Syllabus

Last Updated: August 23, 2024

Course Objectives

After this course, students should be able to

- Understand the formulation of several popular regression, classification and unsupervised machine learning models
- Properly use and tune parameters for these machine learning algorithms
- Understand some of the fundamental concepts in statistical learning, such as the bias-variance trade-off, cross-validation, statistical simulation, nonparametric estimations and optimization
- Use `R` to perform data cleaning and model fitting, evaluation and be able to interpret the results
- Use `RMarkdown` to organize your report with proper visualization of the data and results
- Be able to utilize modern AI tools to assist your analysis and gain learning experience

Course Content

Tentative subjects include:

- Supervised Learning:
  - Linear models and penalization
  - Discriminant analysis, Naive Bayes, logistic regression
  - K nearest neighbor, classification and regression trees, random forest, kernel regression
  - Support vector machine
- Unsupervised Learning:
  - PCA, K-mean and hierarchical clustering, self-organizing maps, spectral clustering
- Concepts:
  - Bias-variance trade-off
  - Variable selection
  - Cross-validation
  - Bootstrap
  - Numerical optimization
- Modeling problems:
  - Personalized Medicine
  - Imaging data

Textbooks

- Supplemental: **SMLR** - Statistical Learning and Machine Learning with R (https://teazrq.github.io/SMLR/). This is my ongoing project that will contain most of the course material.
- Supplemental: **ISL** - An Introduction to Statistical Learning with Applications in R (https://www.statlearning.com/) by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani
- Supplemental: **ESL** - The Elements of Statistical Learning: Data Mining, Inference, and Prediction (https://web.stanford.edu/~hastie/ElemStatLearn/) by Trevor Hastie, Robert Tibshirani and Jerome Friedman. This is a more advanced textbook.

Prerequisites

A course that covers linear regression and uses `R`, such as STAT 420/425. **Good knowledge of probability and statistics (STAT400) and preliminaries of linear algebra (MATH 415) are also assumed**.

Software

`R` and `RStudio` are required software for this course. `R` is a freely available language and environment for statistical computing and graphics. `RStudio` is a free and open-source integrated development environment for `R`. Alternatively you can use `Visual Studio Code` (`VS Code`) as your programming environment, which has more modern programming assisting tools. You must have access to a computer where you are able to install the most up-to-date versions of `R` and `RStudio`, as well as install `R` packages. We will have an R tutorial during the first week.

Homework

There will usually be one homework assignment each week, with the total around 10. The total number may vary depending on the course progression. **The lowest score can be dropped**. All homework assignments must be submitted in `.pdf` format to Gradescope (https://www.gradescope.com/). A PDF report can be created using `R Markdown`, which is a feature in `RStudio` (or equivalently in `VS Code`). During the first week, we will provide a detailed guide for using `R Markdown` in either environment.

Late Policy:

- [***Regular Deadline***] All homework assignments are due at 11:59PM CT Thursday night. Please note that 1 second after the deadline will be treated as late (this will be determined by the `Gradescope` system). So do not wait until the last second.
- [***Late Submission***] You can submit your assignment as late as four days (Mon, 11:59PM) after the deadline. However, for each day of delay **you will lose 5% of the total score**. The late submission

policy **does not apply to the final project**.

- [***Unaccepted***] Anything after the late submission deadline will not be accepted. You will receive 0 score for that homework.

Grading Policy:

- [***Correctness***] 80% of the total score. This includes, but not limited to, providing details of your modeling approach, such as tuning parameters and explaining your decision process.
- [***Interpretation of Results***] 10% of the total score. You should provide the motivation, explain your approach, **summarize and interpret the results**.
- [***R code/output and Organization***] 10% of the total score. Your `R` code should be clean and easy to follow with necessary comments. Figures/Tables should be properly sized/simplified/colored and their labels should be easy to read. Keep in mind that you should submit a report, but not pieces of `R` code that only obtains the numerical results. I encourage you to watch this video (https://mediaspace.illinois.edu/media/t/1_8lrq866x) from the previous semester that discusses related issue.
- [***Gradescope***] You should submit the **compiled report file** (**`.PDF`**) to **Gradescope (https://www.gradescope.com/)**. Make all your `R` code chunks visible for grading.
- [***AI Tools***] In some homework assignments, you are required to report the prompt and results when using advanced AI tools. Pay attention to the specific requirements in the description.

Using AI Tools

As we explore the fascinating realm of programming and machine learning, I encourage you to utilize AI tools in your homework assignments. These cutting-edge technologies can provide you with insights, automate routine tasks, and enhance your learning experience. However, I must stress the importance of understanding both the capabilities and limitations of these tools. While they can be powerful aids, relying solely on them without a deep comprehension of the underlying principles can lead to misconceptions and errors. Be mindful of biases in data and the ethical implications of using AI in various contexts. Embrace these tools as an extension of your skills, but never substitute them for critical thinking and rigorous understanding. I believe that the responsible use of AI tools can greatly enrich your learning.

The above paragraph was generated by GPT-4. The university also has a discussion about ChatGPT (https://citl.illinois.edu/citl-101/instructional-spaces-technologies/teaching-with-technology/chatgpt). If you have questions or concerns, please let me know or talk to an expert.

Final Project

Final report is due 11:59 PM, Dec 12th. You have two options to complete the final project. You can complete the final project with a team. Each team can have **up to three members**. For more details, please see the project page (project.html).

Gradings

Type	Precentage
Homework	60%
Quizzes	10%

Type	Precentage
Final Project	30%

Letter grades

A+	A	A-	B+	B	B-	C+	C	C-	D+	D	D-
TBD	93%	90%	87%	83%	80%	77%	73%	70%	67%	63%	60%

Academic Integrity

The official University of Illinois policy related to academic integrity can be found in Article 1, Part 4 of the Student Code. Section 1-402 (https://studentcode.illinois.edu/article1/part4/1-402/) in particular outlines behavior which is considered an infraction of academic integrity. These sections of the Student Code will be upheld in this course. Any violations will be dealt with in a swift, fair and strict manner. Homework assignments are meant to be learning experiences. You may discuss the exercises with other students, but you **must write up the solutions on your own**. In short, do not cheat, it is not worth the risk. You are more likely to get caught than you believe. If you think you may be operating in a gray area, you most likely are.

Changes

The instructor reserves the right to make any changes he considers academically advisable. Such changes, if any, will be announced on this website and through email. Please note that it is your responsibility to keep track of the proceedings.