THE OHIO STATE UNIVERSITY

COLLEGE OF ARTS AND SCIENCES

SYLLABUS: STAT 6500
STATISTICAL MACHINE LEARNING
SPRING 2021

Course overview

Instructor

Yoonkyung Lee
Email address: yklee@stat.osu.edu or lee.2272@osu.edu
Phone number: 292-9495
Office hours: W 3-4pm; F 11:30am – 12:30pm (Virtual Hours via Carmen Zoom)
Virtual office: Zoom link available on the Carmen website

Grader

Qiuyu Gu
Email address: gu.675@osu.edu
Office hours: by appointment only

Course description

Statistical Machine Learning explores the methodology and algorithms behind modern supervised and unsupervised learning techniques to explore relationships between variables in large, complex datasets. Topics include linear and logistic regression, classification, clustering, resampling methods, model selection and regularization, and non-linear regression. Students will also gain exposure to popular statistical machine learning algorithms implemented in R. A focus will be on understanding the formulation of statistical models and their implementation, and the practical application of learning methods to real-world datasets.

Prerequisites: STAT 6450 (Applied Regression Analysis) or permission of instructor. Familiarity with calculus, linear algebra and linear regression analysis will be assumed. Basic proficiency in a programming language, such as R is required.

Course learning outcomes

- Students will understand the statistical learning framework, including core concepts such as loss, learning, and generalization; they will be able to judge when the framework is applicable and be able to formulate problems within this framework.
- Students will recognize the role of statistical models that are appropriate for a variety of statistical learning problems; they will understand the assumptions, formulation, and evaluation of these models.
- Students will understand the rationale and algorithms behind statistical learning methods, and they will know the merits and limitations of these methods.
- Students will be able to quantitatively evaluate and compare different statistical learning methods.
- Students will be able to apply statistical learning methods for data analysis and be able to interpret the results in the context of the application.

Course materials

Required

- James, Witten, Hastie, Tibshirani: *An Introduction to Statistical Learning with Applications in R*

  (Freely downloadable PDF available at https://statlearning.com/)

Recommended

- Murphy: *Machine Learning: A Probabilistic Perspective* (An electronic version is available for online reading through the OSU library website)

Course technology

For help with your password, university e-mail, Carmen, or any other technology issues, questions, or requests, contact the OSU IT Service Desk. Standard support hours are available at https://ocio.osu.edu/help/hours, and support for urgent issues is available 24-7.

- **Self-Service and Chat support:** http://ocio.osu.edu/selfservice
- **Phone:** 614-688-HELP (4357)

- **Email:** 8help@osu.edu
- **TDD:** 614-688-8743

Baseline technical skills necessary for online courses

- Basic computer and web-browsing skills
- Navigating Carmen

Technology skills necessary for this specific course

- CarmenZoom
- Collaborating in CarmenCanvas
- Recording a slide presentation with audio narration
- Recording, editing, and uploading video

Necessary equipment

- Computer: current Mac (OS X) or PC (Windows 10+) with high-speed internet connection
- Webcam: built-in or external webcam, fully installed
- Microphone: built-in laptop or tablet mic or external microphone

Necessary software

- This class requires you to use the statistical software package called R (The R Project for Statistical Computing; http://www.r-project.org/). This software package is available as Free Software.
    - You can download R for Windows, Mac, and Linux, from the CRAN archive at https://cran.r-project.org.
    - An in-depth introduction to R is available at http://cran.r-project.org/doc/manuals/R-intro.pdf
    - Hands-on tutorials are available in the Swirl system, which you can learn about at http://swirlstats.com/. In particular, "R Programming: The basics of programming in R" is an appropriate first tutorial for students who have never used R.
- An easier to use interface to R is available in the software package RStudio. This package is available for Windows, Mac, and Linux and can be downloaded for free from http://rstudio.org. **Note that RStudio requires R to be installed.**
- Microsoft Office 365 ProPlus All Ohio State students are now eligible for free Microsoft Office 365 ProPlus through Microsoft's Student Advantage program. Each student can install Office on five PCs or Macs, five tablets (Windows, iPad® and Android™) and five phones.

- Students are able to access Word, Excel, PowerPoint, Outlook and other programs, depending on platform. Users will also receive 1 TB of OneDrive for Business storage.
- Office 365 is installed within your BuckeyeMail account. Full instructions for downloading and installation can be found https://ocio.osu.edu/kb04733.

