DSE 501 - Statistical Foundation of Data Science

Salimeh Yasaei Sekeh

salimeh.yasaei@maine.edu

University of Maine, Fall Semesters

Instructor: Salimeh Yasaei Sekeh

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.
- Dirk P. Kroese, Zdravko Botev, Thomas Taimre, Radislav Vaisman, Data Science and Machine Learning: Mathematical and Statistical Methods, CRC Press.
- James D. Miller, Statistics for Data science, Packt Publishing, 2019.
- Steven Skiena, The Data Science Design Manual, 2017
- 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 Brightspace.

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 Brightspace.

Course Prerequisites: College level statistics course.

Grading:

- Homework: 40%
- Midterm exam: 30%

- Final exam: 30%
- Extra credit: 5-10% to students who answer questions in Brightspace and significantly enhance the course experience through their contributions.
- The grading scale for the final mark is as follows:

Letter Grades	**Numerical Range**
A	94-100
A-	90-94
B+	87-90
B	84-87
B-	80-84
C+	77-80
C	74-77
C-	70-74
D+	67-70
D	64-67
D-	60-64
F	0-60

Course Schedule:

The table (next page) provides the initial distribution of topics discussed over the weeks in the semester. This schedule is tentative and subject to change during the semester at the instruction discretion. All changes will be announced in class or on the course website. Students are responsible for making sure they are informed about announcements.

Details:

- **Homeworks:**
  Homeworks will be assigned bi-weekly. The problems include statistical and probability concepts along with simple real-world problem solving.
- **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.
- **Brightspace:**
  Most questions you have about the course, both logistical and technical, should be posted to Brightspace. 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

Week	Class Tue/Thu	Materials
1	01/09 03/9	Syllabus, Introduction on Data Science, Real world problems Basic Probability
2	08/09 10/09	Joint and Conditional Probability Random Variable, Continuous/Discrete RVs
3	15/09 17/09 18/09	Probability Mass/Density Functions Generating Random Variables, Multivariate Random Variables Homework 1 Due Date
4	24/09 18/02	Independence Expectation, Conditional Expectation
5	29/09 01/10 02/10	Variance, Correlation Common Distributions Homework 2 Due Date
6	06/10 08/10	Gaussian Distribution Maximum Likelihood Estimator
7	13/10 15/10 16/10	Histogram, Sample Mean and Variance Sample Covariance Homework 3 Due Date
8	20/10 22/10	Independent identically-distributed sampling Mean Square Error
9	27/10 29/10	Midterm Exam Hypothesis Testing
10	03/11 05/11	Error Types I & II, Confidence Interval Confidence on Mean and Variance
11	10/11 12/11 13/11	Decision Tree, Ensemble Method Bagging, Random Forest Homework 4 Due Date
12	17/11 19/11	Missing Values, Pruning Time-series/Sequential Learning, Random Process
13	24/11 26/11 30/11	Gaussian Process, Markov Chains Thanksgiving Homework 5 Due Date
14	01/12 03/12	Bayesian Decision Naive Bayes
15	08/12 10/12 11/12	Random Walks and Markov Chains Markov Chain Mont Carlo, Application Homework 6 Due Date

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 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 Accessibility Services Statement:**
  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.

- **Course Schedule Disclaimer (Disruption Clause):**
  In the event of an extended disruption of normal classroom activities, the format for this course may be modified to enable its completion within its programmed time frame. In that event, you will be provided an addendum to the syllabus that will supersede this version.
  https://umaine.edu/citl/teaching-resources-2/required-syllabus-information/#Schedule

- **UMaine Student Code of Conduct:**
  All students are expected to conform to the UMaine Student Code of Conduct.
  https://www.maine.edu/board-of-trustees/policy-manual/section-501/

- **Observance of Religious Holidays/Events:**
  The University of Maine recognizes that when students are observing significant religious holidays, some may be unable to attend classes or labs, study, take tests, or work on other assignments. If they provide adequate notice (at least one week and longer if at all possible), these students are allowed to make up course requirements as long as this effort does not create an unreasonable burden upon the instructor, department or University. At the discretion of the instructor, such coursework could be due before or after the examination or assignment. No adverse or prejudicial effects shall result to a student's grade for the examination, study, or course requirement on the day of religious observance. The student shall not be marked absent from the class due to observing a significant religious holiday. In the case of an internship or clinical, students should refer to the applicable policy in place by the employer or site.
  https://umaine.edu/citl/teaching-resources-2/required-syllabus-information/#Observance

- **Sexual Discrimination Reporting:**
  https://umaine.edu/citl/teaching-resources-2/required-syllabus-information/#Reporting_Long

- **COVID-19 Syllabus Statement:**
  Please read the University of Maine COVID-19 Syllabus Statement available on CITL website (link below):
  https://umaine.edu/citl/2020/08/17/suggested-syllabus-language-for-covid19-is-available/

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 Importing, Summarizing, and Visualizing Data

- Structuring Features According to Type
- Summary Tables
- Summary Statistics
- Visualizing Data

1.3 Basic Probability Theory

- Concepts of Events
- Probability of an Event
- Probability Rules
- Joint and Conditional Probability
- Independence

1.4 Random Variable

- Definition of Random Variable (RV)
- Discrete and Continuous RV
- Generating Random Variable
- Probability Mass/Density Function
- Multivariate Random Variable

- Joint and Conditional PMFs/PDFs
- Independence

1.5 Statistical Models

- Gaussian Distribution and its relatives
- Chi-Square
- Gamma and Beta
- Student-t
- Cauchy and Fisher-F

1.6 Expectation and Variance

- Concept of Expectation and Variance
- Data-based Estimators (Sample mean and Sample Covariance)
- Generating Multivariate Random Variables
- Law of Large Numbers and Central Limit Theorem

1.7 Descriptive statistics

- Histogram
- Sample Mean and Variance
- Sample Covariance

1.8 Frequentist Statistics

- Mean Square Error (MSE)
- Likelihood Function
- Parameter Estimation
- Maximum Likelihood Estimator (MLE)
- Application in Data Science - Global Warming

1.9 Monte Carlo Method

- Monte Carlo Sampling
- Monte Carlo Estimation

1.10 More on Gaussian Distribution

- PDF of high-dimensional Gaussian Distribution
- Standard Gaussian Distribution
- Visualization
- Sample Mean and Covariance Matrix - MLE Estimation

1.11 Hypothesis Testing

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

1.12 Confidence Intervals

- Definition of Confidence Intervals
- Confidence on Mean
- Confidence on Variance
- Confidence on Correlation Coefficient

1.13 Bayesian Probability

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

1.14 Decision Trees and Ensembele Methods

- Introduction
- Top-Down Construction of Decision Trees
- Splitting rules
- Termination Criterion
- Binary vs Non-Binary Trees
- Data Pre-processing
- Missing Values
- Pruning

1.15 Time-series/Sequential Data Science

- What is Random Process?
- Independent identically-distributed sequences
- Gaussian Process
- Random Walks and Markov Chains
- Markov Chain Mont Carlo
- Application: Time-dependent Prediction