BIOS 635: Introduction to Machine Learning

Kevin Donovan

Spring 2021

E-mail: kmdono02@ad.unc.edu
Web: https://kmdono02.github.io/
Office Hours: W 10-11am
Class Hours: T/Th 9:30-10:45am
Office/Classroom: https://uncsph.zoom.us/j/8587296876

Course Description

This course is an introductory course to machine learning and statistical learning and is required for MPH students with Data Science concentration. While some technical details will be covered, emphasis will be made on understanding the models, intuitions, and strengths and weaknesses of the various approaches. The primary goal is to provide conceptual understanding of machine learning methods and how these relate back to statistical principles. The secondary goal is to equip students with knowledge of existing tools for data analysis and to get students prepared for more advanced courses in machine learning. Good research principles with respect to machine learning analyses will also be covered (reproducibility, legible and complete code, etc.). Programming language will be R – students will learn how to use the free and powerful software R in connection with each of the methods exposed in the class. For deep learning, Keras/TensorFlow in Python will be introduced if time permits.

Prerequisites

Prerequisites: BIOS 611, BIOS 550, R programming language, college-statistics course

Instructor

Kevin Donovan
PhD Candidate
Department of Biostatistics
Phone: 315-727-3603
Email: kmdono02@ad.unc.edu

Teaching Assistant

Emily Damone
Email: edamone@live.unc.edu
Office Hours: Monday 3:30-4:30PM

Course Materials

Course Website

GitHub: **https://github.com/kmdono02/BIOS_635**
Slack: **...**

Course Texts

Required Textbook:

- James, Witten, Hastie, and Tibshirani, An Introduction to Statistical Learning, Springer (**FREE**).

Recommended Textbooks:

- Grolemund and Wickham, R for Data Science, O'Reilly (**FREE**).
- Irizarry, Introduction to Data Science: Data Analysis and Prediction Algorithms in R, CRC Press (**FREE**).

Advanced Textbooks:

- Hastie, Tibshirani, Friedman, The Elements of Statistical Learning, Springer (**FREE**).
- Efron and Hastie, Computer Age Statistical Inference, Cambridge University Press (**FREE**).

Course Format

The course format will include two weekly lectures via video call interfacing through Zoom at the provided link on page 1. The lecture will be supplemented with in-class exercises, case studies and examples for real data set analysis. Published research in academic journals and non-academic sources will be discussed and critical evaluated.

Course Objectives

1. Gain conceptual understanding of machine learning through a statistical framework
2. Be introduced to standard machine learning methods and when each is most appropriate to use in the analysis
3. Learn how to implement these methods using R (and Python if time permits) through real data analysis

4. Understand complications with applying machine learning methods to data and in society at large (e.g. generalizability, data ownership, group representation in data, etc.)
5. Gain skills in critically evaluating published academic and non-academic research which use machine learning methods

Course Policies and Resources

Recognizing, Valuing and Encouraging Inclusion and Diversity in the Classroom

We share the School's **commitment to diversity**. We are committed to ensuring that the School is a diverse, inclusive, civil and welcoming community. Diversity and inclusion are central to our mission — to improve public health, promote individual well-being and eliminate health inequities across North Carolina and around the world. Diversity and inclusion are assets that contribute to our strength, excellence and individual and institutional success. We welcome, value and learn from individual differences and perspectives. These include but are not limited to: cultural and racial/ethnic background; country of origin; gender; age; socioeconomic status; physical and learning abilities; physical appearance; religion; political perspective; sexual identity and veteran status. Diversity, inclusiveness and civility are core values we hold, as well as characteristics of the School that we intend to strengthen.

We are committed to expanding diversity and inclusiveness across the School—among faculty, staff, students, on advisory groups, and in our curricula, leadership, policies and practices. We measure diversity and inclusion not only in numbers, but also by the extent to which students, alumni, faculty and staff members perceive the School's environment as welcoming, valuing all individuals and supporting their development." In this class, we practice these commitments in the following ways:

- Develop classroom participation approaches that acknowledge the diversity of ways of contributing in the classroom and foster participation and engagement of all students.
- Structure assessment approaches that acknowledge different methods for acquiring knowledge and demonstrating proficiency.
- Encourage and solicit feedback from students to continually improve inclusive practices.