Course delivery

Each week several lecture videos, totaling approximately 165 minutes of lecture, will be posted on the course website. You are responsible for watching the videos and studying the material that is assigned each week. In addition to the lecture videos, quizzes and assignments will be posted on the class website. You will be given ample time to complete the assignments.

The instructor will hold weekly office hours via CarmenZoom. The Zoom link for office hours will be posted on the Carmen website. The instructor will also initiate and manage active discussion boards, also via Carmen.

The vast majority of the course can be completed asynchronously, meaning that you will be able to study materials and work on assessments according to your own schedule. Live CarmenZoom meetings may also take place, if needed. The dates and times of these will be announced later.

Grading and faculty response

Grades

Assignment or category	Percentage
Quizzes (approximately every two weeks)	**10**
Homework assignments (approximately every two weeks)	**30**
Midterm	**30**
Group project	**25**
Participation	**5**
Total	**100**

* *See course schedule below for due dates*

Assignment information

Quizzes

Quizzes will be set up (about every two weeks) on Carmen using the **Quizzes** link. They are questions requiring short answers or multiple-choice questions that are designed to assess your understanding of the course content after watching video lectures posted online. You are strongly encouraged to complete the quizzes soon after watching the associated lectures. The weekly module on Carmen will display the list of lectures and quizzes in order. Most of the quizzes will be due on **Fridays** by **11:59pm**.

Homework assignments

Homework will be assigned regularly (about every two weeks) throughout the semester using the **Assignments** link on Carmen. Assignments will consist of a mix of technical questions to assess students' understanding of the statistical models, and questions asking students to perform analyses of datasets. The grade for the analysis portion of each assignment will be based on both the accurateness and appropriateness of the analysis, as well as the clarity of the description of the analysis and results. Your report should include properly formatted computer output and graphs embedded in text for justification of your answer. For the report, you can use any word processing software of your choice (e.g., Word or LaTeX), but the file format for submission on Carmen should be PDF. You should submit your own individual work. In addition, computer code must be separately submitted as an appendix to each assignment. Your code should include comment statements that indicate what sections of the code correspond to the specific homework questions so that, if needed, the grader can read and check your code for its accuracy. Homework solutions will be posted on the course webpage. Due dates for homework assignments will be announced on Carmen.

Midterm

A take-home midterm exam will be given in the week of March 15-19. The midterm will be completed by each student individually. The exam will be available to you from **9am** on **March 15 (Monday)** using the **Assignments** link and you will need to submit your answers to the exam no later than **9am** on **March 22 (Monday)**. You are expected to work on the exam by yourself. While you can use any notes, textbook or other course materials, you cannot consult with other people during the exam. More details will be provided on the course website as we get closer to the exam dates.

Group project

Students will also complete projects in groups consisting of 4 to 5 members (depending on the enrollment size). The project will consist of selecting a data set (by week 4), performing an exploratory data analysis (EDA, by week 6), making a 5 page proposal with preliminary results

(by week 11), presenting through a recorded video (in week 16), and submitting a 10 page final report (by **noon on April 28**) using the **Assignments** page. The proposal should contain a detailed problem statement that includes questions of interest and a description of what methods will be used and how they will be used to answer questions of interest or solve the problem. More details will be provided on the course website.

Participation

There will be several activities requiring your participation for helping you navigate the course and stay connected with the material or building connections with the instructor and other students (e.g., posting introduction video, proposing datasets for project). These activities will be announced on Carmen for your participation. In addition, active participation in discussion forums with thoughtful questions and answers will be rewarded with extra points.

Late assignments

Typically no late homework assignments will be accepted. If you have documented reasons for missing work or needing extra time, please contact me as soon as possible. Where appropriate, due dates could be extended.

Faculty feedback and response time

I am providing the following list to give you an idea of my intended availability throughout the course. (Remember that you can call **614-688-HELP** at any time if you have a technical problem.)

Grading and feedback

For homework assignments, you can generally expect feedback within **7 days**.

E-mail

I will reply to e-mails generally within **24 hours on school days**.

Discussion board

I will check and reply to messages in the discussion boards every **24 hours on school days**.

