IEMS 304: Statistical Learning for Data Analysis — 2025 Spring

Course Website: https://2prime.github.io/teaching/2024-SL-Da

Instructor Information

Name: Yiping Lu
Office: Tech M237
Email: yiping.lu@northwestern.edu
Office Hours: Monday 10am
TA: Hedieh Sazvar *Email:* hedieh.sazvar@northwestern.edu
Office Hours: Wednsday 12:30–1:30

If you wish to contact me via email, kindly include the tag "[IMES 304]" in the subject line. This will help ensure that I do not overlook your message. Better way to approach me is using campuswire. Please always utilize Campuswire instead of emailing – this helps centralize conversations between us

Campuse Wire: https://campuswire.com/p/G65809778, code:0504
Gradescope Entry Code:VWDZYG (for homework)

About your instructor

I (Dr Yiping Lu) am Assistant Professor of Industrial Engineering & Management Sciences joining Northwestern University in 2024. Before that, I worked as a Courant Instructor at NYU for one year. From 2019 to 2023, I was a Ph.D. student at Stanford University where I obtained my degree in Applied Mathematics, emphasis on Machine Learning (AI) and Numerical analysis (using computers to solve (differential) equations). My research is about using AI to solve hard physics, industrial engineering and system management problems. Our department (IEMS) at Northwestern University is working on how to make important decisions using data based on linear algebra and statistics. My native language is Mandarin, and I also speak Japanese.

Class Information

Time: MWF 9.00AM-9.50AM
Classroom: Tech L251

Course Description

Predictive modeling of data using modern regression and classification methods. Multiple linear regression; logistic regression; pitfalls and diagnostics; nonparametric and nonlinear regression and classification such as trees, nearest neighbors, neural networks, and ensemble methods.

Prerequisites: A previous course in statistics at the level of IEMS 303 plus a course in matrix analysis. Comfort with programming (we will be programming in R) is also necessary.

Course Objectives

- Multiple linear regression basics: model fitting, statistical inference, prediction
- Multiple linear regression: influence, residual diagnostics, multicollinearity, interactions, categorical predictors, variable selection, model evaluation criteria, ridge and lasso regression
- Logistic regression: model fitting and interpretation, statistical inference, diagnostics
- Nonlinear regression basics: maximum likelihood estimation, nonlinear least squares, cross-validation, bootstrapping Classification and regression trees Nearest neighbors for classification and regression Boosted trees and random forests

- Unsupervised Learning: Dimension Reduction, Clustering

Textbook

Required Text:

- James, Gareth, et al. An introduction to statistical learning.
- Stanford CS 229 Lecture Note: https://cs229.stanford.edu/main_notes.pdf

Advanced Topics:

- Hastie, Trevor, et al. The elements of statistical learning: data mining, inference, and prediction.
- Modern Applied Statistics with S, 4th ed. William N. Venables and Brian D. Ripley, Springer, 2002. (classic reference for R).

Pretest

IEMS has implemented a pretest in this course to communicate to students what prerequisite knowledge is relevant for success in the course. The pretest is not intended to be labor-intensive, but it is important you take these pretests seriously to review important concepts for the class. The pretests' data will also be used to identify areas for improvement in the IEMS curriculum.

Passing the pretest is worth 3% of your final course grade. You must achieve a passing score of 70% or higher by Tuesday, April 15 at 11:59 p.m. This deadline will be firmly enforced. It is in your best interest to achieve a passing score on the pretest by the end of the add/change period (April 7) so that you can drop this course if necessary and add a replacement course. You may retake the pretest to achieve the passing score. Feedback will be provided after each attempt, highlighting concepts for review before trying again. If you attempt 15 times without passing, you must request additional attempts from Prof. Wilson (jill.wilson@northwestern.edu). This will require a direct conversation. Note: Informational questions (e.g., "Are you currently declared as an IE major?") do not affect your pretest score.

To take the pretest, go to https://assessments.mccormick.northwestern.edu and log in with your netid and password. If you encounter any problems, contact Prof. Wilson immediately, and include screenshots if possible. You are strongly encouraged to complete this requirement early, so that you can get help with any technical glitches that arise. Tech support will only be provided from the software developers during business hours. Technical issues will not be considered a valid excuse for not passing the pretest by the deadline.