As a student in the class, you are also expected to understand and uphold the following UNC policies:

- **Diversity and Inclusion at the Gillings School of Global Public Health**
- **UNC Non-Discrimination Policies**
- **Prohibited Discrimination, Harassment, and Related Misconduct at UNC**

Accessibility

UNC-CH supports all reasonable accommodations, including resources and services, for students with disabilities, chronic medical conditions, a temporary disability, or a pregnancy complication

resulting in difficulties with accessing learning opportunities. All accommodations are coordinated through the UNC Office of Accessibility Resources & Services (ARS), **https://ars.unc.edu/**; phone 919-962-8300; email **ars@unc.edu**. Students must document/register their need for accommodations with ARS before accommodations can be implemented.

Counseling and Psychological Services

CAPS is strongly committed to addressing the mental health needs of a diverse student body through timely access to consultation and connection to clinically appropriate services, whether for short or long-term needs. Go to their website: **https://caps.unc.edu** or visit their facilities on the third floor of the Campus Health Services building for a walk-in evaluation to learn more.

UNC Honor Code

As a student at UNC-Chapel Hill, you are bound by the university's Honor Code, through which UNC maintains standards of academic excellence and community values. It is your responsibility to learn about and abide by the code. All written assignments or presentations (including team projects) should be completed in a manner that demonstrates academic integrity and excellence. Work should be completed in your own words, but your ideas should be supported with well-cited evidence and theory. To ensure effective functioning of the Honor System at UNC, students are expected to:

1. Conduct all academic work within the letter and spirit of the Honor Code, which prohibits the giving or receiving of unauthorized aid in all academic processes.
2. Learn the recognized techniques of proper attribution of sources used in written work; and to identify allowable resource materials or aids to be used during completion of any graded work.
3. Sign a pledge on all graded academic work certifying that no unauthorized assistance has been received or given in the completion of the work.
4. Report any instance in which reasonable grounds exist to believe that a fellow student has violated the Honor Code.

Instructors are required to report suspected violations of the Honor Code, including inappropriate collaborative work or problematic use of secondary materials, to the Honor Court. Honor Court sanctions can include receiving a zero for the assignment, failing the course and/or suspension from the university. If you have any questions about your rights and responsibilities, please consult the Office of Student Conduct at **https://studentconduct.unc.edu/**, or consult these other resources:

- Honor system module.
- UNC library's plagiarism tutorial.
- UNC Writing Center handout on plagiarism.

Instructor Expectations

Email

The instructor will typically respond to email within 24 hours or less if sent Monday through Friday. The instructor may respond to weekend emails, but it is not required of them. If you receive an out of office reply when emailing, it may take longer to receive a reply. The instructor will provide advance notice, if possible, when they will be out of the office.

Discussion Board

The instructor will be an active reader and will occasionally post throughout the semester to discussion boards on Slack. The group discussion boards will be moderated by the group members unless an issue is brought to the instructor's attention by a fellow group member.

Grading

Assignments, projects and discussion board postings will be graded no more than two weeks after the due date. Assignments that build on the next assignment will be graded within one week of the final due date. Early submissions will not be graded before the final due date.

Collaboration

Collaboration will be expected and required on all group projects; assignments consider group projects will be explicitly marked as such. For weekly homework assignments, collaboration with other students is welcome and encouraged with respect to **all coding-related** aspects of the work. However, collaboration will be **not be allowed** with respect to other aspects of the work (interpretation of results, answers needing written responses, etc.). Examinations will be done remotely with collaboration with others not being allowed in any form. Collaboration with any outside textual resources (online, digital, or otherwise) as well as all course texts (textbooks, notes, etc.) will be **fully allowed** for all assignments and examinations (though written responses must be **written in your own words**).

Syllabus Changes

The instructor reserves to right to make changes to the syllabus, including project due dates and test dates. These changes will be announced as early as possible.

Student Expectations

Appropriate Use of Course Resources

The materials used in this class, including, but not limited to, syllabus, exams, quizzes, and assignments are copyright protected works. Any unauthorized copying of the class materials is a violation of federal law and may result in disciplinary actions being taken against the student. Additionally, the sharing of class materials without the specific, express approval of the instructor may be a violation of the University's Student Honor Code and an act of academic dishonesty,

which could result in further disciplinary action. This includes, among other things, uploading class materials to websites for the purpose of sharing those materials with other current or future students.

Assignments

Submit all assignments through GitHub Classroom or assignment links located in the weekly modules, syllabus link, or assignments link (if made available by your instructor). Emailing assignments is not acceptable unless prior arrangements have been made. If you are having issues submitting assignments, try a different web browser first. If switching browsers does not work, email or call the instructor for guidance.

Late Work

No late work will be excepted. This is a strict rule with no exceptions since the lowest homework score will be dropped

Technical Support

The UNC Information Technology Services (ITS) department provides technical support 24-hours per day, seven days per week. If you need computer help, please contact the ITS Help Desk by phone at +1-919-962-HELP (919-962-4357), or by email at **help@unc.edu**, or by visiting their website at **http://help.unc.edu**, or by UNC Live Chat at **http://its.unc.edu/itrc/chat**.

