BE562 Computational Biology: Machine Learning Fundamentals
Fall 2022 Course Information

Lectures Tu/Th 1:30-3:00, Room LSE B03
Recitations Fri 9:05-9:55, Room LSE B03

In this course we cover the algorithmic and machine learning foundations of computational biology, combining theory with practice. We study principles of machine learning, algorithm design and computational biology; we provide an introduction of important problems in computational biology; and we provide hands on experience analyzing large-scale biological data sets.

Topics (subject to change) include:

Foundational Topics (First Half of Course)

- Sequence Alignment
- Clustering
- Classification
- Hidden Markov Models
- Motif Prediction, EM, and Gibbs Sampling
- Phylogenetics

Advanced Topics (Second Half of Course)

- Gene Prediction
- Generalized HMMs
- Conditional Random Fields
- Logistic Regression
- Baysian Networks
- Sampling Methods and MCMC
- Neural Networks

These topics are grounded in fundamental algorithmic and machine learning techniques including: maximum likelihood, Bayesian analysis, dynamic programming, Gibbs sampling, Expectation Maximization, hidden Markov models, Bayesian networks, and sampling.

Prerequisites: fundamentals of programming and algorithm design (EK 127 or equivalent), basic molecular biology (BE 209 or equivalent), statistics and probability (BE 200 or equivalent).

Course Staff

Lecturer: James Galagan, LSEB 1002, jgalag@bu.edu, 617-875-9874

TA: TBD

Website

The course website is being hosted via BU Blackboard 8 at http://learn.bu.edu/. You should be able to access the site if you are registered for the course. If you have any problem with this, please email us.

Grading

Your grade in this course will be based on the following:

- Problem sets (35%)
- Midterm Exam (20%)
- Final Project (40%)
- Scribing (5%)

Problem Sets

There will be three problem sets during the first half of the semester. The problem sets will include both theoretical and programming problems. For programming problems, we will provide skeleton code in Python, but you may use a different programming language if you so choose.

Midterm Exam(s)

There will be two midterm exams during the course, both of which will cover all material up until that point. There will be no final exam.

Final Project

You will complete a final project during the second half of the semester. You may either work alone or with one partner. Teams will be expected to undertake more ambitious projects. In previous years, approaches to the final project have included:

- Compare several computational biology algorithms for solving the same problem, by implementing them, applying them to some dataset, and evaluating the results.
- Design and apply a novel computational biology algorithm and evaluate its performance and effectiveness.

We will distribute more detailed project expectations and suggested project topics as the term progresses.

Final Project Proposals

Prior to beginning work on the final project there will be a proposal submission and review process modeled after the NIH grant/fellowship application process. You will submit a proposal following guidelines that mimic those of an NIH grant/fellowship. You will then be assigned three submitted proposals to critique and your proposal will be submitted to three (anonymous) reviewers. A revised proposal incorporating feedback will be due a week after you receive feedback from the staff and a summary of the critiques of your own proposal written by other students, and your final project will be due three weeks thereafter.

Scribing

Each student will be required to scribe for one lecture. Several students may be assigned to work together on each lecture, depending on course enrollment. As a scribe, you should strive produce a self-contained narrative of the lecture. However, the slides for each lecture will be available on the course web site, so you should pay particular attention to issues that the slides don't convey well on their own.

The scribe notes from past years will be made available for you to use as a starting point - if the lecture was given last year. Your goal is not to recreate the scribe notes from scratch, but rather to improve upon the existing notes by adding content for any new topics, clarifying confusing topics, and expanding upon topics that require more explanation.

Recitations

A weekly recitation will be held on Fridays, during which the TA will discuss additional aspects of the lecture material and hold Q&A. The TA will also hold office hours, time and location TBA.

Textbooks

Although no textbook is mandatory for this course, we recommend the following textbooks as references.

- Richard Durbin, Sean R. Eddy, Anders Krogh and Graeme Mitchison, Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.
- Neil Jones and Pavel Pevzner, An Introduction to Bioinformatics Algorithms.
- Michael Waterman, Introduction to Computational Biology
- Richard Duda, Peter Hart, David Stork, Pattern Classification.
- Daphne Koller and Nir Friedman, Probabilistic Graphical Models, Principles and Techniques
- Uri Alon, An Introduction to Systems Biology: Design Principles of Biological Circuits

Collaboration Policy

You are welcome to collaborate on problem sets and the final project. However:

- You must work independently on each problem before you discuss it with others.
- You must write the solutions on your own.
- You must acknowledge outside sources and collaborators.