Attendance, participation, and discussions

Student participation requirements

Because this is a distance-education course, your attendance is based on your online activity and participation. The following is a summary of everyone's expected participation:

- **Logging in**: **AT LEAST ONCE PER WEEK**
  Be sure you are logging in to the course in Carmen each week, including weeks with holidays or weeks with minimal online course activity. (During most weeks you will probably log in many times.) If you have a situation that might cause you to miss an entire week of class, discuss it with me *as soon as possible.*
- **Office hours and live sessions**: **OPTIONAL OR FLEXIBLE**
  All live, scheduled events for the course, including office hours, are optional. For live presentations, I will provide a recording that you can watch later. If you are required to discuss an assignment with me, please contact me at the beginning of the week if you need a time outside my scheduled office hours.
- **Participating in discussion forums**: **OPTIONAL**
  Participation in discussion forums is optional. However, you are strongly encouraged to post questions in discussion forums and to answer questions posted by other students. Active participation and substantive class discussion on the week's topics will be a great way to learn the material.

Discussion and communication guidelines

The following are my expectations for how we should communicate as a class. Above all, please remember to be respectful and thoughtful.

- **Writing style**: While there is no need to participate in class discussions as if you were writing a research paper, you should remember to write using good grammar, spelling, and punctuation. Informality (including an occasional emoticon) is fine for non-academic topics.
- **Tone and civility**: Let's maintain a supportive learning community where everyone feels safe and where people can disagree amicably. Remember that sarcasm doesn't always come across online.
- **Citing your sources**: When we have academic discussions, please cite your sources to back up what you say. (For the textbook or other course materials, list at least the title and page numbers. For online sources, include a link.)
- **Backing up your work**: Consider composing your academic posts in a word processor, where you can save your work, and then copying into the Carmen discussion.

Other course policies

Health and safety

The Ohio State University Wexner Medical Center's Cornavirus Outbreak site (https://wexnermedical.osu.edu/features/coronavirus) includes the latest information about COVID-19 as well as guidance for students, faculty and staff.

Guidelines and requirements for campus safety from the University's COVID-19 Transition Task Force can be found at https://safeandhealthy.osu.edu. They include the following:

- A daily health check to report body temperature and health status will be required for all faculty, staff and students each day they intend to be on Ohio State's campuses in the autumn.
- Face masks must be worn in indoor settings, including classrooms.
- Members of the campus community will be required to sign a pledge to affirm their understanding of what is needed to help fight the spread of the virus and their intention to do their part.
- Accountability measures will be in place for those who refuse to abide by required health and safety guidelines.

Potential disruptions to instruction

Since this course offering is online, prolonged student absences due to health reasons could be managed easily. Students will have access to material online if they are unable to attend class because of positive diagnosis, symptoms, or quarantine required following contact tracing.

If the instructor is unable to be present in person because of positive diagnosis, symptoms, or quarantine following contact tracing, an alternate instructor will be assigned to the course. Details will be given on the course website.

Student academic services

Student academic services offered on the OSU main campus http://advising.osu.edu/welcome.shtml.

Student support services

Student support services offered on the OSU main campus http://ssc.osu.edu.

Academic integrity policy

Policies for this online course

- **Quizzes and exam**: You must complete the midterm yourself, without any external help or communication. Quizzes are included as self-checks with only small points attached, and they are expected to be completed on your own.
- **Written assignments**: Your written assignments, including discussion posts, should be your own original work. In formal assignments, you should follow **[MLA/APA/Chicago]** style to cite the ideas and words of your research sources. You are encouraged to ask a trusted person to proofread your assignments before you turn them in--but no one else should revise or rewrite your work.
- **Reusing past work**: In general, you are prohibited in university courses from turning in work from a past class to your current class, even if you modify it. If you want to build on past research or revisit a topic you've explored in previous courses, please discuss the situation with me.
- **Falsifying research or results**: All research you will conduct in this course is intended to be a learning experience; you should never feel tempted to make your results look more successful than they were.
- **Collaboration and informal peer-review**: The course includes many opportunities for formal collaboration with your classmates. While study groups and discussions on the course topics are encouraged, remember that comparing answers on a quiz or assignment is not permitted. If you're unsure about a particular situation, please feel free just to ask ahead of time.

Ohio State's academic integrity policy

It is the responsibility of the Committee on Academic Misconduct to investigate or establish procedures for the investigation of all reported cases of student academic misconduct. The term "academic misconduct" includes all forms of student academic misconduct wherever committed; illustrated by, but not limited to, cases of plagiarism and dishonest practices in connection with examinations. Instructors shall report all instances of alleged academic misconduct to the committee (Faculty Rule 3335-5-487). For additional information, see the Code of Student Conduct http://studentlife.osu.edu/csc/.

Copyright disclaimer

The materials used in connection with this course may be subject to copyright protection and are only for the use of students officially enrolled in the course for the educational purposes associated with the course. Copyright law must be considered before copying, retaining, or disseminating materials outside of the course.