Class Attendance and Participation

It is essential to your success in this course that you attend each lecture and participate in the discussions. Therefore, you are expected to attend each lecture and to show up on time. Should you need to miss a class for any reason, you are to contact the instructor in a timely manner. Reasons for missing lecture must be documentable and presented, if requested. You are responsible for any material covered, any work assigned, or any course changes made during the lecture. *Do not* expect the instructor to provide notes from any class that you might miss. More than three unexcused absences from lectures could result in receiving an 'F' in the course. Furthermore, excessive lateness will also count as an absence. If you are dismissed from lecture due to problems during the lecture, e.g. disruptive behavior or unauthorized cell phone use, then this dismissal will be recorded as an absence.

Grading

The course grade is determined by the following components:

Midterm 1	20%
Midterm 2	20%
Final	20%
Homework	37%
Pretest	3%

Grading Disputes

All questions and requests regarding the amount of points taken off in grading of any assignment (HW, Labs, Exams, etc.) must be submitted in writing within one week of when the grades are posted. Please prepare a clear, detailed written description of the issue, including what your answer was, what the correct answer was, and why you think the amount of points taken off should be adjusted. Also include a photo of the relevant portion of your assignment. For all assignments, the statute of limitations is one week from the time the graded assignment is returned.

Grade Scale

Final grades will be assigned according to the following scale:

A	93 – 100	C+	77 – 79
A–	90 – 92	C	73 – 76
B+	87 – 89	C–	70 – 72
B	83 – 86	D	60 – 69
B–	80 – 82	F	0 – 59

Homework and Lab Assignments

Homework will usually be assigned on an approximately weekly basis. There will be NO CREDIT given for late homeworks or lab assignments, but over the course of the quarter you are allowed to drop your two lowest homework or lab assignment scores (e.g., either two HWs, two labs, or one HW and one lab). HW solutions will be posted on Canvas, so HW assignments may not be graded in detail. For each homework assignment, a randomly chosen subset of the problems will be graded.

IMPORTANT:

All of your HW and Lab assignments must be submitted electronically on Canvas as R Markdown files. Exams: The midterm and final exams will place heavy emphasis on concepts, in addition to interpreting data analysis results from software. They will be closed book and closed notes, except for the following. You are allowed one page of "cheat sheet" notes (8-1/2" by 11", front and back) for the midterm and two pages for the final exam. In addition, if needed, you are allowed to bring a printout of the pdf handout of distribution tables from the textbook appendices (no notes written on the back/margins/etc. of these, though). The pdf handout will be posted on Canvas, if needed. If you are unable to take the midterm and/or final exam at the scheduled times due to travel or other plans, then you should drop this course and take it the next quarter it is offered.

Exams

The two midterms and final exam will place heavy emphasis on concepts, in addition to interpreting data analysis results from software. They will be closed book and closed notes, except for the following. You are allowed one page of "cheat sheet" notes (8-1/2" by 11", front and back) for the midterm and two pages for the final exam. In addition, if needed, you are allowed to bring a printout of the pdf handout of distribution tables from the textbook appendices (no notes written on the back/margins/etc. of these, though). The pdf handout will be posted on Canvas, if needed. If you are unable to take the midterm and/or final exam at the scheduled times due to travel or other plans, then you should drop this course and take it the next quarter it is offered.

Accommodations

I am happy to provide accommodations, understanding that they may be necessary for student success. Students who may need academic accommodation based on the impact of a disability must initiate the request with the AccessibleNU. Students should contact AccessibleNU as soon as possible since timely notice is needed to coordinate accommodations.

Tentative Schedule

The following is a *tentative* schedule for the course.

Week of...	Sections
04/01	Introduction
04/07	Simple Linear Regression
04/14	Multiple Linear Regression
04/21	Multiple Linear Regression
04/28	Model and Variable Selection, Shrinkage, and Multicollinearity
05/05	Midterm & Basic Nonlinear regression/classification
05/12	Neural Networks and Trees
05/19	Ensemble/Committee Methods
05/26	Unsupervised Learning(Dimension Reduction,Clustering)
06/02	Final Review