STAT 0218: Statistical Learning

Fall 2025

Instructor:	Christian Stratton	**Time:**	TR 11:15 – 12:30
Email:	cstratton@middlebury.edu	**Place:**	75 Shannon St 203
Office:	Warner 203	**Office hours:**	TBD Also by appointment

Course description: This course is an introduction to modern statistical, machine learning, and computational methods to analyze large and complex data sets that arise in a variety of fields, from biology to economics to astrophysics. The theoretical underpinnings of the most important modeling and predictive methods will be covered, including classification, clustering, and tree-based methods. Student work will involve implementation of these concepts using open-source computational tools.

Correspondence: My goal is to maximize my availability for help and discussion throughout the semester. Office hours will be determined via poll during the first week of class, but please feel free to contact me via email at anytime. Additionally, I am happy to meet outside of office hours by appointment.

Meeting format: Class will generally be used to learn new statistical concepts through a mixture of lecture and in-class activities. Most class periods will feature a short lecture introducing a new concept, followed by an in-class guided activity to be worked on in small groups. You will need to have access to a laptop during class See more details below.

Learning objectives: Through this course, students will:

- Identify the goal of statistical learning and recognize the difference between inference and prediction
- Recognize the difference between supervised and unsupervised learning techniques
- Understand how to develop loss functions for model assessment
- Develop familiarity with a number of common statistical learning procedures, including KNN, LDA, tree-based methods, dimension reduction techniques, and clustering methods.

Textbook and materials: There is nothing that need be purchased for this class; all materials are free.

- The website for this course is on Middlebury Canvas. Please check Canvas often for assignments, deadlines, resources, and announcements.
- Students must have access to a laptop with the statistical computing language R, which can be downloaded for free at https://cran.rstudio.com/. Additionally, I recommend using RStudio as an integrated development environment (IDE) for interfacing with R. RStudio may be downloaded for free at https://posit.co/download/rstudio-desktop/.
  - Laptops with R/RStudio pre-installed are available to borrow from the Davis Family Library, which are a good option for those without access to a laptop or those experiencing short-term issues with your laptop. Please talk to me or the front desk of the Davis Library for more info.
- We will use a free online textbook: *An Introduction to Statistical Learning* by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. This book may be downloaded at https://www.statlearning.com/.

Academic integrity: You are bound by Middlebury College's honor code, including its policies on plagiarism and cheating. Violation of these rules is ground for failure. To avoid charges of plagiarism, cite all the sources used to complete your assignments/homework, including any peers with whom you collaborated. I encourage you to seek help in understanding the concepts and problems in your assignments from various sources, including peers, instructors, peer tutors, class notes, textbooks, and online sources.

Use of LLM and generative AI: Large language models (LLM) and generative AI, such as ChatGPT, are powerful tools enabled by statistics and data science techniques that may be used to enhance your learning of statistics and coding languages. As such, the use of large language models (LLM) and generative AI, such as ChatGPT, is permitted in this class and may be used on all assignments, unless explicitly prohibited by the assignment. However, **you may not copy responses verbatim from these tools, nor may you use these tools to generate complete responses or assignments.** Additionally, if content from generative AI is used on an assignment, **you must provide appropriate citation.** To clarify this policy, examples of acceptable and unacceptable prompts for ChatGPT are provided below.

Acceptable:

- What is the intuition behind k-nearest neighbors regression?
- What do principle components represent?
- The following code keeps giving me an error: ...

Unacceptable:

- Write a KNN algorithm to predict the uploaded data.
- Answer the following question: *copy-paste from assignment*

Disclaimer: I am compelled to note that while generative AI can be a powerful tool, it is not infallible. Consider the exchange provided at the end of the syllabus, conducted on ChatGPT 4o mini on 2024/09/01. It is possible that generative AI will provide you with incorrect information, and it is your responsibility to use generative AI critically. "ChatGPT said so," is not sufficient justification for an answer, and I am unlikely to be sympathetic to such comments on assignments.

Late policy: Consistent engagement with the course material is essential for your learning and academic growth. However, I understand that unforeseen circumstances may occasionally arise:

- When you become aware that you won't be able to make a deadline, please notify me and inform me of what day in the next week you anticipate completion of the assignment. You do not need to disclose why you are missing the deadline. So long as you communicate to me **before** the deadline, no late penalty will be applied.
- **If you do not communicate with me before the deadline, late submissions will receive no credit.**

Course assessment: Your grade will be determined by in-class activities, readiness assessments, statistical reports, and exams. Each category is loosely defined as follows:

10%	**Activities**	Most class days will feature an in-class activity that demonstrates the day's concepts; the activity must be completed by the next class day. **Graded on completion**.
10%	**Readiness assessment**	Most class periods will begin with a short assessment based on the assigned reading for the day. These assessments are designed to be very easy if you completed the reading. **Graded on accuracy**.
40%	**Exams**	There will be two exams in this class: the midterm and the final. Both exams will be completed out of class, and **on your own**.
40%	**Statistical reports**	Throughout the semester, you will be asked to write approximately three statistical reports. Full details will be provided as each report arises.

