IND E 427: Data Analytics for Systems Engineering

Instructor: Prof. Shuai Huang
Office: AERB 141B
Phone: 206-685-2953
Email: shuaih@uw.edu

TA: Ameer Hamza Shakur
Email: ahamza@uw.edu

Office hour:

- Online: T/Th 1:30-2:30 (Shuai). Please use the link: https://washington.zoom.us/meeting/98756328896
- In person: You are also welcomed to stay after class to discuss with Shuai
- Online: W/F 1pm-2pm (Ameer). Please use the link: https://washington.zoom.us/j/98556170142
- Other time options available by email communication

Focus of the course: This course is about *principles* and *techniques* of statistical modeling and machine learning, designed mainly as a beginner class for senior undergraduate students and graduate students. It primarily focuses on the theories, ideas, and algorithms behind statistical learning methods for analytic decision-makings in applications. Data analysis examples in R will be mentioned and demonstrated.

Prerequisite: IND E 315, STAT 390, STAT 391, or other stats class

Recommended prep: Basic programming skills; experience with R, Python, or Matlab, etc.

Textbook: **Data Analytics: A Small Data Approach**, by Shuai Huang and Houtao Deng. https://www.amazon.com/Data-Analytics-Approach-Chapman-Science/dp/0367609509.

A free online version: https://dataanalyticsbook.info/.

Homework: The students will form study groups, and each group should submit one copy of the homework. Students in the same group receive the same grade. Homework must be submitted in CANVAS as one PDF file (scanned copy or photo is fine, but make sure it has a good quality). Homework submitted late will be penalized by 20% of the total points. Homework will NOT be accepted more than 24 hours after it is due.

Exams: There will be a take-home midterm exam and a take-home final exam.

Bonus assignment: You can have 2 bonus points rewards added to your final total score if you write up a blog article for CI^2 (https://ci2.us/) about whatever you have learned in this class. A brief introduction of CI^2 can be found here: https://ci2.us/about/. You can write solo or with your group members. All the authors receive the 2 bonus points. If you want to write about your project, you need to use discretion about the completeness of your article as a story, so your blog article will be a re-telling of the story in your project report. An example is here: https://ci2.us/post/2020/08/30/galton-machine-and-normal-distribution/. If you want to contribute such an article, please contact Shuai anytime before Dec 9th 2021 with a draft of

your article. He will discuss with you about your timeline and if your timeline is reasonable (note that the grading deadline is 5pm Dec 22th 2021, https://www.washington.edu/students/reg/2021cal.html), he will help you develop this article. Note that by agreeing to submit your article to CI^2 you transfer the copyright to CI^2.

How the Study Group Works:

Mechanism

- Students will self-organize into study groups the first day of class. Groups should be at most 3 students.
- Each group will have a leader. Broadly, leaders will be responsible for coordinating HW tasks, organizing project tasks, and attending project update meetings.
- **Leaders are required to attend project update meetings held during the Professor's office hours**

Group Project

- It will be a fun project! The instructor will work closely with you to design the project, monitor your progress in biweekly meetings, and provide feedbacks for you so you could have a rebuttal opportunity for one week to revise your report based on the instructor's comments on your submitted report and your presentation. See the appendix for details.

Grading:

Project	40%
Homework	30%
Exams	30%

Course Topics and Schedule:

09/30	Course Intro & Linear regression model	Chapter 1-2	
10/05	Decision tree	Chapter 2	**Group form**
10/07	Logistic regression	Chapter 3	
10/12	Bootstrap	Chapter 4	**HW1 due**
10/14	Random forest	Chapter 4	
10/19	Cross-validation and out-of-bag (OOB) errors	Chapter 5	
10/21	Residual analysis	Chapter 6	**HW2 due**
10/26	Clustering	Chapter 6	
10/28	**Deliverable: Presentation to start the project**		
Take home midterm exam (weekend 10/30-10/31)			
11/02	Clustering	Chapter 7	
11/04	LASSO	Chapter 7	
11/09	Variable importance in tree models	Chapter 7	**HW3 due**
11/11	Support vector machine	Chapter 8	
11/16	Support vector machine	Chapter 8	
11/18	Principal component analysis	Chapter 7	
11/23	Principal component analysis	Chapter 7	**HW4 due**
11/25	**No class on Thanksgiving**		
11/30	Other regression models: kernel regression	Chapter 9	
12/02	Review for final exam	Chapter 9	
12/07 & 12/09	**Deliverable: Presentation to conclude the project & HW5 due**		
Take home exam (weekend 12/11-12/12)			

Schedule for Project Meetings

Week 3 **(optional)** (10/08)	Group leaders attend the office hour hosted by the instructor, discuss the project topic
Week 4 **(optional)** (10/17)	Project update. Group leaders attend the office hour hosted by the instructor
Week 5 (10/28)	Deliverable: kick-off presentation of each team to introduce their projects to the class
Week 7 **(optional)** (11/07)	Project update. Group leaders attend the office hour hosted by the instructor
Week 9 **(optional)** (11/21)	Project update. Group leaders attend the office hour hosted by the instructor
Week 10 (12/07, 12/09)	Deliverable: presentation of each team to conclude their projects to the class

Changes: The syllabus is an arrangement of the course activities. The instructor reserves the right to make changes to the syllabus during the course. Any necessary changes will be announced in class and posted on the website.

