STATISTICAL ANALYSIS AND LEARNING
PA 397C (61005)/ EER 396 (25830)/ INF 397 (27455)/ME 397 (17924), Spring 2018

Professor Varun Rai
Associate Professor, LBJ School of Public Affairs
Associate Professor, Mechanical Engineering

OFFICE:
Sid Richardson Hall, Unit 3, Room: SRH 3.232
Email: raivarun@utexas.edu. *Phone*: 512-471-5057
Office hours: Mon 10:30-noon and by appointment
Teaching Assistants:
Sanjit Paliwal (sanjit.paliwal@utexas.edu); Office hours: Mon noon-2; Fri 10-noon
Vivek Khetan (vivek.khetan@utexas.edu); Office hours: Tue 11a-1p and Thu noon-2p
Faculty Assistant: Danielle Kahikina (tel: 512-475-8691;
danielle.kahikina@austin.utexas.edu)

COURSE SCHEDULE

17 Jan - 2 May, Wednesday from 9 am - 12 pm. Room: SRH 3.122 (LBJ School)

COURSE DESCRIPTION

Large datasets are increasingly becoming available across many sectors such as healthcare, energy, and online markets. This course focuses on methods that allow learning from such datasets to uncover underlying relationships and patterns in the data. The course starts with a review of basic statistical concepts and linear regression. But the course will focus mostly on *classification and clustering based on non-regression techniques* such as tree-based approaches, support vector machines, and unsupervised learning. This course is intended for first and second year Masters students. Ph.D. students with an interest in non-regression based quantitative methods may also find this course useful.

Topics to be covered: See Course Outline on page 4.

In covering the material from the assigned textbook (see below), this course will emphasize both on formulaic and conceptual understanding of the discussed methods. As necessary, the instructor will draw on material from outside the textbook for driving conceptual clarity.

PREREQUISITES

Basic grasp of linear regression would be helpful. However, all relevant concepts will be reviewed in class. Problem sets will include applied problems, often but not always from the textbook. As needed, you may use either Python or R for solving the problems. In the beginning, the instructor and TA(s) will conduct optional lab sessions in R and Python to provide the necessary background and toolsets in R and Python that will be necessary in solving the problem sets. Timing for the lab sessions will be set after the first class.

REQUIREMENTS

1. Required readings: Students are expected to complete the required readings (when assigned) each week *prior* to the class meeting for the unit and to contribute to the class discussion.

2. Problem sets: There will be 5 problem sets (PS). Assignments will be announced via email and posted on Canvas. The due date of each assignment will be noted on the assignment. Typically, problem sets will be due before 9 am on the next Wednesday following the day when a problem set is assigned, i.e., the beginning of the next class unit. Submit your work (except your code) in hard copies at the beginning of the class. Submit your code via Canvas.

3. In-class tests: Two 1-hour tests on 3/7 and 4/18.

4. Pop-quizzes: There will be four pop-quizzes (i.e., dates will not be announced in advance), with the objective to test grasp of material introduced in the class. Best three scores will count.

5. Projects: Each student will work in a group of 3 to 5 students on a term project to develop a predictive modeling analysis using real-world dataset(s). A project proposal will be due in late February (date TBD). Each team will prepare a brief report (due on the last day of class) and deliver a short (~15 min) presentation in class.

REQUIRED READINGS

We will use material from the following textbook:

An Introduction to Statistical Learning, by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. A free pdf of the text and all associated datasets are available at: http://www-bcf.usc.edu/~gareth/**ISL**/

The Course Outline below provides further details and schedule of the specific material to be covered from the book.

Additional required readings: Relevant material from the textbook for each unit may be augmented with 2-3 additional readings, typically highly accessible and relevant journal articles selected by the instructor. These readings will be posted on Canvas at least one week prior to the unit when these readings will be discussed. This will happen only for a few select units.

GRADING

Final grades will be determined by the following formula:

1. Class participation	5%
2. Assignments (5)	20%
3. In-class tests (3/7 and 4/18)	30%
4. Project[1]	30%
5. Pop-quizzes (best 3 of 4)	15%

[1] All students in the same project group will, by default, get the same grade for the project. However, if a group feels that this may be unfair for their project (for example, because of greatly different workload/contribution), the group must let me know by the last class day (2 May).

SPECIAL NEEDS

The University of Texas at Austin provides upon request appropriate academic accommodation for qualified students with disabilities. To determine if you qualify, please contact the Dean of Students at 417-6259; 471-4641 TTY. If they certify your needs, we will work with you to make appropriate arrangements.

STUDENT CONDUCT AND ACADEMIC INTEGRITY

Please see detailed information provided by the Office of the Dean of Students:
http://deanofstudents.utexas.edu/conduct/

ACCOMMODATIONS FOR RELIGIOUS HOLIDAYS

By UT Austin policy, you must notify me of your pending absence at least fourteen days prior to the date of observance of a religious holy day. If you must miss a class or a work assignment in order to observe a religious holy day, you will be given an opportunity to complete the missed work within a reasonable time after the absence.

COURSE OUTLINE[2]

(ChX = Chapter X in Textbook *ISL*[3])

Unit 1: Course Overview and Introduction (Jan 17)
- Ch1 + Other (outside textbook) introductory material

Unit 2: Statistical Learning (Jan 24)
- Ch2 + Other relevant basics

Unit 3: Linear Regression - 1 (Jan 31)
- Ch3.1-3.2

Unit 4: Linear Regression - 2 (Feb 7)
- Ch3.3-3.5

Unit 5: Classification - 1 (Feb 14)
- Ch4.1-4.3

Unit 6: Classification - 2 (Feb 21)
- Ch4.4-4.5

Unit 7: Resampling Methods (Feb 28)
- Ch5

In-class test #1 + Project discussion (Mar 7)

Unit 8: Linear Model Selection and Regularization - 1 (Mar 21)
- Ch6.1-6.2

Unit 9: Linear Model Selection and Regularization - 2 (Mar 28)
- Ch6.3-6.4

Unit 10: Tree-Based Methods (Apr 4)
- Ch8

Unit 11: Support Vector Machines – 1 (Apr 11)
- Ch9.1-9.2

Unit 12: Support Vector Machines – 2 (Apr 18)
- Ch9.3-9.5
- *In-class test #2*

Unit 13: Unsupervised Learning (Apr 25)
- Ch10

Unit 14: Project presentations (May 2)

[2] As noted under 'Required Readings', some of the units might also have 2-3 assigned papers as required readings. The papers will be posted on Canvas a week prior to the class discussion on those readings.
[3] ISL: *An Introduction to Statistical Learning*