EM 655: MANAGEMENT OF INFORMATION SYSTEMS AND BIG DATA ANALYTICS USING STATISTICAL LEARNING

Spring 2024

Instructor: SangWoo Park, Assistant Professor in Mechanical & Industrial Engineering
Office: Fenster Hall 266
Email: sangwoo.park@njit.edu
Lecture: Wednesday 6PM – 8:50 PM, CKB 315
Class website: Canvas
Office hours: Monday 10:30 AM – 12PM, Fenster 266

1 Course Objectives

Modern day organizations use information systems at all levels of operation to collect, process, and store data. Over the past few decades, technological advances have led to a tremendous increase in the amount of recorded data. In this course, we will learn about information flow in an organization as an integrated system and management resource: techniques of data analysis, design, and processing in the big data regime; data acquisition, storage, retrieval, and transmission to decision-makers. Furthermore, we will discuss computational methods that work on high-dimensional data and are scalable to large data sets; dealing with messy data and how to tackle nonlinearity in models; the course will deliver a comprehensive treatment of the mathematical and statistical models useful for analyzing data sets arising in various disciplines that use information systems, like finance, health care, energy, bioinformatics, security, education, and social services.

One of the main objectives of this course is to teach data analytical tools and statistical/optimization skills using R and hands-on activities to solve engineering problems and to convert real-world large data sets into useful information for decision making. Mastering these tools not only offers tremendous career opportunities for students, but it also enables engineers and managers to formulate and solve complex problems on their computers to aid in decision making.

We will focus on the use of R in data analysis and modeling. The goal is to develop analytical skills to interpret data, make better decisions, and have excellent programming skills. Topics include linear regression, classification, resampling methods, model selection and regularization, nonlinear regression, support vector machines, and deep learning. In addition to R, we will have the opportunity to learn how to use Python for statistical analyses (given time).

Student Learning Outcomes:

- Explore datasets using R and Python.
- Import, clean, store, sort, and filter data using R and Python.
- Build data model, transform raw data, and deliver interactive data visualization.
- Understand the fundamental concepts of statistical learning methods and be able to use them for real-world applications.

2 Required Textbook and Materials

a. Textbook. An Introduction to Statistical Learning (2nd edition), James, Witten, Hastie, Tibshirani **(Required)**

Free Online: `https://hastie.su.domains/ISLR2/ISLRv2_corrected_June_2023.pdf.download.html`

(Optional): Data Mining and Business Analytics with R (1st edition), Johannes Ledolter

b. R. Download the R software and R Studio. If you are an Apple Mac user, please make sure that your laptop can run R scripts before coming to class.

Free Online: [R software] `https://www.r-project.org/`
[R Studio] `https://posit.co/download/rstudio-desktop/`

If you run into issues, please notify the TA or the instructor ahead of time.

3 Grade Determination

Your grade will be determined on the basis of your performance on the activities identified below. One midterm exam and a final exam will be given. Students are expected to complete eight assignments (the majority of which will require R or Python) to get a passing grade from the course. No make-ups for exams will be given. If an individual misses a lecture and fails to participate in the in-class exercises, up to one bonus assignment will be given to make up for that. Additional quizzes or other assignments may be given to everyone in class with or without notice in advance at the instructor's discretion.

a) Point distribution
When preparing your assignments and solutions for the exam, pay attention to the content, cleanliness, and organization of the document. They all contribute to your grade. You will be required to upload R (or Python) codes, and images of the results showing your work. Semester grades will be based on the four main scores:

Component	**Percentage**
Midterm Exam	20%
Final Exam	30%
Assignments (8)	35%
Participation	15%
Total	100%

* Note that final grades will be calculated using the grading scheme above, and so the Total grade column shown in Canvas, which is automatically calculated, does not reflect your final grade. Do not use the percentage grades you see in canvas to calculate your overall grade.

b) Grading policy
Letter grades will be assigned based on the following criteria as a percentage of total points:

c) Exams
One midterm exam and a final exam will be given. The midterm exam will be held on Wednesday, March 6, 2024. The final exam will be held on the final exam week, on the day determined by the school. All exams are closed-book and closed-notes. The exams include a set of problems/questions that are based on the graded problem sets and in-class problems we have covered. Conceptual/theoretical questions will also be included. The final exam will be cumulative. Both midterm and final exams will be proctored. Exam time and dates are set; they will not be changed.