In-person class: *This class is conducted in-person. Students are expected to participate in class to fully benefit from course activities and meet the course's learning objectives. Students should only register for this class if they are able to attend in-person. To protect their fellow students, faculty, and staff, students who feel ill or exhibit possible COVID symptoms should not come to class. When absent, it is the responsibility of the student to inform the instructor in advance (or as close to the class period as possible in the case of an unexpected absence), and to request appropriate make-up work as per policies established in the syllabus. What make-up work is possible, or how assignments or course grading might be modified to accommodate missed work, is the prerogative of the instructor. For chronic absences, the instructor may negotiate an incomplete grade after the 8th week, or recommend the student contact their academic adviser to consider a hardship withdrawal (known as a Registrar Drop).*

Academic Misconduct*: The university's policy on plagiarism and academic misconduct is a part of the Student Conduct Code, which cites the definition of academic misconduct in the WAC 478-121. (WAC is an abbreviation for the Washington Administrative Code, the set of state regulations for the university. The entire chapter of the WAC on the student conduct code is here.) According to this section of the WAC, academic misconduct includes "Cheating"—such as "unauthorized assistance in taking quizzes", "Falsification"*

"which is the intentional use or submission of falsified data, records, or other information including, but not limited to, records of internship or practicum experiences or attendance at any required event(s), or scholarly research"; and "Plagiarism" which includes "[t]he use, by paraphrase or direct quotation, of the published or unpublished work of another person without full and clear acknowledgment."

Accommodation: *Your experience in this class is important to me. If you have already established accommodations with Disability Resources for Students (DRS), please communicate your approved accommodations to me at your earliest convenience so we can discuss your needs in this course.*

Washington state law requires that UW develop a policy for accommodation of student absences or significant hardship due to reasons of faith or conscience, or for organized religious activities. The UW's policy, including more information about how to request an accommodation, is available at Religious Accommodations Policy (https://registrar.washington.edu/staffandfaculty/religious-accommodations-policy/). Accommodations must be requested within the first two weeks of this course using the Religious Accommodations Request form (https://registrar.washington.edu/students/religious-accommodations-request/).

Appendix: Team Project

A team (with 3 students as the maximum) will conduct the project. It will be fine if you plan to work alone. Each team will submit one report and present their project at the middle and the end of the course. You are encouraged to choose a project related to your own research interests, and please feel free to discuss your project with the instructor. The objective of a class project is to help you gain experience and to relate what you learn in this course to real life problems.

Be aware of the following:

1. A Project Proposal presentation (3-5 pages of slides) should be submitted. You will need to provide the following information.
    a. Your name(s)
    b. Project description
    c. How and where you obtained the data
    d. Questions you may want to address using the data and corresponding data mining & statistical learning methods
2. The Final Project: You will need to submit two deliverables:
    a. The presentation file. Each presentation is around 20 minutes (= 15 + 5 QA).
    b. A final summary report of your class project. The final summary report shall not be longer than 15 pages, and the body of the report (without appendix and figures/tables) is generally 4 ~ 8 pages. Only very relevant plots and tables shall be included in the body of the report, and the rest should go to Appendix.
3. Grading: all team members will receive the same grade on the project. Your grade on the project will depend on you selecting and adhering to a logical and readable format for the report (10 points); on the appropriate use of statistical technique (20 points); on the appropriateness in the conclusions of your report (20 points); and on the readability and understandability of the report when technical material is needed (30 points – including the presentation of your project in delivering your ideas). Finally, the report as a whole will be evaluated (20 points). You will have a rebuttal opportunity for one week to revise your report based on the instructor's comments on your submitted report and your presentation.
4. Datasets: You can collect the data by yourself, use the data set from your own research or the public domain. The followings are some examples of online datasets:
    a. Kaggle data science competitions: https://www.kaggle.com/competitions
    b. DREAM Challenges: http://dreamchallenges.org/
    c. KDD cups: http://kdd.ics.uci.edu/ or http://archive.ics.uci.edu/ml/. One example is the KDD cup 1999 data at

http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.More KDD cup data can be found at http://www.sigkdd.org/kddcup/index.php

d. PHM data competitions: e.g., the challenge in 2016 could be found here https://www.phmsociety.org/events/conference/phm/16/data-challenge

e. UCI data repository: http://archive.ics.uci.edu/ml/

5. Here is a suggested format for your summary report.

a. Title Page: Project Title, author(s) (your name and email address), the submission date;

b. Abstract: informative summary of the whole report (100-300 words).

c. Introduction includes problem description and motivation, data challenge(s), problem solving strategies, accomplished learning from the applications and outline of the report.

d. Problem Statement or Data Sources: cite the data sources, and provide a simple presentation of data to help readers understand the problem or challenge(s).

e. Proposed Methodology: explain (and justify) your proposed analysis strategies.

f. Analysis and Results: present key findings when executing the proposed analytic methods. For the benefit of readability, detailed results should be placed in the Appendix. Reference of R codes to implement your proposed methods should be given.

g. Conclusions: Draw conclusions from your practice. Unfinished or possible future work could be included (with proper explanation or justification). Please add a subsection for lessons you learned from this project.

h. Appendix: This section only includes needed documents to support the presentation in the report. Feel free to divide it into several subsections if necessary. Do NOT dump all computer outputs unorganized here.

i. Bibliography and Credits.