Statement on title IX

Title IX makes it clear that violence and harassment based on sex and gender are Civil Rights offenses subject to the same kinds of accountability and the same kinds of support applied to offenses against other protected categories (e.g., race). If you or someone you know has been sexually harassed or assaulted, you may find the appropriate resources at http://titleix.osu.edu or by contacting the Ohio State Title IX Coordinator at titleix@osu.edu

Accessibility accommodations for students with disabilities

The university strives to make all learning experiences as accessible as possible. In light of the current pandemic, students seeking to request COVID-related accommodations may do so through the university's request process, managed by Student Life Disability Services. If you anticipate or experience academic barriers based on your disability (including mental health, chronic, or temporary medical conditions), please let me know immediately so that we can privately discuss options. To establish reasonable accommodations, I may request that you register with Student Life Disability Services. After registration, make arrangements with me as soon as possible to discuss your accommodations so that they may be implemented in a timely fashion. SLDS contact information: slds@osu.edu; 614-292-3307; http://slds.osu.edu; 098 Baker Hall, 113 W. 12th Avenue.

Accessibility of course technology

This online course requires use of Carmen (Ohio State's learning management system) and other online communication and multimedia tools. If you need additional services to use these technologies, please request accommodations with your instructor.

- Carmen (Canvas) accessibility
- Streaming audio and video
- Synchronous course tools

Your mental health

As a student you may experience a range of issues that can cause barriers to learning, such as strained relationships, increased anxiety, alcohol/drug problems, feeling down, difficulty concentrating and/or lack of motivation. These mental health concerns or stressful events may lead to diminished academic performance or reduce a student's ability to participate in daily

activities. The Ohio State University offers services to assist you with addressing these and other concerns you may be experiencing. If you or someone you know are suffering from any of the aforementioned conditions, you can learn more about the broad range of confidential mental health services available on campus via the Office of Student Life's Counseling and Consultation Service (CCS) by visiting ccs.osu.edu or calling 614- 292-5766. CCS is located on the 4th Floor of the Younkin Success Center and 10th Floor of Lincoln Tower. You can reach an on call counselor when CCS is closed at 614-292-5766 and 24 hour emergency help is also available through the 24/7 National Suicide Prevention Hotline at 1-800-273- TALK or at suicidepreventionlifeline.org

Disclaimer

This syllabus should be taken as a fairly reliable guide for the course content. However, you cannot claim any rights from it and in particular we reserve the right to change due dates or the methods of grading and/or assessment if necessary. Any changes will be communicated to you through official course announcements.

Course schedule (tentative)

Week	Dates	Topics, Readings, Assignments, Deadlines
1	**1/11-1/15**	Introduction to Statistical Learning, Linear Regression
2	**1/18-1/22**	1/18(M): Martin Luther King Jr Day Classification: Logistic regression, Gaussian LDA Project: Dataset Proposal
3	**1/25-1/29**	Quadratic DA, Comparison of Methods, Evaluation Criteria Homework assignment 1 Due
4	**2/1-2/5**	Cross-Validation and Bootstrap Project: Dataset Selection
5	**2/8-2/12**	Linear Model Selection and Regularization: Subset Selection, Shrinkage Methods, Dimension Reduction Method Homework assignment 2 Due
6	**2/15-2/19**	Basis Expansion Approach: Splines, Smoothing Splines Project: EDA
7	**2/22-2/26**	2/23(T)-2/24(W): Instructional Break Local Regression, Generalized Additive Models Homework assignment 3 Due
8	**3/1-3/5**	Support Vector Machines and Maximal Margin Classifier
9	**3/8-3/12**	Kernels, Relationship to Logistic Regression Homework assignment 4 Due
10	**3/15-3/19**	Kernels for Nonlinear SVM, SVMs with More than Two Classes **Midterm**
11	**3/22-3/26**	Clustering Methods: k-Means, Hierarchical Clustering **Project: Proposal Due**
12	**3/29-4/2**	3/31(W) – 4/1(R): Instructional Break Tree-based method: Classification and Regression Trees
13	**4/5-4/9**	Bagging, Random Forest, Boosting Homework assignment 5 Due
14	**4/12-4/16**	PCA, Gaussian Mixture, EM
15	**4/19-4/21**	Neural Networks
16	**4/26-4/30**	**Project presentation video due by noon 4/26 (Mon)** **Project report due by noon 4/28 (Wed)**