Letter grades: Letter grades will be assigned according to the following scale. Note that I may adjust thresholds at the end of the semester, but they will only ever be adjusted *down*.

F	**D**	**C-**	**C**	**C+**	**B-**	**B**	**B+**	**A-**	**A**
$[0, 60)$	$[60, 70)$	$[70, 74)$	$[74, 77)$	$[77, 80)$	$[80, 84)$	$[84, 87)$	$[87, 90)$	$[90, 94)$	$[94, 100)$

Accommodations for disabilities: Students who have Letters of Accommodation in this class are encouraged to contact me early in the semester to ensure that such accommodations are implemented in a timely fashion. For those without Letters of Accommodation, assistance is available to eligible students through the Disability Resource Center (DRC).

Disclaimer: This syllabus is a dynamic document and may change throughout the semester as a result of our conversations and decisions as a class.

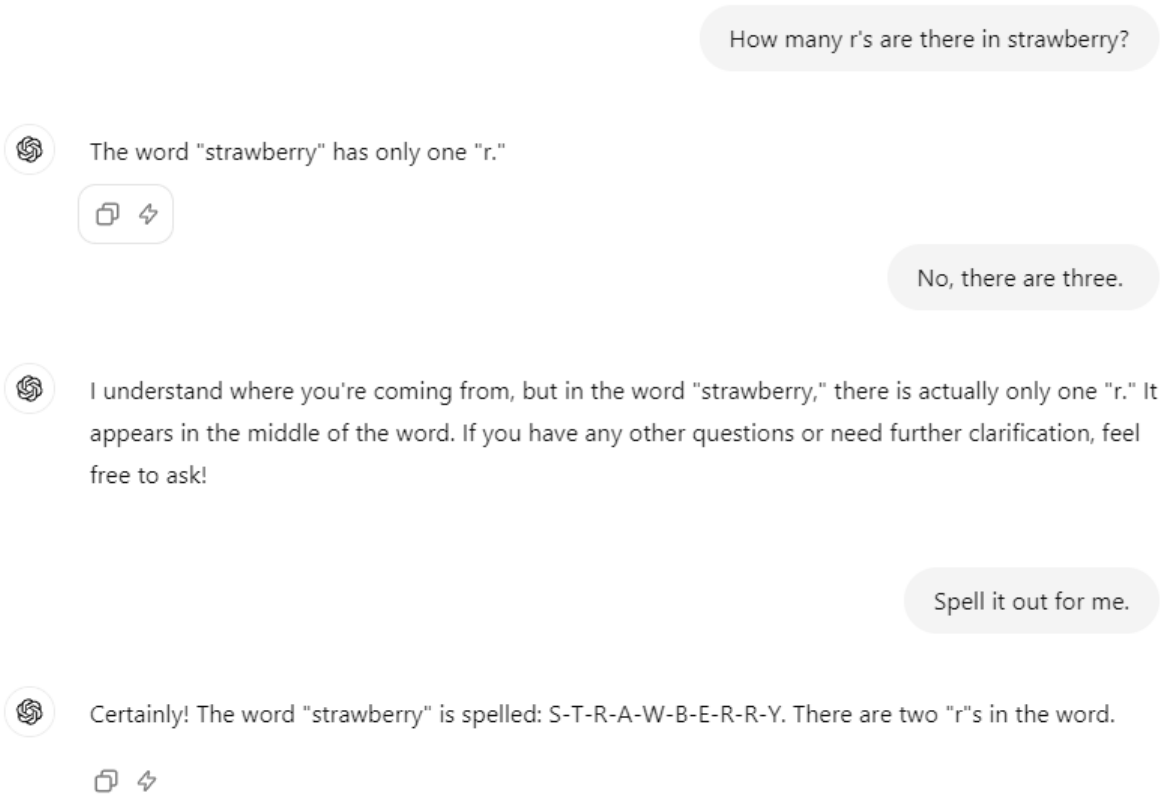

Figure 1: Prompt provided to ChatGPT 4o mini on 2024/09/01.

Tuesday	Thursday
Sep 9th **1** R/RStudio What is statistical learning	11th **2** **Supervised learning** Intro to KNN
16th **3** KNN cont.	18th **4** KNN cont.
23rd **5** Model assessment in supervised settings	25th **6** Model assessment cont.
30th **7** Intro to LDA	Oct 2nd **8** LDA cont.
7th **9** LDA cont.	9th **10** Intro to tree-based methods
14th **11** Tree-based methods cont.	16th **12** Tree-based methods cont.
21st **13** Tree-based methods cont.	23rd **14** **Unsupervised learning** Intro to dimension reduction and PCA
28th **15** PCA cont.	30th **16** PCA cont.
Nov 4th **17** K-means clustering	6th **18** K-means clustering cont.
11th **19** Alternatives to K-means clustering	13th **20** Hierarchical clustering
18th **21** Hierarchical clustering cont.	20th **22** Model assessment in unsupervised settings
25th *Thanksgiving break*	27th *Thanksgiving break*
Dec 2nd **23** Special topics: Deep learning	4th **24** Special topics: Deep learning
9th Reading days	11th Finals week
16th Finals week	18th