Machine Learning in Earth & Environmental Sciences

Course Syllabus – Mohammed Ombadi (ombadi@umich.edu)

Course Description

The rapid increase in environmental data acquisition over the last few decades has provided us with an unprecedented opportunity to draw insights from big data on the behavior of environmental systems. This course aims to introduce students to statistical methods, ranging in complexity from autoregression to machine learning models.

The course covers the basic theory behind machine learning and provides hands-on experience in building machine learning models. Students will learn to apply these models for both prediction and hypothesis formulation purposes. The methods will be taught through example applications in environmental sciences, with a specific focus on climate and hydrologic applications.

Examples include short-term forecasts of temperature and precipitation, streamflow forecasting in selected hydrologic basins, understanding the relative contributions of temperature and precipitation in snowmelt trends, regional clustering of precipitation patterns and trends, and climatic teleconnections in regulating regional precipitation patterns.

Learning Outcomes

After successfully completing this course, students will be able to:

- Identify the basic statistical features of environmental data.
- Understand the differences between parametric/non-parametric and linear/nonlinear models and select the most appropriate statistical models based on the research problem and/or availability of data.
- Conduct exploratory analysis on data and develop hypotheses regarding variable interactions.
- Develop data-driven models for prediction and evaluate their adequacy.
- Gain an understanding of the basic theory behind machine learning models.

Teaching Method and Philosophy

Teaching Methods and Philosophy: The teaching method for this course combines lectures, small group work, and active (experiential) learning. Lectures will introduce concepts, methodologies, and examples of applications in climate and environmental sciences.

Small group work is primarily focused on the final project, where each group of 2 to 4 students will apply some of the methods introduced in the course to a problem of their own interest. Experiential learning is integrated into homework assignments, where students will receive Python codes and data files. They will be tasked with applying the methods and answering a set of questions.

Class Climate & Inclusivity

Diversity is not only appreciated but celebrated in this classroom. The teaching staff maintains a zero-tolerance policy towards any form of discrimination. Students are expected to demonstrate respect, civility, and considerate conduct. Additionally, recognizing differences in language, culture, and personal viewpoints is encouraged. Feedback on issues related to diversity in the classroom is welcomed.

Required Computing Software

Coding scripts for the data analysis methods will be provided in Python. Students are welcome to use R or MATLAB to complete assignments if they prefer, but they must write their own scripts.

Grade Breakdown

Participation in class / Quizzes	10%
Homework Assignments	60%
Project	30%

Course Schedule

Week	Topic	Application Example
1	**Review of basic statistics:** Random variables, measures of central tendency (mean, median, mode), variance, higher moments (skewness and kurtosis), sample vs population statistics, probability distributions	Summarize the statistics of temperature and precipitation in Ann Arbor, Michigan
2	**Hypothesis testing and statistical dependence:** one- and two-sided hypothesis testing, statistical significance, confidence interval, Pearson correlation, nonparametric correlations (e.g., Spearman), partial correlation	Explore changes in global mean surface temperature (compare two periods), Explore the relationship between snowmelt, precipitation and temperature over selected watersheds in Western US
3	**Trend Analysis:** Linear regression, non-parametric methods (Mann-Kendall test), stationarity	Explore trends in global mean surface temperature.
4	**Clustering algorithms:** K-means, density-based clustering, expectation-maximization methods	Identify regional precipitation patterns in the Contiguous United States
5	**Multiple linear regression:** Maximum likelihood parameter estimation, loss functions (L1 and L2 norms), model evaluation and goodness of fit.	Prediction of streamflow using input variables of precipitation, evapotranspiration and soil moisture. Build models with different number of inputs

		and evaluate their goodness of fit.
6	**Intro to machine learning:** Gradient descent and backpropagation, feedforward neural networks, Multilayer Perceptron	Repeat application in 5 using feedforward neural networks.
7	**Decision Trees:** Random Forest models	Repeat application in 5 using random forests with/without discretization.
8	**Recurrent Neural Networks:** Applications to temporal dynamics, long-short term memory	Snowmelt prediction
9	**Convolutional Neural Networks:** Applications to image processing (e.g., remotely sensed and gridded datasets)	Short-term forecast of precipitation using gridded observations
Students start forming groups and discuss project topics with instructors		
10	**Autoencoders and representation learning**	Applications for dimensionality reduction of atmospheric pressure data
11	**Physics-informed machine learning:** Enforcing simple physics-based constraints, nonlinear dynamics-based neural networks	Modeling Evapotranspiration using water balance as a constraint
12	**Information theory:** Entropy, Mutual information, conditional mutual information, KL Divergence	Explore the relationship between snowmelt, precipitation and temperature over selected watersheds in Western US.
13	**Causal Inference:** Granger and predictive causality, Transfer Entropy, Graph-based methods	Evaluating the relative contribution of environmental variables in driving evapotranspiration
14	**Graphical models and Graph Neural Networks**	Applications to streamflow and water quality modeling