Student Evaluation

Course Assignments and Assessments

This course will include graded assignments and/or exams, with the student's final grade based on the following scale:

- Homework (30%)
- Article Evaluations (10%)
- Midterm (30%)
- Group Project (30%)

Grading Scale

For graduate and professional students, final course grades will be determined using the following UNC Graduate School grading scale. The relative weight of each course component is shown in the list above.

Letter	Description	Points Range
H	High Pass: Clearly excellent graduate work	≥ 93
P	Pass: Entirely satisfactory graduate work	≥ 80
L	Low Pass: Inadequate graduate work	≥ 70
F	Fail	< 70

For undergraduate students, final course grades will be determined using the following UNC Undergraduate School grading system.

Letter	Description	Points Range
A	Mastery of course content at the highest level of attainment that can reasonably be expected of students at a given stage of development. The A grade states clearly that the students have shown such outstanding promise in the aspect of the discipline under study that he/she may be strongly encouraged to continue.	≥ 93
B	Strong performance demonstrating a high level of attainment for a student at a given stage of development. The B grade states that the student has shown solid promise in the aspect of the discipline under study.	≥ 80
C	A totally acceptable performance demonstrating an adequate level of attainment for a student at a given stage of development. The C grade states that, while not yet showing unusual promise, the student may continue to study in the discipline with reasonable hope of intellectual development.	≥ 70
D	A marginal performance in the required exercises demonstrating a minimal passing level of attainment. A student has given no evidence of prospective growth in the discipline; an accumulation of D grades should be taken to mean that the student would be well advised not to continue in the academic field.	≥ 60
F	For whatever reason, an unacceptable performance. The F grade indicates that the student's performance in the required exercises has revealed almost no understanding of the course content. A grade of F should warrant an advisor's questioning whether the student may suitably register for further study in the discipline before remedial work is undertaken.	< 60

Assignment Descriptions

- Homework (30%)

Criteria	Fully Met	Partially Met	Not Met
Amount (20 points)	20 points Completed 100% of the problems assigned	15-19 points Completed 80-99% of the problems assigned	0-14 points Completed <80% of the problems assigned
Accuracy (80 points)	72-80 points	60-71 points	0-59 points

- Article Evaluations (10%)

Criteria	Fully Met	Partially Met	Not Met
Amount (10 points)	10 points Thoughtful and thorough evaluation of article given	5-9 points Complete response given, though may show incorrect or limited understanding of article	0-4 points Incomplete or no response given, shows misunderstanding of article.

- Midterm (30%)
  Point values will be assigned to each question. Each question will be graded based on whether the question is answered correctly.
- Group Project (30%)
  The students will work on an open-ended challenge problem, set up as a competition.

Criteria	Fully Met (10)	Partially Met (6-9)	Not Met (0-5)
Content (10 points)	Analysis code written clearly and efficiently with complete documentation. Project directory on GitHub is clearly organized containing all project files with reproducibility of the analysis easily facilitated and achieved	Analysis code runs without errors though may be lacking in clarity and efficiency in certain aspects. Project directory may lack in organization clarity or completeness, harming reproducibility	Analysis code runs with one or more errors or is severely lacking in clarity and efficiency, and documentation is incomplete or poorly done. Project directory may be poorly organized or incomplete, with analysis not meeting reproducibility standards in any fashion.
Subject Knowledge (10 points)	The project demonstrated knowledge of the course content by integrating major and minor concepts throughout.	The project partially demonstrated knowledge of the course content by integrating major and minor concepts throughout.	The project demonstrated very little knowledge of the course content by major and minor concepts throughout.

Course Schedule

The instructor reserves to right to make changes to the syllabus, including project due dates and exam dates. These changes will be announced as early as possible.

Week/Session	Topic and Competency
Session 1	Intro, assessing model accuracy, bias and variance tradeoff
Session 2	Regression & classification: linear regression, logistic regression, linear/quadratic discriminant analysis, naïve bayes
Session 3	Nonlinearity: polynomial regression, spline, smooth spline
Session 4	Resampling: cross-validation, bootstrap
Session 5	Model selection: penalized regression, regularization
Session 6	Tree-based methods: bagging, boosting, random forest
Session 7	Kernel methods, support vector machine
Session 8	Unsupervised learning: dimensionality reduction, clustering
Session 9	Neural networks, deep learning, and big data

More detailed schedule available **here**