Spring 2025
STAT 385: Elementary Statistical Techniques for Machine Learning and Big Data

Time: Mon & Wed & Fri, 9:00-9:50 am (Dr. Zhong)
Mon & Wed & Fri, 2:00-2:50 pm (Dr. Embers)

Locations: Lectures (Monday and Wednesday):
Center Building A, Room A005 (9:00-9:50am, Dr. Zhong)
Or
Taft Hall 117 (2:00-2:50pm, Dr. Embers)

Computer Labs (Friday):
Center Building A, Room A005 (9:00-9:50am, Dr. Zhong)
Or
Taft Hall 117 (2:00-2:50pm, Dr. Embers)

Course Modality: In-person On-Campus Instruction and Computer Labs

Instructors: Ping-Shou Zhong /Dale Embers

Office Hours: Mon and Fri, 1:30-2:30pm online or in person at SEO 501
and by appointment (Dr. Zhong)
Mon and Wed, 1:00-1:50 pm,
Fri 11-11:50 am in person at SEO 309 (Dr. Embers)

E-mail: pszhong@uic.edu (Dr. Zhong)
dembers@uic.edu (Dr. Embers)

Teaching Assistant/Grader: Linxin Liu (lliu89@uic.edu)
Office hours: Tuesdays from 9-11 am at MSLC (Math & Science Learning Center)

Textbook:

An Introduction to Statistical Learning: with Applications in R written by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani (Springer 2013)

Book website with free PDF book: https://hastie.su.domains/ISLR2/ISLRv2_website.pdf

Reference:

Data Science for Business: Fundamental principles of data mining and data analytic thinking written by Foster Provost and Tom Fawcett (O'Reilly, 2013).

Book website: http://data-science-for-biz.com/

Prerequisite: Grade C or better in STAT 382; and consent of the instructor. Students in the BS in Data Science must satisfy the prerequisite with grade of C or better in IDS 462 instead of STAT 382.

Credit Hours: 3 hours.
Objectives and Course Content

STAT-385 is an introductory course to statistics concepts, techniques, and methods related to data science. It is an intro class for UIC's program in data science. You need to learn to program using R or concurrently be taking R programming in other courses. This course will introduce data analysis tools and statistical techniques for extracting information from data to solve problems better and improve decision-making. We will introduce the fundamental principles and techniques of machine learning and discuss the applications of machine learning techniques in real-world example contexts. The course will include lectures, computer labs, discussions, class exercises, and real-world examples.

The material will be introduced in the course includes sampling algorithms; nonparametric tests; classification; clustering; classification and regression trees; least absolute shrinkage and selection operator (Lasso); cross-validation; principal component analysis; dimension reduction; neural network and R-packages for big data analysis. The emphasis is on understanding the fundamental concepts, mechanics, and applications of these methods.

Computer Labs

All computing for the course will be done in R. You may find the software on the website https://www.r-project.org/. The software may be downloaded for free. We will discuss the implementation of the methods introduced in the class. Real-world data sets and examples will be used for demonstration. The data sets and lab material will be available on Blackboard.

We will have labs on Friday each week this semester. In each lab, we will prepare data examples and hands-on practice problems, in particular, we will include some data examples with social impacts. You will have an opportunity to access the data examples and analyze them in computer labs. Our emphasis is on implementation, interpretation, and addressing practical issues related to each data set. For the hands-on practice problems in the labs, we encourage you to work with your group members and other classmates to understand what you need to do and discuss possible solutions. However, you need to write the R code on your own, not copy and change code written by someone else. You will need to submit your solution to some lab questions and part of them will be reviewed or graded. Half of the computer lab credits are based on your participation in computer labs.

Homework

Homework assignments will be posted on Blackboard with the due date included. Each homework comprises questions to be answered or real-world applications. Homework should be submitted online on the specified due date. You should work out your homework independently. *Unexcused late submissions are subject to a 10% points reduction (0-24 hours) and a 20% points reduction (24-48 hours). Homework assignments submitted two days after the due date would not be accepted*. In addition, some reading material and exercises from the textbook will be assigned. These problems will not be collected and graded.

Classroom Exercises

Some classroom exercises will be posted before lectures. You are encouraged to complete the exercises before the lectures. These exercises will be discussed in the class. While these exercises will not be graded, it's important to note that they may be compiled as a means of documenting attendance and participation records.

Team Project

A team project is required and will be prepared by student teams. We will give you instructions on how to form your teams. The instructor will discuss with teams about the team projects. The requirements and details of the team project will be discussed in class.

Exams, Grading and Attendance Policy

Your final score will be based on scores in homework, computer labs, one midterm exam, one final project, and your participation. The percentage of each part to the final score is listed below:

Homeworks	25%
Mid-Term Exam	25%
Time: Mar 5, 2024,	
Locations: Center Building A, Room A005 (9:00 - 9:50am, Dr. Zhong) or Taft Hall 117 (2:00-2:50pm, Dr. Embers)	
Team Project	30%
Computer Labs	10%
Participation/Contribution	10%

If you are unable to attend an exam, you must contact the instructor as early as possible before the day of the exam. All excuses will be verified. The make-up exams will be given only under exceptional circumstances.

A tentative criterion for the final grade is

A(>=85%), B(84.9-70%), C(69.9-60%), D(59.9-40%), F(39.9-0%).

