SUNY Albany

AMAT565 Fall 2022

Applied Statistics for Data Science

Instructor: Felix Ye

Lectures: TTh 1:30PM-2:50PM in ES 146.

Instructor Office Hours: T 12:30PM-1:20PM, Th 3pm-4pm

E-Mail Address: xye2@albany.edu
Email will be a major line of communication between the student and the instructor. I will send urgent announcements and important information via email. Please check your university email regularly.

Web Page: Check the course page in blackboard regularly. Homework assignments, course announcements, and grades will be posted there.

Course Description: A course in statistical methods for students with some knowledge of statistics. Topics include multiple regression, analysis of variance and nonparametric statistical techniques. Emphasis on data analysis and statistical methodology. This course is about the linear model. It is mainly a course about applied statistics, using the linear model to illustrate important concepts.

This course is significantly harder than AMAT 554, so students should expect to devote substantial time every week on this course. As a remark, students are expected to devote at least 20 hours on each homework set.

In addition, the students will practice basic programming skills to use software tools in applied statistics. The programming language in this course will be Jupyter notebook.

All latest notebook can be downloaded in https://github.com/yexf308/AppliedStatistics.git. I will keep updating this git repository as class progresses, so please fetch the update regularly. I will also leave a copy (not most updated) version in blackboard.

Prerequisite: An introductory course in probability or statistics AND AMAT 502.

Textbook: The elements of statistical learning
You can download the latest version from: https://hastie.su.domains/ElemStatLearn/

Probabilistic Machine Learning: An Introduction
You can download the latest version: https://probml.github.io/pml-book/book1.html. I will loosely follow this book, however, I will put some additional material into it.

Grading Policy:

Homework	100%

Incomplete grade: This class will not give any incomplete grade. If the work cannot complete in the current semester, the student can choose to retake this class in the following year.

Homework: Homework assignments will be assigned every three weeks. There are 5 sets of homework in total.

Late assignment turn-in is not permitted. Any assignment turned in after the deadline will NOT be graded.

Attendance: Although attendance will not be taken, I strongly encourage you attend and participate in every lecture. This is one of the best ways to ensure success in the course.

Academic Misconduct: The strength of the university depends on academic and personal integrity. In this course, you must be honest and truthful. Ethical violations include cheating on exams, plagiarism, reuse of assignments, improper use of the Internet and electronic devices, unauthorized collaboration, alteration of graded assignments, forgery and falsification, lying, facilitating academic dishonesty, and unfair competition.

In addition, specific ethics guidelines for this course are as follows: Students may discuss homework. However, all solutions MUST be written up and submitted individually. The same rules apply to computer programs. Basic ideas may be discussed but detailed codes should not be copied or shared. Finally, exams must represent the result of individual effort and communication is permitted only with the instructor.

Report any violations you witness to the instructor. You may consult the associate dean of student affairs and/or the chairman of the Ethics Board beforehand.

Tentative Course Outline and Schedule:

- Week 1 (Aug 23 & Aug 25): Overview. Review of univariate models. Homework 1 assigned.
- Week 2 (Aug 30 & Sep 1): Multivariate models: Multivariate Gaussians.
- Week 3 (Sep 6& Sep 8): MLE for Gaussian and for linear regression. Homework 1 due. Homework 2 assigned.

- Week 4 (Sep 13 & Sep 15): Linear regression.
- Week 5 (Sep 20 & Sep 22): Cross-validation and Bootstrap.
- Week 6 (Sep 27 & Sep 29): Linear Model Selection and Regularization.
- Week 7 (Oct 4 & Oct 6): Linear Model Selection and Regularization. Homework 2 due. Homework 3 assigned.
- Week 8 (Oct 13): Robust linear regression.
- Week 9 (Oct 18 & Oct 20): Generalized linear models. Homework 3 due. Homework 4 assigned.
- Week 10 (Oct 25 & Oct 27): Mixture models and EM algorithm.
- Week 11 (Nov 1 & Nov 3): More on EM algorithm.
- Week 12 (Nov 8 & Nov 10): Tree-based models: Decision Trees. Homework 4 due. Homework 5 assigned.
- Week 13 (Nov 15 & Nov 17): Bagging, Random Forests.
- Week 14 (Nov 22): Boosting.
- Week 15 (Nov 29 & Dec 1): Multiple Testing. Homework 5 due.