Course Information

Course Number:	STAT 639/ECEN 758/CSCE 676
Course Title:	Data Mining and Analysis
Section:	TBA
Time:	TBA
Location:	TBA
Credit Hours:	3

Instructor Details

Instructor:	Yang Ni
Office:	BLOC 458E
Phone:	N/A
E-Mail:	yni@stat.tamu.edu
Office Hours:	TBA

Course Description

This course is an introduction to concepts, methods, and practices in statistical data mining. We will provide a broad overview of topics that are related to supervised and unsupervised learning. See the tentative schedule on the last page for details. Emphasis will be placed on applied data analysis.

This course will also have distance section and all the instructions for the distance and on campus students will be the same (see the Grading Policy below).

Course Prerequisites

Familiarity with programming language R and knowledge of basic multivariate calculus, statistical inference, and linear algebra is expected. Students should be comfortable with the following concepts: probability distribution functions, expectations, conditional distributions, likelihood functions, random samples, estimators and linear regression models.

Special Course Designation

Cross-listed as ECEN 758 and CSCE 676.

Course Learning Outcomes

Students will learn how and when to apply statistical learning techniques, their comparative strengths and weaknesses, and how to critically evaluate the performance of learning algorithms. Students who successfully complete this course should be able to apply basic statistical learning methods to build predictive models or perform exploratory analysis, and make sense of their findings.

Suggested Textbooks

(ISLR) An Introduction to Statistical Learning with Applications in R by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani
http://www-bcf.usc.edu/~gareth/ISL/index.html

(PRML) Pattern Recognition and Machine Learning by Christopher M. Bishop
https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf

(DMA) Data Mining and Analysis: Fundamental Concepts and Algorithms by Mohammed J. Zaki and Wagner Meira, Jr.
http://www.dataminingbook.info/pmwiki.php

(ESL) The Elements of Statistical Learning: Data Mining, Inference, and Prediction by Trevor Hastie, Robert Tibshirani and Jerome Friedman
https://web.stanford.edu/~hastie/ElemStatLearn/

Grading Policy

- Midterm Exam --- 40%
  - On-campus and distance student will take the same exam.
  - On-campus students will take the exam in the classroom.
  - Distance students will take the exam remotely, proctored via ZOOM.
- Project --- 60%
  - On-campus and distance student will have the same project.
  - On-campus students will present their projects in the classroom.
  - Distance students will present their projects via recorded videos.

The final letter grade will be assigned according to the following scheme for both on-campus and distance students:

Course Grade	Points Needed
A	90-100%
B	80-89%
C	70-79%
D	60-69%
F	0-59%

Late Work Policy

No late work will be accepted. Work submitted by a student as makeup work for an excused absence is not considered late work and is exempted from the late work policy. (See Student Rule 7.)

Course Schedule (Tentative)

Day	Date	Topic	Due
Tue	1/18	Introduction I	
Thur	1/20	Introduction II	
Tue	1/25	Linear Regression	
Thur	1/27	Classification I	
Tue	2/1	Classification II	
Thur	2/3	Classification III	
Tue	2/8	R Session: Classification	
Thur	2/10	Resampling I	
Tue	2/15	Resampling II	
Thur	2/17	R Session: Resampling	
Tue	2/22	Regularization	
Thur	2/24	R Session: Regularization	
Tue	3/1	Tree I	
Thur	3/3	Tree II	
Tue	3/8	Tree III	
Thur	3/10	R Session: Tree	
Tue	3/22	Midterm	
Thur	3/24	Neural Networks	
Tue	3/29	Clustering I	
Thur	3/31	Clustering II	
Tue	4/5	Clustering III	
Thur	4/7	PCA	
Tue	4/12	R Session: Clustering & PCA	
Thur	4/14	Community Detection I	
Tue	4/19	Community Detection II	Project Slides
Thur	4/21	Project Presentation I	
Tue	4/26	Project Presentation II	
Thur	4/28	Project Presentation III	Project Report

Tentative Topics

Regression and classification
Bootstrap and cross-validation
Regularization
Decision tree, bagging, random forest, boosting
Neural network
Clustering
Principle component analysis
Community detection

Computing

Statistical Software: we will primarily use the open source statistical software R.

