Introduction to Machine Learning for Biology (BIOS 26122)

winter 2024

Dmitry Kondrashov

Instructors:

Dmitry Kondrashov dkon@uchicago.edu

Course logistics

- Course meetings: Tue/Thu 9:30 - 10:50 am; BSLC TBA
- Lab sections: Fri 1:30-3:30 pm; BSLC TBA
- Office hours: Wed 9:30 - 11 am; Fri 9:30 - 11 am; BSLC 301A

Course Description

Machine learning techniques are essential in many fields of biology that rely on large amounts of data. This course is intended to introduce key concepts in this field and illustrate their applications to biological questions. Students will learn about methods for supervised and unsupervised learning; regression and classification algorithms, and dimensionality reduction approaches. With every method we will emphasize model selection and validation on real data sets. Computational labs are an integral part of the course for students to work on applying these methods using R in the Quarto document system.

Reading resources

- Textbook: Introduction to Statistical Learning by James, Witten, Tibshirani, and Hastie
- R for Data Science, 2e by Wickham, Çetinkaya-Rundel, and Grolemund.
- Introduction to Modern Statistics Çetinkaya-Rundel and Hardin.

Learning goals

In this course, you will…

- Understand concepts for applying machine learning methods, such as cross-validation and bias-variance tradeoff
- Know when to use regression and when to use classification for supervised learning
- Wrestle with unsupervised learning and dimensionality reduction
- Use resampling and bootstrapping for validation and multiple hypothesis testing
- Apply these methods to a variety of biological data sets
- Critically evaluate methods section of a paper that uses these methods

How to get help

Students are expected and encouraged to use the in-person interactions in the classroom to ask questions and discuss the information introduced in the course. Please attend and participate in the in-class coding activities to get immediate help!

The instructor will hold office hours as indicated above; please use them to ask questions about the course, for planning the final project, or for general communication.

We will use the Ed Discussion platform as the primary means of answering questions outside of the classroom. It is available on the left-hand side menu of the Canvas course. You can post questions, either using your name or anonymously, and they may be answered by other students or myself in a time frame of no more than a few hours.

You are welcome to email the instructors with personal requests, but please ask any course material questions through Ed Discussion so others may benefit.

Weekly Schedule

Week	Day	Topic	Date
Week 1	Thursday	Learning from data	4-Jan-24
Week 2	Tuesday	Bias-variance tradeoff (chapter 2 ISLR)	9-Jan-24
Week 2	Thursday	K nearest neighbors (chapter 2 ISLR)	11-Jan-24
Week 3	Tuesday	Linear regression (chapter 3 of ISLR)	16-Jan-24
Week 3	Thursday	Linear regression (chapter 3 of ISLR)	18-Jan-24
Week 4	Tuesday	Classification: Naïve Bayes (chapter 4 of ISLR)	23-Jan-24
Week 4	Thursday	Classification: logistic regression (chapter 4 of ISLR)	25-Jan-24
Week 5	Tuesday	Cross-validation (chapter 5 of ISLR)	30-Jan-24

Week	Day	Topic	Date
Week 5	Thursday	Bootstrapping (chapter 5 of ISLR)	1-Feb-24
Week 6	Tuesday	Linear model selection (chapter 6 of ISLR)	6-Feb-24
Week 6	Thursday	Regularization and regression in high dimensions (chapter 6 of ISLR)	8-Feb-24
Week 7	Tuesday	Decision trees (chapter 8 of ISLR)	13-Feb-24
Week 7	Thursday	Random forests (chapter 8 of ISLR)	15-Feb-24
Week 8	Tuesday	PCA and dimensionality reduction (chapter 12 of ISLR)	20-Feb-24
Week 8	Thursday	Clustering methods (chapter 12 of ISLR)	22-Feb-24
Week 9	Tuesday	Multiple hypothesis testing (chapter 13 of ISLR)	27-Feb-24
Week 9	Thursday	False discovery rate and resampling (chapter 13 of ISLR)	29-Feb-24

Assessments

The final course grade will be calculated as follows:

Category	Work	Final grade percentage
R labs	8 assignments	40%
Weekly quizzes	take on Canvas	20%
Midterm	in-class exam	10%
Final exam	in-class exam	20%
Surveys	complete on Canvas	10%

Course policies

general expectations

The ethos of the course is collaborative rather than competitive. I will provide flexibility with timing if needed, and I am happy to offer one on one support. I ask that you facilitate this process by communicating promptly in case there is a need for help - the more communication, the better!

If there is any portion of the course that is not accessible to you due to challenges with technology or the course format, please let me know so we can make appropriate accommodations.

I will not take attendance or penalize you for missing classes, but I ask that you show up and participate in discussion and coding work. Being helpful and respectful of your fellow learners is, of course, expected.

collaboration and academic integrity

Students are expected to work in accordance with the University Policy of Academic Integrity, which can be found here.

Please abide by the following guidelines as you work on assignments in this course:

- You are welcome to work together with others, but what you turn in must have been typed or written by you, and not copied from another. Evidence that a student has copied another one's work will result in a grade of 0 for the assignment, and may trigger disciplinary action.
- Reusing code: You may make use of online resources (e.g. StackOverflow, ChatGPT, or other LLMs) for coding examples on assignments. If you directly use code from an outside source, you must explicitly cite where you obtained the code.

late work & extensions

The due dates for R assignments are there to help you keep up with the course material and to ensure the we can provide feedback within a timely manner. If circumstances come up to prevent your submitting an assignment by the deadline, please notify the instructor as soon as you know and we will work with you to find accommodations. Untimely requests or lack of communication may make it impossible to provide accommodations.

Each R assignment will have an initial deadline and a resubmission deadline. If the work is submitted by the initial deadline, even if incomplete, you will receive feedback and a chance to revise the work by the resubmission deadline. If you don't submit an initial version, you may still submit the work by the final deadline, but you will not get a chance to resubmit.