STAT 442: Statistical Learning and Data Mining

Professor: Anna Neufeld

Spring 2025

---

1 General Course Information

Lectures: Monday/Thursday, 2:35-3:50, Wachenheim 116.

Office Hours: Tentative! If students are unable to attend these times, I may adjust them based on availability. Homework assignments will generally be due on Monday at 11:59pm.

- Thursdays 11am-noon, Wachenheim 239.
- Mondays 1:15pm-2:30pm, Wachenheim 239.
- By appointment, at this link: https://calendar.app.google/tGQKsNqSETcPAmRv6.

Teaching assistant: Sarah Hartman, sah4@williams.edu.

TA session hours: Sunday, 7pm-9pm, room TBD.

Prerequisites: Stat 341 and Stat 346, or equivalent. Students are expected to be very comfortable with linear regression, programming in R, probability, and random variables.

Course description: We are surrounded by data, which continues to grow in size and complexity. Scientific progress often depends on extracting insights from data, such as developing predictive models, identifying relationships, drawing causal conclusions, or uncovering hidden structures. Many of these tasks can be tackled with familiar statistical tools like linear and logistic regression. However, as data becomes larger and more complex, these methods may not be the best option. Furthermore, advances in computing have enabled us to fit much more complex models. In response, recent decades have seen the rapid development of new algorithms designed for learning from diverse and complex datasets. In this course, we will explore a variety of these modern statistical learning algorithms. Beyond understanding how they work, we will develop the skills to compare, critically evaluate, and refine these methods to improve their effectiveness.

Learning objectives: At the end of this course, students will be able to:

- Understand a broad set of classical statistical learning methods and develop the skills to learn new algorithms as needed.
- Evaluate a task, research question, or dataset to determine the most appropriate models or algorithms and compare candidate methods to make an informed selection.

- Articulate the fundamental bias-variance tradeoff and explain its impact on algorithm performance.
- Critically assess a machine learning pipelineâĂŤfrom data collection and cleaning to statistical learning and downstream decisionsâĂŤwhile identifying potential limitations and harms.
- Digest and then communicate complex statistical concepts.

Technology Computing is an essential part of statistics and data analysis. In this course, we will be using R and RStudio. If you have access to a personal laptop that you will be using throughout the semester, it will be convenient to download the latest versions of R and RStudio. Otherwise, please check out the Library's laptop lending page for information about borrowing a laptop from OIT https://libguides.williams.edu/c.php?g=916778&p=8530683. R and RStudio should be already installed on all school computers.

2 Assessments

Here is a summary of the types of assessments you will find in this class, and how much they are worth for your grade.

2.1 Homework (15%)

There will be 7-8 homework assignments this semester. Problem sets will involve coding problems in R (conduct a simulation study to compare methods, implement a method, or analyze a dataset), theoretical or conceptual problems, and occasional writing questions (describe an idea or respond to a reading assignment).

Submit all homework as a polished PDF, preferably created in R Markdown. Please answer all questions in order in full sentences. All plots should have axis labels and titles. To avoid cluttering your main document, you may want to put your R code in an appendix. Review your PDF before submission to ensure it is well-formatted and resembles a formal report.

2.2 Teaching presentation (15%)

Each student will give one presentation during the semester. These will take place often, starting in Week 3. The idea is that the student will give a 20 minute "primer talk" at the start of lecture that introduces one of our topics for the course (often, but not always, this will mean introducing the basics of a statistical learning algorithm).

I could spend 12 weeks this semester teaching you the fundamentals of 12 different statistical learning algorithms. However, this would be a disservice to you: new statistical learning algorithms are developed constantly, and if you treat this class as a place to learn a fixed set of algorithms, you will not be prepared for the new algorithms that you will encounter in the real world. Thus, I want to give you practice with the skill of independently learning about a new algorithm from various resources and then processing and distilling the "key points" of this algorithm to be communicated to others. These teaching presentations are a chance for each of you to practice this skill once during the semester!

You will sign up for your day/topic during the first week of the semester. You should then set up a time to meet with me (using my individual office hours sign up link) 1.5-2 weeks before your presentation so that we can discuss resources and scope. You are welcome to also schedule a meeting

1-2 days before your presentation if you want feedback on your final plan. Slides are encouraged but not required.

Please take these seriously! When you are presenting, you will be graded on content and delivery. When you are in the audience, you should be engaged and ask questions! General engagement with these presentations will be reflected in your participation grade.