- Go to http://cran.r-project.org to download R for free.
- We strongly recommend downloading R-Studio from http://www.rstudio.com and working in that environment. It is free, and it runs on Windows, Mac and Linux operating systems.
- Also make sure to install ISLR package (https://CRAN.R-project.org/package=ISLR), which includes the datasets used in the course book.

University Policies

Attendance Policy

The university views class attendance and participation as an individual student responsibility. Students are expected to attend class and to complete all assignments.

Please refer to Student Rule 7 in its entirety for information about excused absences, including definitions, and related documentation and timelines.

Makeup Work Policy

Students will be excused from attending class on the day of a graded activity or when attendance contributes to a student's grade, for the reasons stated in Student Rule 7, or other reason deemed appropriate by the instructor.

Please refer to Student Rule 7 in its entirety for information about makeup work, including definitions, and related documentation and timelines.

Absences related to Title IX of the Education Amendments of 1972 may necessitate a period of more than 30 days for make-up work, and the timeframe for make-up work should be agreed upon by the student and instructor" (Student Rule 7, Section 7.4.1).

"The instructor is under no obligation to provide an opportunity for the student to make up work missed because of an unexcused absence" (Student Rule 7, Section 7.4.2).

Students who request an excused absence are expected to uphold the Aggie Honor Code and Student Conduct Code. (See Student Rule 24.)

Academic Integrity Statement and Policy

"An Aggie does not lie, cheat or steal, or tolerate those who do."

"Texas A&M University students are responsible for authenticating all work submitted to an instructor. If asked, students must be able to produce proof that the item submitted is indeed the work of that

student. Students must keep appropriate records at all times. The inability to authenticate one's work, should the instructor request it, may be sufficient grounds to initiate an academic misconduct case" (Section 20.1.2.3, Student Rule 20).

You can learn more about the Aggie Honor System Office Rules and Procedures, academic integrity, and your rights and responsibilities at aggiehonor.tamu.edu.

Americans with Disabilities Act (ADA) Policy

Texas A&M University is committed to providing equitable access to learning opportunities for all students. If you experience barriers to your education due to a disability or think you may have a disability, please contact Disability Resources in the Student Services Building or at (979) 845-1637 or visit disability.tamu.edu. Disabilities may include, but are not limited to attentional, learning, mental health, sensory, physical, or chronic health conditions. All students are encouraged to discuss their disability related needs with Disability Resources and their instructors as soon as possible.

Title IX and Statement on Limits to Confidentiality

Texas A&M University is committed to fostering a learning environment that is safe and productive for all. University policies and federal and state laws prohibit gender-based discrimination and sexual harassment, including sexual assault, sexual exploitation, domestic violence, dating violence, and stalking.

With the exception of some medical and mental health providers, all university employees (including full and part-time faculty, staff, paid graduate assistants, student workers, etc.) are Mandatory Reporters and must report to the Title IX Office if the employee experiences, observes, or becomes aware of an incident that meets the following conditions (see University Rule 08.01.01.M1):

- The incident is reasonably believed to be discrimination or harassment.
- The incident is alleged to have been committed by or against a person who, at the time of the incident, was (1) a student enrolled at the University or (2) an employee of the University.

Mandatory Reporters must file a report regardless of how the information comes to their attention – including but not limited to face-to-face conversations, a written class assignment or paper, class discussion, email, text, or social media post. Although Mandatory Reporters must file a report, in most instances, a person who is subjected to the alleged conduct will be able to control how the report is handled, including whether or not to pursue a formal investigation. The University's goal is to make sure you are aware of the range of options available to you and to ensure access to the resources you need.

Students wishing to discuss concerns in a confidential setting are encouraged to make an appointment with Counseling and Psychological Services (CAPS).

Students can learn more about filing a report, accessing supportive resources, and navigating the Title IX investigation and resolution process on the University's Title IX webpage.

Statement on Mental Health and Wellness

Texas A&M University recognizes that mental health and wellness are critical factors that influence a student's academic success and overall wellbeing. Students are encouraged to engage in healthy self-care by utilizing the resources and services available from Counseling & Psychological Services (CAPS). Students who need someone to talk to can call the TAMU Helpline (979-845-2700) from 4:00 p.m. to 8:00 a.m. weekdays and 24 hours on weekends. 24-hour emergency help is also available through the National Suicide Prevention Hotline (800-273-8255) or at suicidepreventionlifeline.org.