Percent	**Grade**
92.0% or above	A
85.0 - 91.9 %	B+
80.0 - 84.9 %	B
70.0 - 79.9 %	C+
65.0 - 69.9 %	C
60.0 - 64.9 %	D
Lower than 60.0 %	F

Please make all your arrangements based on the exam dates. No make-up exams will be given, so missing an exam will result in a zero grade for the exam. However, **well-documented** special circumstances (e.g., severe illness or injury, death of a close family member) could be considered to provide a make-up exam with the instructor's prior approval.

d) Homework Policy

There will be software assignments where R or Python will be used to solve problems discussed in class. All assignments must be submitted via the Canvas "Assignments" tab by the deadline. Deadlines are based on Eastern Standard Time; if you are in a different time zone, please adjust your submittal times accordingly.

You should attempt to solve the questions yourself. If you are stuck, you can discuss problems with me or your classmates. However, you should provide your own solutions and code file. Plagiarism, i.e., copying somebody else's work will not be tolerated.

Lateness Policy. I encourage you to submit all homework by the due date specified. Late homework will be accepted for up to three days past the due date, but the late penalty will be as follows (note even half-an-hour lateness of the due date will be considered as a day late):

Days Late	**Late Penalty**
1	10%
2	20%
3	30%

e) Attendance

Participation includes the following: regular attendance, timely arrival (at least 5 minutes before the class time to set up the computer), and participation in in-class problem-solving. Regular attendance and participation in class are critical to learning the class material and will be, therefore, a part of your overall grade.

Class participation will account for 15% of the grades. Absences and tardiness may lower your grade.

In-class problem-solving. An essential part of the attendance grade will be determined based on the submissions of your work during in-class problem-solving sessions. I expect you to work on class exercise problems and submit your work by the end of the class through Canvas under the Assignments tab. I often give a chance for students to interact with each other, discuss solution strategies, and learn from each other during in-class group working/problem-solving sessions.

4 Academic Integrity

Academic Integrity is the cornerstone of higher education and is central to the ideals of this course and the university. Cheating is strictly prohibited and devalues the degree that you are working on. As a member of the NJIT community, it is your responsibility to protect your educational investment by knowing and following the academic code of integrity policy that is found at:

`http://www5.njit.edu/policies/sites/policies/files/academic-integrity-code.pdf`.

Please note that it is my professional obligation and responsibility to report any academic misconduct to the Dean of Students Office. Any student found in violation of the code by cheating, plagiarizing or using any online software inappropriately will result in disciplinary action. This may include a failing grade of F, and/or suspension or dismissal from the university. If you have any questions about the code of Academic Integrity, please contact the Dean of Students Office at dos@njit.edu

More on Cheating:

1. Cheating will result in the student receiving a zero grade for the assignment and may result in a failing grade for the class.
2. Turning in an item you did not create is cheating.
3. Copying another person's digital item or work is cheating.
4. Allowing (intended or not intended) someone else to copy your work or the digital item is considered cheating and will result in a failing grade for the assignment.
5. You must do your own work, and do not exchange your work with another student.
6. Having someone complete a homework assignment for you is cheating.

5 Students with Disabilities

If you have a disability or a particular need for which you are or may be requesting accommodations, please contact both the Office of Accessibility Resources and Services (OARS) and me as early as possible in the semester. The official website is `https://www.njit.edu/accessibility/`. You must submit appropriate documentation to the instructor before accommodations can be granted. OARS will review your concerns and determine, with you, what accommodations are necessary and appropriate for you. All information and documentation of your disability is confidential and will not be released by OARS without your written permission.

6 Class Schedule

Please see the **Tentative Class Schedule**:

Class Date	**Topics**
Week 1	Introduction and Syllabus, Basics of R
Week 2	Data import and analysis using R
Week 3	Chapter 2 (What is Statistical Learning?)
Week 4	Chapter 3 (Linear Regression) - part 1
Week 5	Chapter 3 (Linear Regression) - part 2
Week 6	Chapter 4 (Classification) - part 1
Week 7	Chapter 4 (Classification) - part 2 Midterm review
Week 8	Midterm exam
Week 9	Spring break
Week 10	Chapter 5 (Resampling Methods)
Week 11	Chapter 6 (Linear Model Selection and Regularization)
Week 12	Chapter 7 (Moving beyond Linearity)
Week 13	Chapter 8 (Tree-based Methods)
Week 14	Chapter 9 (Support Vector Machines)
Week 15	Chapter 10 (Deep Learning)
Week 17	Final exam