# STAT 6500 Statistical Machine Learning

**Term:** Spring 2024
**Lectures:** MWF 3:00–3:55 PM (3 credit hours) in Cockins Hall 312

**Instructor:** Yoonkyung Lee
Office: 440H Cockins Hall
Office Hours: M 4:00–5:00 PM and F 11:00 AM–12:00 PM or by appointment
Email: yklee@stat.osu.edu or lee.2272@osu.edu

**Grader:** Zhizhen Zhao
Office Hours: by appointment only
Email: zhao.3053@osu.edu

**Course Website:** `https://carmen.osu.edu`

**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.

**Expected 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 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.

**Computing and Software:**
One of the goals of the course is to train students in the computing aspects of statistical machine learning and develop the skills to implement machine learning algorithms. Many homework assignments will have a computing and programming component (knowledge of software packages such as `Stata`, `SAS` or `JMP` will *not* be sufficient). There will be example codes provided, primarily written in the language `R`.

**Textbooks:**

- Required:
  James, Witten, Hastie, Tibshirani: *An Introduction to Statistical Learning with Applications in R*, 2nd edition. (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)

**Coursework:**

- **Homework:** Homework will be assigned regularly (about every two weeks) throughout the semester using the Assignments page 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. The assignments and assignment solutions will be posted on the course website. Late submissions will NOT be accepted (unless an alternative arrangement has been made with the instructor prior to the deadline for valid reasons).

- **Midterm:** A take-home midterm exam will be given. The midterm will be completed by each student individually.

- **Group Project:** Students will also complete projects in groups consisting of 2 to 3 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 8), making a 5 page proposal (by week 12), presenting (in week 15), and submitting a 10 page final report (by April 29). 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 given later.

- **Participation:** You are expected to attend all lectures. In addition to regular class participation, there will be several activities requiring your participation for building connections with other students or formulating potential projects (e.g., posting introduction video, proposing datasets for project). These activities will be announced in class and on Carmen.

**Grading:**
Grades will be assigned on the basis of the following components.

- Homework (35%)

- Take-home midterm exam (30%)

- Group project (30%)

- Participation (5%)

**Tentative Course Schedule:**
Homework due dates and midterm schedule in the table are tentative. Please refer to in-class announcements (also on Carmen) for official dates.

| Week | Dates | Topics, Assignments, Deadlines |
|---|---|---|
| 1 | 1/8–1/12 | Intro to Statistical Learning, Review of Linear Regression |
| 2 | 1/15–1/19 | 1/15(M): Martin Luther King Jr Day<br>Classification: Logistic regression, Gaussian LDA<br>Project: Dataset Proposal |
| 3 | 1/22–1/26 | Quadratic DA, Comparison of Methods, Evaluation Criteria<br>Homework assignment 1 due |
| 4 | 1/29–2/2 | Application*, Resampling methods: Cross-validation, Bootstrap<br>Project: Dataset Selection |
| 5 | 2/5–2/9 | Linear Model Selection and Regularization: Subset Selection, Shrinkage Methods, Dimension Reduction Method<br>Homework assignment 2 due |
| 6 | 2/12–2/16 | Linear Model Selection and Regularization, Basis Expansion Approach |
| 7 | 2/19–2/23 | Splines, Smoothing Splines<br>Homework assignment 3 due |
| 8 | 2/26–3/1 | Local Regression, Generalized Additive Models, Application*<br>Project: EDA |
| 9 | 3/4–3/8 | Support Vector Machines and Maximal Margin Classifier<br>Homework assignment 4 due |
| | 3/11–3/15 | Spring Break |
| 10 | 3/18–3/22 | Kernels for Nonlinear SVM, Application*<br>Homework assignment 5 due<br>Midterm assigned |
| 11 | 3/25–3/29 | Tree-based method: Classification and Regression Trees<br>Midterm due |
| 12 | 4/1–4/5 | Bagging, Random Forest, Boosting<br>Project: Proposal |
| 13 | 4/8–4/12 | Application*, PCA, Matrix Completion |
| 14 | 4/15–4/19 | Clustering Methods: $k$-Means, Hierarchical Clustering, Neural Networks |
| 15 | 4/22–4/26 | 4/22 (M): Last day of class<br>Neural Networks<br>Project presentation video due by 4/26 (F) |
| 16 | 4/29–5/3 | Project report due by 4/29 (M) |

## 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.

### Academic integrity policy

Although students are encouraged to work together on assignments, each student is expected to write and submit individual solutions to homework problems. The midterm is to be completed on your own without any external help or communication.

Academic misconduct will not be tolerated and will be dealt with procedurally in accordance with university 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/`.

## Accessibility accommodations for students with disabilities

The university strives to maintain a healthy and accessible environment to support student learning in and out of the classroom. If you anticipate or experience academic barriers based on your disability (including mental health, chronic, or temporary medical conditions), please let the instructor know immediately so that we can privately discuss options. To establish reasonable accommodations, we may request that you register with Student Life Disability Services. After registration, make arrangements with the instructor as soon as possible to discuss your accommodations so that they may be implemented in a timely fashion.
If you are isolating while waiting for a COVID-19 test result, please let me know immediately. Those testing positive for COVID-19 should refer to the Safe and Healthy Buckeyes site for resources. Beyond five days of the required COVID-19 isolation period, the instructor may rely on Student Life Disability Services to establish further reasonable accommodations. You can connect with them at slds@osu.edu; 614-292-3307; or slds.osu.edu.

### Religious accommodations

It is Ohio State's policy to reasonably accommodate the sincerely held religious beliefs and practices of all students. The policy permits a student to be absent for up to three days each academic semester for reasons of faith or religious or spiritual belief.

Students planning to use religious beliefs or practices accommodations for course requirements must inform the instructor in writing no later than 14 days after the course begins. The instructor is then responsible for scheduling an alternative time and date for the course requirement, which may be before or after the original time and date of the course requirement. These alternative accommodations will remain confidential. It is the student's responsibility to ensure that all course assignments are completed.