SDS323 (55430): Statistical Learning and Inference – Fall 2019

Instructor: Giorgio Paulon - `giorgio.paulon@utexas.edu`
TA: Ciara Nugent - `ciara.nugent@utexas.edu`

Class Website: **Canvas**
Class Hours: T TH 2:00 – 3:30pm, PHR 2.114
Instructor Office Hours: TH 3:30 – 5:30pm, GDC 6.828
TA Office Hours: W 9:00 – 11:00am, MEZ 1.104

Course Description

This course is an introduction to statistical inference, broadly construed as the process of drawing conclusions from data, and of quantifying uncertainty about said conclusions. The goal is to introduce the basic ideas of statistical learning and predictive modeling from a statistical, theoretical and computational perspective, together with applications to real data. Topics cover the major schools of thought that influence modern scientific practice, including classical frequentist methods, machine learning and Bayesian inference. The course aims to provide a very applied overview of some classical linear approaches such as *Linear Regression, Logistic Regression, Linear Discriminant Analysis,* as well as some non-linear methods such as *K-Means Clustering, K-Nearest Neighbors, Generalized Additive Models, Decision Trees, Boosting, Bagging and Support Vector Machines.*

This course may be used to fulfill the mathematics component of the university core curriculum and addresses the following three core objectives established by the Texas Higher Education Coordinating Board: communication skills, critical thinking skills, and empirical and quantitative skills. This course carries the Quantitative Reasoning flag. Quantitative Reasoning courses are designed to equip you with skills that are necessary for understanding the types of quantitative arguments you will regularly encounter in your adult and professional life. You should therefore expect a substantial portion of your grade to come from your use of quantitative skills to analyze real-world problems.

Prerequisites/Corequisites

Formal prerequisites are M408D (or M408M), SDS321 (or M362K) or the equivalent. 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. This course will make extensive use of the statistical software `R` and build on knowledge of introductory probability and statistics, as well as multiple regression. If you have any doubt about your preparation for this course, feel free to chat with me on the first day.

Textbook and Materials

- Required textbook: *An Introduction to Statistical Learning with Applications in R* by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. Available online: **http://faculty.marshall.usc.edu/gareth-james/ISL/**

- Recommended textbook: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* by Trevor Hastie, Robert Tibshirani and Jerome Friedman. The required text is a simplified version of this book. Please refer to it if you want to go more into depth on any topic covered in the course text, particularly from a theoretical perspective. Available online: **https://web.stanford.edu/~hastie/ElemStatLearn/**

- Statistical Software: this class will primarily use the open source statistical software `R`. `R` has several advantages; in addition to supporting all of the statistical learning methods that will be covered, it is also the choice for research statisticians.
  - Go to **https://www.r-project.org** to download `R` for free
  - Downloading and getting familiar with `R Studio` from **https://www.rstudio.com/products/rstudio/download/** is strongly recommended. It is free, and it runs on Windows, Mac and Linux operating systems
  - Make sure to install the ISLR package, which includes the datasets used in the course book.

- An excellent introduction to `R` is the book *Using R for Introductory Statistics* by J. Verzani. This book is freely available at **https://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf**

Course Structure

Lectures

The class meets twice a week for 75-minutes each time. Lectures are distributed amongst different topics and `R` sessions accordingly. Students are encouraged to participate in the lecture by asking questions and making comments.

Homework assignments

Bi-weekly homework will be posted on Canvas (usually on Thursdays) and due 2 weeks later, before midnight. Students are allowed to work in groups of up to 4 people and submit one copy for each group. All group members are expected to work on all problems, either together or individually. Use the group to compare solutions and combine them into one common written assignment, not to split the problems. Upload your solution as a single PDF document on Canvas. Latex (a template will be provided) and `R` markdown files are strongly recommended. If you submit hand-written copies that are not legible, they will NOT be graded!

Pop Quizzes

Pop quizzes will be distributed during some of the classes. These will be mostly in the format of multiple choice questions and their aim is to test how well you have mastered the concepts learnt so far. Each quiz will take no more than 15 minutes. The impact of these quizzes on the final grade is minimal, they are simply used to gauge the general level of the class.

Midterm Exams

Two midterms will be held in class as a closed book/notes exam. The format will be similar to homework and quiz questions, without programming in `R`. However, you might be tested on interpreting `R` commands and outputs.

Take-home Final Exam

There will be a take-home final exam after the last day of class. This final exam will be in the format of a data analysis problem using `R` and it will be due 48 hours after the assignment. You are encouraged to use every available resource, including class materials and the internet. However, you must work individually without help from anyone. The final deliverable includes well-documented `R` code and a report to summarize your results and findings.

