Grad 5100: Fundamentals of Data Science

Introduction

Grad 5100 is a foundational course in the MS in Data Science Program. It is designed to provide the essential background in programming, statistics, linear algebra, and multivariate calculus that the core and elective courses in the program rely on. The course will run intensively during the first few weeks of the semester and then drop back to a regular weekly schedule.

References

We will rely on the following materials.

- The anaconda machine learning environment for python and associated libraries
- The R language, Rstudio IDE, and associated libraries
- The VScode IDE

There is no formal textbook for the course. Some useful references include the following, all of which are either open source or available through the UConn Library.

- An Introduction to Statistical Learning by James, *et. al.* Note that the original version of this book uses the R language, but a new edition available in Summer 2023 uses Python.
- R for Data Science by Wickham and Grolemund.
- Python for Data Science, 3E by Wes McKinney.
- *Practical Statistics for Data Scientists* by Bruce, Bruce, and Gedeck. Note that this is available for free to UConn students through the UConn library's subscription to the O'Reilly Learning Platform.
- *Statistical Practice for Data Science* by Bar, Ravishankar, and Asha (draft).
- Mathematics for Machine Learning (draft) by Teitelbaum.
- *Fluent Python* by Ramalho. Available to UConn students through the UConn Library's subscription to the O'Reilly Learning Platform.

More advanced technical references include:

- The Elements of Statistical Learning by Hastie, *et. al.*
- Pattern Recognition and Machine Learning by Bishop

CampusWire

We will use the campuswire Q&A site for class discussions and question asking/answering. Campuswire is like a private version of stackoverflow. Make sure to register for the site.

Assessment

Course grades will be based on:

- biweekly homework assignments (60 %)
- a final exam (40 %)

Disclaimer

The instructor reserves the right to modify or adapt this syllabus to account for disruption due to COVID-19 or other unexpected circumstances.

University Policies

Students with disabilities should work with the Center for Students with Disabilities to request academic accommodations. The CSD is located in Wilbur Cross, Room 204 and can be reached at (860)-486-2020 or at csd@uconn.edu.

Students are bound by the university's policies on academic misconduct. Academic misconduct is dishonest or unethical behavior that includes but is not limited to misrepresenting mastery in an academic area (e.g. cheating), failing to properly credit information, research, or ideas to their rightful originators or representing such information, research, or ideas as your own (e.g. plagiarism).

Students, faculty, and staff are bound by the university's policy against discrimination, harassment, and related interpersonal violence.

Course Outline

1. Setting up a data science working environment
2. Probability and Statistics: the normal distribution
3. Working with Data in R and Python
4. Linear Algebra: vectors, matrices, the dot product
5. Partial derivatives and the gradient; matrix calculus
6. Slicing and dicing data in R and Python
7. Statistical Models
8. Hypothesis Testing
9. Data Structures and Object Oriented Programming
10. Visualization tools in R and Python
11. Version Control
12. Databases
13. Additional topics as time permits