Fall 2021

Machine Learning and AI with Bioinformatics Applications

Prof. Konstantinos Krampis

Wednesdays 10am - 12pm, starting August 25, 2021
Online via Zoom : https://us02web.zoom.us/j/87135234123

Course purpose and description.

The purpose of this course, is to cover the basic theory behind machine learning and artificial intelligence (AI) algorithms for biology students, in addition to a range of practical applications of the algorithms for the analysis of genomic and bioinformatics data. The field of machine learning and artificial intelligence algorithms has grown exponentially in the last decade, and recently multiple applications of these algorithms for the analysis of genomic data became available in the literature.

The topics will range from the biological paradigms followed in building artificial neural networks, to the statistical underpinnings of machine learning, and hands-on examples using machine learning to analyze genomic data. Prior coding experience in Python or similar scripting language is a plus but not required, as long as students are willing to think computationally and interpret the code examples is sufficient. Furthermore, willingness to delve into statistical and mathematical concepts that are key for understanding artificial intelligence is required. Topics that will be covered also include Next-Generation Sequencing (NGS) data formats as these are used as input to the algorithms, in addition to programming concepts with AI frameworks such as Google Tensorflow (in Python).

This is an interdisciplinary course for students interested in a data-centric view of biological systems, and computational approaches to understand biology through large-scale genomic data and machine learning / AI algorithms. Towards the end of the semester we will also cover the latest topics in machine learning / AI (that have not found their way in bioinformatics and genomic data analysis literature yet), with the goal of possibly inspiring students to pursue research in this area as part of their thesis.

Course topics by week.

1. Computational representation of a neuron, overview of data representation and statistical learning processes.
2. Single layer perceptron artificial neural networks and linear regression, genomic applications and limitations.
3. Unsupervised and data clustering algorithms, principal components, k-means, support vector machines, genomic applications and limitations.
4. Foundations part A: basic matrix algebra concepts, introduction to Python / Tensorflow.

5. Foundations part B: Python continued, basics of Next-Generation Sequencing (NGS) data formats.
6. Multi-layer artificial neural networks introduction, structure, function and advantages in genomic data analysis and classification.
7. Structure and algorithmic processes for deep learning / data classification, core analysis concepts required such as back-propagation, gradient descent optimization.
8. Convolutional neural network structure and learning, applications for genomic data classification and analysis.
9. Recurrent neural network structure and learning, models for genomic sequence analysis.
10. Hybrid neural network structures, applications for genomic data classification and analysis, aditional topics such as autoencoders for data noise removal / dimension reduction.
11. Advanced topics: deep reinforcement data learning and classification algorithms, potential applications to genomic data.
12. Advanced topics: adversarial data learning and classification with artificial neural networks.