2.3 Midterm exam (25 %)

The week before spring break, we will have a midterm with an in-class component (on Thursday) and a take-home component (likely posted on Monday and due on Friday).

2.4 Final project (35%)

In groups of approximately two students, you are responsible for choosing a topic that we did not cover in class. This might involve a type of data we did not consider (e.g. network data, text data, image data, survival data, genomic data), a statistical concept that we did not cover in depth (e.g. conformal inference, semi-supervised learning, multiple testing, multitask learning, double descent), an algorithm that we did not cover (e.g. auto-encoders, transformers, graph clustering), or something else.

You will then prepare a thorough report where you introduce this topic and either apply it (if it is a concept), analyze it (if it is a type of data), and compare it to alternatives (if it is an algorithm). I expect that all projects will involve a substantial coding portion (a data analysis, implementation, or simulation study) as well as a substantial literature review portion (so that you can write about the topic in detail).

We don't have a final exam in this class, and so I expect these projects to be very substantial and well-executed. For example, if you are picking a "type of data", I expect that the project will go above and beyond a simple data analysis- I expect that you will really learn a lot about this type of data and the different strategies for analyzing it!

You will submit a proposal shortly after spring break where you will explain your idea. At this time, we will work together to make sure that your project has the appropriate amount of depth. You will then turn in a progress report in Week 10. You will give a final presentation during our last class meeting, and you will turn in a final report and a final reflection during reading period. While the majority of the grade will be a group-grade, a portion of the grade will reflect your individual contributions to the project.

2.5 Participation (10%)

A high participation grade requires consistent attendance and/or communication about absences. Beyond attendance, I will consider your engagement, including whether you ask questions during other students' "teaching presentations". A few class sessions will include discussions of readings. Active participation in these discussions is expected.

3 Calendar

Here is an approximate schedule of topics.

- Lecture 0: Welcome and intros; what is statistical learning?

- Lectures 1-3: Supervised learning: regression.
- Lectures 4-5: Supervised learning: classification.
- Lectures 6-10: Non-linear models and how to fit them; including splines, GAMS, neural networks, trees, and SVMs.
- Lecture 11: Midterm review.
- Lecture 12: Midterm
- Spring break.
- Lectures 13-16: Model selection; more on the bias-variance tradeoff, but also ensemble methods and interpretability.
- Lectures 17-19: Ethics, fairness, and social implications of ML.
- Lectures 20-23: Unsupervised learning; clustering and dimension reduction.
- Lecture 24: Final project presentations!

4 Calendar

Please see GLOW for the schedule of topics and for all due dates!

5 Course Policies

5.1 Honor Code

As an institution fundamentally concerned with the free exchange of ideas, Williams College has always depended on the academic integrity of each of its members. A student who enrolls at Williams thereby agrees to respect and acknowledge the research and ideas of others in his or her work and to abide by those regulations governing work stipulated by the instructor. Any student who breaks these regulations, misrepresents his or her own work, or collaborates in the misrepresentation of another's work has committed a serious violation of this agreement. See the description of Williams's honor code and system at https://sites.williams.edu/honor-system/.

The Williams Honor Code applies to all graded work in this class. Please see the sections on collaboration and AI for more details on how the honor code applies to assignments in this course. If you have any questions or uncertainties about what something means, or whether something is okay, please contact me: I am very happy to talk to you about this.

5.2 Collaboration

I encourage you to work with classmates, visit office hours, and visit TA sessions for help with all assignments in this class. However, your final written work must be your own. One way to be sure you are not violating the honor code is to refrain from typing up your final assignment until you are on your own and working independently. When you submit your assignments, please write down the names of any of your collaborators, write down which TA sessions / office hours / etc. you attended.

5.3 Use of Artificial Intelligence

Generative AI is an exciting tool. I hope that you will use AI as a tool to support your learning, but not use it to replace your learning.

Below, I will list three ways in which you are allowed to use Generative AI and three ways in which you are not allowed to use generative AI in this course.

Allowed:

- Ask for help debugging code, understanding code error messages, or making code faster.
- Ask for code to carry out a specific subtask, such as: "what function can I use to apply KNN to a dataset in R" or "how do I add a legend to a ggplot in R"?
- Ask for help making a paragraph of a written report more concise, or ask for grammar/proof-reading help with written reports. In other words, ask for help editing a draft of writing.