Grading Policy

The typical UT grading scale will be used. The instructor reserves the right to curve the scale dependent on overall class scores at the end of the semester. Any curve will only ever make it easier to obtain a certain letter grade. The grade will count the assessments using the following proportions:

- **30%** of your grade will be determined by homework assignments
- **10%** of your grade will be determined by the in class pop quizzes
- **30%** of your grade will be determined by the two midterm exams
- **30%** of your grade will be determined by the take-home final exam

The ± system in assigning final grades will be used, and the grade cutoffs will be:

Course Grade	Points Needed
A	93%
A-	90%
B+	87%
B	83%
B-	80%
C+	77%
C	73%
C-	70%
D+	67%
D	63%
D-	60%

Late assignments and missed exams or quizzes will receive a grade of zero except under the following circumstances:

1. You are away from UT as part of a UT-sponsored activity including athletics. Check with the instructor if you are uncertain whether your absence qualifies.
2. The quiz, lab, or exam is in conflict with a religious observance, notify the instructor by the 12th day of class.
3. You suffer from a chronic, documented illness or an emergency that results in your missing an exam or lab. Under these circumstances, contact the instructor as soon as possible to discuss a course of action.

Course Policies Student Accommodations

Students with a documented disability may request appropriate academic accommodations from the Division of Diversity and Community Engagement, Services for Students with Disabilities, 512-471-6259 (voice) or 1-866-329-3986 (video phone). **http://ddce.utexas.edu/disability/about/**

- Please request a meeting as soon as possible to discuss any accommodations
- Please notify the instructor as soon as possible if the material being presented in class is not accessible
- Please notify the instructor if any of the physical space is difficult for you

Emergency Evacuations

Occupants of buildings on The University of Texas at Austin campus are required to evacuate buildings when a fire alarm is activated. Alarm activation or announcement requires exiting and assembling outside. Familiarize yourself with all exit doors of each classroom and building you may occupy. Remember that the nearest exit door may not be the one you used when entering the building. Students requiring assistance in evacuation shall inform the instructor in writing during the first week of class. In the event of an evacuation, follow the instruction of faculty or class instructors. You will be notified via Canvas for follow up directions regarding missed class time. Do not re-enter a building unless given instructions by the following: Austin Fire Department, The University of Texas at Austin Police Department, or Fire Prevention Services office. More information can be found at: **www.utexas.edu/emergency**

If You Need Help

CMHC Crisis Line is a confidential service of CMHC that offers an opportunity for UT-Austin students to talk with trained counselors about urgent concerns. A counselor is available every day of the year, including holidays. You can call when you want, at your convenience. Telephone counselors will spend time addressing your immediate concerns. CMHC 24/7 Crisis Line: 512- 471-CALL (2255).

Academic Honesty

The course is built upon the idea that team-based learning is important and a powerful way to learn. Students are encouraged to study together. However, there are times when you need to demonstrate your own ability to work and solve problems. Your take-home final is to be completed on your own, without discussion with your peers. You can work with other students to complete your homework, but you cannot copy answers directly from someone else (including resources on-line). Students who violate these expectations can expect to receive a failing grade on the assignment and be reported to the Student Judicial Services for academic dishonesty.

Tentative schedule and weekly learning goals

The following schedule is tentative and will be updated and posted on Canvas throughout the course.

Day	Date	Topic	Assignment	Due	Readings
TH	8/29	Introduction	HW0		1
T	9/3	Statistical Learning overview			2
TH	9/5	*R Session: Introduction to R*			
T	9/10	Introduction to the Linear Model			3.1
TH	9/12	Multiple Linear regression and potential problems	HW1	HW0	3.2, 3.3
T	9/17	*R Session: Linear Regression*			
TH	9/19	Classification			4.1, 4.2, 4.3
T	9/24	Classification			4.4, 4.5
TH	9/26	*R Session: Classification*	HW2	HW1	
T	10/1	Resampling methods			5.1, 5.2
TH	10/3	*R Session: Resampling methods*			5.2
T	10/8	Linear Model selection			6.1
TH	10/10	Linear model regularization		HW2	6.2, 6.3
T	10/15	**Midterm Exam 1**			
TH	10/17	*R Session: Model selection*	HW3		
T	10/22	Moving beyond linearity			7.1, 7.2, 7.3, 7.4
TH	10/24	Moving beyond linearity			7.5, 7.6, 7.7
T	10/29	*R Session: Moving beyond linearity*			
TH	10/31	Tree based methods	HW4	HW3	8.1
T	11/5	Tree based methods			8.2
TH	11/7	*R Session: Tree based methods*			
T	11/12	Support Vector Machines			9.1, 9.2, 9.3
TH	11/14	*R Session: Support Vector Machines*	HW5	HW4	
T	11/19	**Midterm Exam 2**			
TH	11/21	Unsupervised Learning			10
T	11/26	Unsupervised Learning			10
TH	11/28	**Thanksgiving**		HW5	
T	12/3	*R Session: Unsupervised Learning*			
TH	12/6	Special topic TBD			

Special topics may include: Bayesian statistics, Neural networks, …