BE 700: Foundations of Biomedical Data Science and Machine Learning (Spring 2023)

Instructors: Michael Economo (mne@bu.edu) & Brian DePasquale (bddepasq@bu.edu)
TA: Harrison Fischer (hfisher@bu.edu)

Class: T/Th 1:30-3:15
Recitation: F 2:30-3:20
Programming help: F 3:20-4:15

Office Hours: By appointment. 24 Cummington St, Room 201 (Economo) & TBD (DePasquale)

Course documents: on Blackboard (https://learn.bu.edu) and course GitHub repository

Course description: This course will cover conceptual and practical aspects of data science and introductory machine learning for biomedical engineers. This course serves as a foundational course in data analytics for BME Ph.D. students. It is designed to follow a graduate-level introductory programming course and will prepare students for graduate-level courses and research focused on more advanced applications of machine learning and data science. This course will cover the theory and practical applications of hypothesis testing, model fitting and parameter estimation, classification, clustering, dimensionality reduction, and machine learning.

Course Goals: This course takes a practical approach to the analysis of biomedical data. By doing so, three goals are strived for. First, students will become familiar with the necessary theoretical background of different analysis methods, empowering them to understand why certain methods are appropriate in certain contexts and why others are not. Second, students will acquire practical, hands-on skills necessary for analyzing biomedical data including data management, algorithm development, and proper codebase development. These skills will prepare students for independent research projects both within academic research and within industry. Third, students will learn how to interpret, visualize, and summarize the results of analysis, once completed. Applying analysis methods is only half the challenge of scientific discovery. The third goal of this course is to train students to collect the results of scientific analysis into a format that allows scientific discoveries to be shared with and understood by other researchers.

Prerequisites: BE 601 (Linear Algebra), or equivalent familiarity with linear algebra (e.g., linear systems of equations, eigendecomposition, bases, orthogonality, matrix decomposition, etc.), and BE 604 (Statistics & Numerical Methods) or equivalent familiarity with probability and statistics (e.g., random variables, conditional and marginal probability, Bayes' rule, etc.). ENG BE 500 (Programming Fundamentals for Biomedical Engineering Data Analysis with Python) or equivalent programming experience.

Grading, homework, and exams: 80% homework completion and recitation participation, 20% per-module take-home exam. Homework set every 2-3 weeks. There will be a take-home exam at the end of each module (it will be similar in form to the homework sets). Programming exercises to be completed in Matlab or Python, or a language of your choice with instructor permission. Solutions to be pushed to the course GitHub repository on the assigned due date. Collaboration on the homework is allowed (and encouraged!) but your submitted work must be your own.

Recitation and programming help: Recitations will cover questions from lecture, example problems from the week's material, and introduce homework material. Then, there will be an optional programming help session where we will review basic programming concepts, discuss implementation differences between languages (e.g., Matlab vs. Python), and provide support for those learning advanced programming packages (e.g., Tensorflow, PyTorch, SciPy, etc.).

Textbook: Pattern Recognition and Machine Learning. 2006. Christopher Bishop. (PDF)

Tentative topics (Spring 2023)

Module & Lectures	Topics
Math and programming review (1-2), Model fitting/estimation (3-6)	• Programming review • Review of probability • Review of statistics • Classic statistics • Resampling-based statistical methods (e.g., bootstrap, Monte Carlos methods, etc.) • Fitting models to data • Generalizability and validity • Regularization • Maximum likelihood/MAP • Linear regression
Classification (7-8), Clustering (9-10)	• Concept of classifiers • Types of errors • ROC • Logistic regression • SVM • Partitioning vs. hierarchical clustering • Distance measures • Stochastic vs. deterministic methods • K-means
Bayesian classifiers/detectors (11-14), Dimensionality reduction (15-18)	• Naïve Bayes classifiers • EM algorithm • Gaussian mixture models, latent variables • Hidden Markov models • Multiclass problems • Linear methods (PCA, Factor analysis, etc.) • Nonlinear methods (Isomap, t-SNE, etc.) • Two-dimensional data
Neural networks (19-22), Decision trees/forests (23-24)	• Perceptron learning • Gradient descent • Backpropagation • Programming neural networks (Tensorflow, PyTorch) • Decision trees • Shannon entropy/information theory • Random forests