Please note that it is your responsibility to attend the class and keep track of the proceedings. Although we will not take attendance in classes, we may use classroom exercises as a means of documenting attendance records.

Important Dates

January 13, M	Classes begin
January 20, M	Martin Luther King Jr. Day. No classes.
March 7, F	Eight-week Part of Term A ends
March 10, M	Eight-week Part of Term B begins.
March 24-28, M-F	Spring vacation. No classes.
May 2, F	Instruction ends

Disability Statement

Students with disabilities must inform the instructor of the need for accommodations. Those who require accommodations for access and participation in this course must be registered with the Disability Resource Center. Please contact ODS at 312-413-2183 (voice) or 312-413-0123 (TTY).

Course Schedule (January 13, 2025 - May 2, 2025)

Week	Mon	Wed	Fri (Computer Labs)	Homework
Week 1 (Jan 13- Jan 17)	Introduction to course or statistical learning (Section 2.1)	Introduction to statistical learning (Section 2.1)	A review and practice on R programming (Lab 1)	Assign homework 1 on Wed (Jan 15)
Week 2 (Jan 20- Jan 24)	MLK, Jr. Day	Assessing Model Accuracy (Section 2.2)	A review and practice on linear algebra with exercise (Lab 2)	Homework 1 due on Wed (Jan 22) Homework 2 assigned on Fri (Jan 24)
Week 3 (Jan 27- Jan 31)	Simple linear regression model (Section 3.1)	Multiple linear regression (Section 3.2)	A lab on multiple linear regression models (Lab 3)	Homework 2 due on Fri (Jan 31) Homework 3 assigned on Fri (Jan 31)
Week 4 (Feb 3- Feb 7)	Subset selection for linear regression models (Section 6.1)	Classification or Logistic regression models (Section 4.1-4.3)	A lab on model selection in linear models (Lab 4)	Homework 3 due on Fri (Feb 7)
Week 5 (Feb 10- Feb 14)	Logistic regression models (Section 4.3)	Linear discriminant analysis (LDA) (Section 4.4)	A lab on logistic regression models (Lab 5)	Homework 4 Assigned on Mon (Feb 10)
Week 6 (Feb 17- Feb 21)	LDA, Bayes Theorem, Naive Bayes (Section 4.4.4)	K nearest Neighbor methods (Section 2.2.3 and 3.5)	A lab on various of classification methods (Lab 6)	

Week 7 (Feb 24- Feb 28)	Cross-validation (Section 5.1)	Bootstrap (Section 5.2)	A lab on cross-validation (KNN) and bootstrap (Lab 7)	Homework 4 due on Tues (Feb 25)
Week 8 (Mar 3- Mar 7)	A review on midterm exam	Midterm Exam	Introduction to optimization and practice	
Week 9 (Mar 10- Mar 14)	Shrinkage methods (Section 6.1), Ridge regression and LASSO (Sections 6.1 and 6.2)	Tuning parameter selection in LASSO/ PCA (Section 6.2.3)	A lab on ridge regression, subset selection and lasso (Lab 8)	Homework 5 assigned on Mon (Mar 10)
Week 10 (Mar 17- Mar 21)	PCA and Partial least squares (Section 6.3)	Partial least squares (Section 6.3)	Introduction to final project and a lab on PCA and PLS (Lab 9)	Homework 5 due on Mon (Mar 17) Homework 6 assigned on Wed (Mar 19)
Week 11 (Mar 31 - Apr 4)	Classification trees (Section 8.1)	Classification trees and bagging (Section 8.1 and 8.2.1)	A lab on classification trees (Lab 10)	Homework 6 due on Fri (Apr 4)
Week 12 (Apr 7- Apr 11)	Random Forest (Section 8.2.2)	Boosting (Section 8.2.3)	A lab on Bagging, Boosting and Random Forest (Lab 11)	Homework 7 assigned on Mon (Apr 7)
Week 13 (Apr 14- Apr 18)	Support vector machine (linear) (Section 9.1 and 9.2)	Kernelized SVM (Section 9.3)	A lab on SVM (Lab 12)	Homework 7 due on Mon (Apr 14) Homework 8 assigned on Wed (Apr 16)
Week 14 (Apr 21 - Apr 25)	K-means clustering (Section 12.4.1)	Hierarchical clustering (Section 12.4.2)	A lab on clustering (Lab 13)	Homework 8 due on Fri (Apr 25)
Week 15 (Apr 28- May 2)	Final presentations	Final presentations	Final presentations	

Homework Assignments

Homework 1: Statistical learning, model accuracy and bias-variance trade-off
Homework 2: R programming and Linear algebra practice
Homework 3: Linear regression models
Homework 4: Various classification methods
Homework 5: Shrinkage methods like lasso and ridge regression (include some optimization exercises)
Homework 6: PCA, PCR and PLS
Homework 7: Ensemble methods: bagging, boosting and random forests
Homework 8: SVM

Classroom Environment

UIC values diversity and inclusion. Regardless of age, disability, ethnicity, race, gender, gender identity, sexual orientation, socioeconomic status, geographic background, religion, political ideology, language, or culture, we expect all members of this class to contribute to a respectful, welcoming, and inclusive environment for every other member of our class. If aspects of this course result in barriers to your inclusion, engagement, accurate assessment, or achievement, please notify us as soon as possible. If your name does not match the name on the class roster, please let us know.

The instructors reserve the right to make any changes that he deems academically advisable. Such changes, if any, will be announced in class or online.