peer reviews, you can obtain as many bonus points as 30% of the final project's grade (10% per constructive peer review). Your final project will be made publicly-accessible on the course's website at the end of the semester.

3.1 Timeline

- Weeks 1-2: Pick a relatively large (if possible, more than 1000 samples) environmental dataset linked to a scientific question you are passionate about (see Section 3.4 for more guidance). **Don't hesitate to discuss possible projects with the instructor, the TA, and your classmates during labs or office hours**, especially if you are struggling to find an environmental dataset, define a scientific question, or choose an appropriate ML algorithm for the task at hand. We will schedule dedicated, one-on-one office hours to help you kickstart your final project.
- Weeks 3-8: Work on your final project at the end of labs and outside of the classroom. **Don't hesitate to discuss pitfalls and brainstorm solutions with the instructor, the TA, and your classmates during labs or office hours.**
- Week 8: Please submit a draft of your final project (even if still in-progress) one day before the in-class peer review. **This deadline is especially important** as the in-class peer-review process will not be possible without everyone's submission.
- Please submit the final report in PDF format via Moodle (the deadline is on Moodle).

3.2 Deliverables

1. A written 4-page[1] final report addressing a well-defined scientific question by applying one of the machine learning (ML) algorithms introduced in class. Use this template on Overleaf (LaTeX tutorial at this link) and include:
   - An informative title, and an abstract with no more than 6 sentences (follow this link for tips),
   - At least 2 Figures to communicate your methodology and your results,
   - At least 2 Tables with (1) the range of your hyperparameter search and the hyperparameters you chose using your validation set; and (2) at least 2 carefully-chosen performance metrics evaluated over the training, validation, and test sets.
   - A link to the dataset you used,
   - A link to a well-documented Python notebook/script on GitHub for your project.
2. A short[2] in-class presentation of your project, clearly communicating your scientific question, introducing your dataset, explaining your methodology and the reason you chose a particular ML algorithm, and summarizing your findings.

3.3 Evaluation & Peer-Review Process

The final evaluation and the in-class peer-review process will both use the same rubric at this link.

3.4 Resources

We encourage you to use data from your Masters or PhD research. If you are still looking for the right dataset, consider reaching out to your Masters/PhD thesis' advisor. You may also browse Kaggle datasets, this list of benchmark datasets (maintained by Pangeo), the linked datasets from last year's final projects, the environmental datasets from the course's syllabus (refer to the course's GitHub repository), and this list of open geography datasets and final projects suggestions kindly provided by Dr. Yu.

[1] excluding references

[2] the exact duration will be based on the total number of presentations and will include time for questions

4 Resources and Ethics

4.1 UNIL/FGSE Resources for Students

- Disability resources. If you need academic support, please email me (preferentially before the class starts) so that I can request/provide the appropriate services.
- English resources: Many of us are not native English speakers, and UNIL provides a wealth of resources to practice English, including free consultations/workshops for essay/paper writing, which may come in handy when writing up your final project.
- Financial support resources.
- Confidential and free mental health resources provided by the university's hospital.

4.2 Diversity and Inclusion in the Classroom

The University of Lausanne is committed to equal opportunity and stands firm against all forms of discrimination, including discrimination based on based on race, gender, religion, country of origin, ethnicity, socioeconomic status, sexual orientation, and disability. There are confidential resources if you feel harassed, and advice/mediation resources.

In the context of our classroom, this means:

- Choosing how you would like to be addressed by indicating your preferred name and pronouns in the initial course survey,
- Openly discussing and asking about concepts we struggle with to normalize difficulties in learning and applying course materials,
- Being kind and understanding towards each other: Especially in an interdisciplinary and international environment, concepts that seem obvious to you may be unknown to others or have different names depending on your sub-field,
- Emailing me or the equal opportunity office if you feel that students are not treated evenhandedly, or if the context/structure of the course is negatively impacting your learning experience and performance,
- All recognizing and working on our implicit biases by actively listening to each other.

4.3 Late Work Policy

Late work is eligible for partial credit of 50% until the official end of the semester.

4.4 Academic Integrity

At UNIL, we all share strict rules on academic integrity, which can be found at this link (in French). In the context of this course, the following behavior can lead to an automatic failure of the class (grade of 0%):

1. Plagiarism. To avoid plagiarism, always cite your sources: at the bottom of your slides during the final presentation, including for photos/schematics, and using bibtex when writing your final report using Overleaf. Please do not take credit for someone else's work and do not have someone write in your name (this also applies to guiding questions during readings).
2. Unauthorized collaboration. Even if you collaborate with some of your peers on the final project, you must answer distinct questions and write separate reports. Please transparently acknowledge any help you received from your peers (coding, research ideas, writing, proofreading, data, citations, etc.) in the acknowledgments section of your final report. During graded quizzes in class, please do not copy your peers' responses. Even if collaboration is highly encouraged, do not copy your peers' code during computer labs. Between classes, do not copy your peers' answers to the readings' guiding questions.
3. Data fabrication or falsification. Please do not fabricate the data reported in the analyses, figures, and tables of your final report. Being transparent about the shortcomings of a method or a dataset is always helpful to the community.