COS 598 - Statistics Foundation of Data Science

Salimeh Yasaei Sekeh

salimeh.yasaei@maine.edu

University of Maine, Fall 2020

Instructor:

Salimeh Yasaei Sekeh - Tue/Thu 2:00-3:50pm, Boardman Hall, Rm 136

Office Hours:

Thu 4:00-5:30pm in Boardman Hall, second floor, Rm 247.

Required textbook: None.

Recommended texts:

- Carlos Fernandez-Granda, Probability and Statistics for Data Science, Center for Data Science in NYU, 2017, available online.
- Avrim Blum, John Hopcroft, and Ravi Kannan, Foundations of Data Science, Cambridge University Press, March 2020, available online.
- James D. Miller, STATISTICS FOR DATA SCIENCE,Packt Publishing, 2019.
- Davy Cielen, Arno D. B. Meysman, and Mohamed Ali, Introducing Data Science, Manning Publication, 2016, available online.

The pdf format of textbooks will be uploaded to Blackboard.

Additional references/reading:

- Michael J. Evans and Jeffrey S. Rosenthal, Probability and Statistics The Science of Uncertainty, University of Toronto, 2009.
- Joseph C. Watkins, An Introduction to the Science of Statistics: From Theory to Implementation, Preliminary Edition.

More textbooks and additional readings will be uploaded to Blackboard.

Grading:

- Homework: 40%, submission via blackboard. Lowest 1 dropped.
- Midterm exam: 30%, TBA

- Final exam: 30%, TBA
- Extra credit: 5-10% to students who answer questions in Blackboard and significantly enhance the course experience through their contributions.

Details:

- **Homeworks:**
  Homeworks will be assigned bi-weekly. Applications will be developed through programming exercises, including face recognition, spam filtering, handwritten digit recognition, image compression, and image segmentation. Python is the supported languages in the course. You must complete your programming assignments in this language.
- **Exams:**
  You may use two cheat sheets (front and back), and no other materials are allowed. Please notify us the first week of class if you have a conflict.
- **Final Project:**
  There will be a final project. You will be asked to form groups of maximum 2 people. The project must explore a methodology or application not covered in the lectures. You will be asked to select a paper on a methodology not covered in class, and implement the method. You will grade each others projects using grade's features.
- **Blackboard:**
  Most questions you have about the course, both logistical and technical, should be posted to Blackboard. Questions about how to solve homework problems are encouraged, but responses should provide hints as opposed to detailed answers. You may indicate that your questions is for instructors only your question is of a sensitive nature or may disclose solutions to the class.
- **Standards of Conduct:**
  Please,
  Arrive to class on time. If you must enter a class after lecture has clearly begun, please do so quietly. Focus on class material during class time. Sleeping, doing work for another class, checking email, and exploring the internet are unacceptable and can be disruptive. Only use personal electronic devices during class for viewing or taking notes. If you elect to use a laptop during class, please type very quietly. Few things are more annoying than sitting next to someone who is pounding on their keyboard during a lecture. Refrain from eating during class. Avoid audible and visible signs of restlessness. These are both rude and disruptive to the rest of the class. Don't pack your bags to leave until the instructor has dismissed class. Thank you.
- **Collaboration on homeworks:**
  Each student will prepare the final write-up/coding of his or her homework solutions without reference to any other person or source, aside from the student's own notes or scrap work. Students may consult classmates for the purpose of brainstorming, but not for obtaining the details of solutions. Under no circumstances may you copy solutions or code from a classmate or other source.

Campus Policies:

- **Academic Honesty Statement:**
  Academic honesty is very important. It is dishonest to cheat on exams, to copy term papers, to submit papers written by another person, to fake experimental results, or to copy or reword parts of books or articles into your own papers without appropriately citing the source. Students committing or aiding in any of these violations may be given failing grades for an assignment or for an entire course, at the discretion of the instructor. In addition to any academic action taken by an instructor, these violations are also subject to action under the University of Maine Student Conduct Code. The maximum possible sanction under the student conduct code is dismissal from the University.

- **Students with Disabilities:**
  If you have a disability for which you may be requesting an accommodation, please contact Student Accessibility Services, 121 East Annex, 581.2319, as early as possible in the term. Students who have already been approved for accommodations by SAS and have a current accommodation letter should meet with me privately during the first two weeks of class. All discussions will remain confidential.

- **Equal opportunity:**
  The Faculty of the COS are committed to a policy of equal opportunity for all persons and do not discriminate on the basis of race, color, national origin, age, marital status, sex, sexual orientation, gender identity, gender expression, disability, religion, height, weight, or veteran status. Please feel free to contact your instructor with any problem, concern, or suggestion. We ask that all students treat each other with respect.

1 Tentative Course Syllabus

Topics to be covered:

1.1 Introduction

- What is Big Data?
- Difference between Big Data and Smart Data
- What is Data Science?

1.2 Basic Probability Theory

- Concepts of Events
- Probability of an Event
- Probability Rules
- Independence

1.3 Bayesian Probability

- Prior Probability
- Posterior Probability
- Bayes Rule
- Applications in Data Science (Classification)

1.4 Random Variable

- Random Variable
- Discrete and Continuous PDFs
- Cumulative density function
- Multivariate Random Variable
- Joint and Conditional PDFs
- Independence
- Common Discrete and Continuous Distributions

1.5 Expectation and Variance

- Concept of Expectation and Variance
- Data-based Estimators (Sample mean and Sample Covariance)
- Generating Multivariate Random Variables

1.6 Linear Regression

- Linear Models
- Mean Square Error (MSE)
- Likelihood Function
- Parameter Estimation (MLE)
- Over-fitting/Under-fitting
- Application in Data Science - Global Warming

1.7 Cross Validation

1.8 Supervised and Unsupervised Learning

- Labeled Data and Unlabeled Data
- Clustering

1.9 Gaussian Distribution

- PDF of Gaussian Distribution
- Standard Gaussian Distribution
- Visualization
- Sample Mean and Covariance Matrix - MLE Estimation
- Application - Modeling Noisy Data

1.10 Hypothesis Testing

- The hypothesis-testing framework
- Parametric Testing
- Non-Parametric Testing
- Multiple Testing

1.11 Time-series/Sequential Data Science

- What is Random Process?
- Independent identically-distributed sequences
- Gaussian Process
- Random Walks
- Application: Time-dependent Prediction

1.12 Processing and Summarizing Data

- Real-world Big Data Processing
- Practical Examples
- Practical Statistical Analysis