STAT 639 Data Mining and Analysis

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.

Learning Objectives

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.

Prerequisites/Corequisites

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.

Suggested Textbooks

- **(ISLR) An Introduction to Statistical Learning with Applications in R** by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani.
  Available online: **http://www-bcf.usc.edu/ gareth/ISL/index.html**
- **(PRML) Pattern Recognition and Machine Learning** by Christopher M. Bishop.
  Available online: **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.
  Available online: **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. This is a more advanced and more comprehensive version of the ISLR.
  Available online: **https://web.stanford.edu/ hastie/ElemStatLearn/**

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 dowloading `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, which includes the datasets used in the course book. **https://CRAN.R-project.org/package=ISLR**
- Some resources if you are unfamiliar with `R`:
  - An excellent introduction to `R` is the book

    Using R for Introductory Statistics by J. Verzani. The book is freely available at **http://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf**.
  - Additional (excellent) tutorials are the

    R Tutorial Videos, **http://dist.stat.tamu.edu/pub/rvideos/**. The site includes `R` scripts and data sets to follow along.
  - Other tutorials: **An Introduction to R**, **R for Beginners**, **simpleR**, **Princton**, **Quick R**, **R Reference Card**, **R Manuals**, **R Wiki**.

Homework

Bi-weekly homework will be posted on eCampus and due on Thursday before midnight. Students are allowed to work in groups up to 4 people and submit one copy for each group. All group members are expected to work on all problems. Upload your solution as a single PDF document on eCampus.

Pop Quiz

We will have pop quiz randomly during class. Mostly in format of multiple choices, just to test how well you master the concepts we have learned so far. Each quiz will take no more than 10 minutes.

Grading Policy

- Homework — 20%
- Quiz — 20%
- Midterm Exam — 30%
- Final Exam — 30%

The final letter grade will be assigned according to the following scheme. I may curve up the scores depending on overall class scores at the end of the semester.

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

Course Policies

Copyright Notice

The handouts used in this course are copyrighted. By "handouts", I mean all materials generated for this class, which include but are not limited to syllabi, quizzes, exams, lab problems, in-class materials, review sheets, and additional problem sets. Because these materials are copyrighted, you do not have the right to copy the handouts, unless I expressly grant permission.

Absence

Only university excused absences will be accepted for missing homework or exams. A student is responsible for providing satisfactory evidence to the instructor to substantiate the reason for absence. If you know that you will miss an exam for a valid reason, please communicate with me over email at your earliest convenience. Check **http://student-rules.tamu.edu/rule07** for what constitutes a university excused absence.

Statement of Disabilities

The Americans with Disabilities Act (ADA) is a federal anti-discrimination statute that provides comprehensive civil rights protection for persons with disabilities. Among other things, this legislation requires that all students with disabilities be guaranteed a learning environment that provides for reasonable accommodation of their disabilities. If you believe you have a disability requiring an accommodation, please contact Disability Services, currently located in the Disability Services building at the Student Services at White Creek complex on west campus or call 979-845-1637. For additional information, visit **http://disability.tamu.edu**.

Statement of Plagiarism

As commonly defined, plagiarism consists of passing off as one's own ideas, words, writing, etc., which belong to another. In accordance with that definition, you are committing plagiarism if you copy the work of another person and turn it in as your own, if you have the permission of that person. Plagiarism is one of the worst academic sins, for the plagiarist destroys the trust among colleagues without which research cannot be safely communicated. If you have any questions regarding plagiarism, please consult the latest issue of the Texas A&M University Student Rules, under the section "Scholastic Dishonesty".

Academic Integrity Statement

"An Aggie does not lie, cheat, or steal or tolerate those who do." The Aggie Honor Council Rules and Procedures are available at the website: **aggiehonor.tamu.edu**.