Not allowed:

- Ask for code to carry out an entire assignment question, such as "can I have R code that uses cross validation to try out KNN for values of K ranging from-10 and plots the error as a function of K".
- Ask it to generate a first draft of any written material or code. First drafts should always be written yourself!
- Ask it to summarize your topic for your teaching presentation. For your teaching presentation, I really want you to practice the skill of synthesizing complex topics and putting them into your own words.

While the list above is not exhaustive, I hope it gives you an idea of how AI may be used in this class. If you are ever in doubt about how AI can be used, please just ask!

As long as you **document exactly how/why you used AI** in any place that you used it, I will not consider it an honor code violation: the worst case scenario is that if it seems that you used AI for an entire problem (rather than a subtask), you may receive a 0 for that problem. And there will be no points deductions for appropriate uses of AI. On the other hand, if you use AI inappropriately and **do not document its use**, I will consider that an honor code violation. Therefore, to be safe, **any time that you use AI in this course, you should write down exactly how/why you used it**.

Finally, remember that ChatGPT is not actually that smart, and that you are 100% responsible for your final answers, which might be incomplete or wrong if you rely fully on ChatGPT!

5.4 Extensions or late work

Every student has three late days that you may use throughout the semester on homework assignments. If you are using a late day on a homework, just note this in the comments on your GLOW submission. These late days are meant to cover cases where minor illness, coursework for other classes, other campus commitments, or unexpected difficulty on the assignment gets in the way of finishing the assignment on time. For major illnesses or other personal circumstances that may require additional late days, please email me.

5.5 Inclusion and classroom climate

It is my intent that students from all diverse backgrounds and perspectives be well served by this course, that students' learning needs be addressed both in and out of class, and that the diversity that students bring to this class be viewed as a resource, strength and benefit. I am dedicated to presenting materials and activities that are respectful of diversity in gender, gender identity, gender expression, sexual orientation, age, socioeconomic status, ethnicity, beliefs, race, culture and educational background, and other visible and non-visible categories. I welcome and appreciate your suggestions. Please let me know (by email or by anonymous survey) if you see ways to improve the effectiveness of the course for you personally or for other students or student groups. arrangements for you.

5.6 Names and Pronouns

In this class, we use the name and gender pronouns that individuals ask us to use as a sign of mutual respect. I will gladly honor your request to address you by an another name or gender pronoun. Please inform me of your preferences early in the semester so I can update my records accordingly.

6 Resources

6.1 GLOW discussion board

If you have a clarifying question about something that we covered in class, or something on an assignment or a solution key, please ask this on the GLOW discussion board. This board will be monitored by both the instructor and the TAs. This way, other students can see the question and the response, and may benefit from it.

6.2 Email

If you have an individual matter that needs attention, please email me at acn2@williams.edu. Here are some matters to keep in mind when sending an email:

- Please include "Stat 442" in the subject line.
- If you are asking a simple clarification question about the course, I may respond "please ask this on the discussion board instead".
- I will do my best to respond to emails within 24 hours on the weekdays and 48 hours on the weekend. If I have not responded to an email within this time frame, please send me a polite reminder: things can certainly get lost in my inbox!
- If you are asking for an individual meeting, please first try to schedule via my calendar-link on GLOW. This will reduce the need for back-and-forths!

6.3 General support and feedback

Your health and well-being are the most important things to us. Please let me know if you encounter any issues and difficulties in engaging in the course activities. You can reach out over email, post on the GLOW discussion board, or use the anonymous course feedback form that is posted on GLOW.

6.4 Course help sessions

Please attend any office hours or TA sessions! You are welcome to drop in with any questions related to the course or the study of statistics in general, or just stick around and work on the homework during the help session. I especially welcome you to come to my office hours to introduce yourself in the first week of the semester. If my scheduled office hours and appointment times truly do not work for your schedule, please email me!

6.5 Accommodations

We are committed to supporting the learning of all students in our classes. Students with disabilities or disabling conditions who experience barriers in this course are encouraged to contact me to discuss options for access and full course participation. The Office of Accessible Education is also available to facilitate the removal of barriers and to ensure access and reasonable accommodations. Students with documented disabilities or disabling conditions of any kind who may need accommodations for this course or who have questions about appropriate resources are encouraged to contact the Office of Accessible Education at oaestaff@williams.edu.