QTM 385 - Introduction to Statistical Learning

Spring 2022

Instructor: Kevin McAlister

Email: kevin.mcalister@emory.edu

Synchronous Session/Lecture Time: W 6:00 PM - 9:00 PM EST in PAIS 220

Synchronous Sessions: https://emory.zoom.us/j/95246416756

Office Hours: Mondays 10:00 - 11:15 AM Online, Thursdays 4:00 - 5:15 PM in PAIS 579

Office Hour Sessions: https://emory.zoom.us/j/97594430265

Course Description: This course is designed to introduce students to the field of statistical learning, an essential toolset for making sense of the vast and complex data sets that have emerged in fields ranging from biology to finance to marketing to astrophysics in the past twenty years. This class will present a number of important modeling and prediction techniques that are staples in the fields of machine learning, artificial intelligence, and data science (more broadly). Unlike a standalone machine learning class, special attention will be given to the statistical underpinnings of common methods. This class will consist of 4 parts:

1) A review of probability theory and an introduction to maximum likelihood estimation/loss minimization: probability theory review, Bayes' rule, maximum likelihood estimation, model fitting, model comparison, predictive accuracy, overfitting

2) Regression as a predictive task and general model fitting: review of linear regression, cross validation and leave-one-out cross validation, bootstrapping, alternative prediction methods for continuous outcomes, sparse regression (Ridge, LASSO, and Elastic Net), non-linear methods, tree-based methods

3) Classification methods: K-nearest neighbors, naïve Bayes classifiers, linear discriminants, logistic regression, support vector machines, deep learning

4) Unsupervised methods: Principal components analysis, clustering (generative and geometric), hierarchical clustering, factor analyzers, semi-supervised methods

A Note on The Start of the Semester: We're starting the semester online. This is not ideal in any way, shape, or form. Since this class is designed to be a 3 hour session per week with a significant collaborative component, I have reworked the front end of the class to minimize the number of assignments in the first three weeks. Instead of meeting for the full three hours in the remote setting, we'll aim to meet for 2 hour sessions the first three weeks. Once we're back in person, the semester will proceed as the syllabus outlines.

Teaching and Learning During the Pandemic: Due to the unusual nature of the semester, communication is important. I will try my very best to respond to emails and questions posted to Canvas within 48 hours. In order to facilitate efficient communication, I ask that you check Canvas and the syllabus before asking a question. If you have a question of a personal nature then please email me. I will likely be slower on weekends and it is usually not a great idea to ask questions on a Friday night or right before something is due.

If your situation changes regarding health, housing, or in any other regard with respect to your ability to participate in the class, please contact the appropriate Emory student support organization first and then me as soon as feasible. It is easier for me to address your needs if I know about them as soon as they arise. This does not mean I can successfully respond to every request for consideration, but I emphasize that my goal is to treat you all equitably and do what I can to help you succeed in this course.

Though attendance and participation in lectures is expected, there is no formal grade associated with these aspects of the class. That said, your previous experience should have shown a strong positive correlation between a student's final grade and active participation in the lectures.

If you are not feeling well, please do not attend lectures in person! If you are sick, understand that I will be flexible about attendance and keeping up with work. If you expect that illness or other circumstances will prevent you from attending more than a single lecture, please make sure to email me so that we can discuss your individual circumstances. In order to remove the incentive for students to attend lectures while sick, I will be simulcasting the lecture via the class Zoom room during the regular lecture time. I will also be recording each lecture and posting the recording to the course's Canvas site. Together, these two options should remove the incentive to come to class while sick.

However, these recordings are not intended to be a substitute for the actual lectures! Much of the content in this course centers on discussion and group problem solving that occurs during the lab portion of the lectures. There will be no efforts made to record this part of the class. If you cannot attend lectures due to illness, let me know and I can try to set up a remote option for you to join in with other students while completing the lab exercises. If I determine that attendance has dipped below an acceptable level, I will stop posting the recordings and distribute them to students by request only. If everyone uses these recordings responsibly (as I expect will be the case), then this common good will benefit the class greatly.

Textbooks: There will be two main text resources for this course:

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning, 2nd Edition - ISL 2nd edition is available from the authors for free here
- Friedman, J., Hastie, T., & Tibshirani, R. (2017). The Elements of Statistical Learning, 2nd Edition - ESL 2nd edition available from the authors for free here

Over the course of the semester, we will also be using a number of other resources that will posted to the course's Canvas site.

The reading for each week of the course is strongly recommended and should be completed at some point. However, there is no need to rush to complete each week's reading prior to the class meetings. There will be a few times over the semester where readings will be referenced heavily in the course discussions and I will make it clear in those weeks that they should be completed prior to the next course meeting.

Beyond ISL and ESL, there are numerous other references on the topics covered in this course that are freely available online. In many cases, these sites/books do a better job of explaining the material than the chosen textbooks. I highly recommend seeking out these materials to better your understanding of the topics in this course. If you find anything particularly good, please send them my way so that I can post them for all the students in the course!

Prerequisites: As stated in the course listing, students must have completed all required QTM core courses before taking this class. This course will heavily leverage tools and topics in calculus and linear algebra, so competence with these topics is required. There is also an assumption that students have extensive exposure to probability, statistical inference, and regression methods. Students must also have some exposure to programming in `R`, `Python`, or another numerical computing language. While coding expertise is not required (you only get better with practice and this class will provide a lot of good directed practice!), familiarity with basic programming concepts is needed. I will be working mostly in `R`, but all assignments and projects can be completed in any appropriate programming language.

Course Structure: This course will be made up of lectures/discussions, a prerequisite refresher assignment (Problem Set #1), weekly homework assignments, two midterm data analysis projects, and a final project.

1. **Synchronous Lectures** - There will be no formal attendance grade, but attendance is expected and strongly encouraged. The textbook and suggested readings are great resources for learning the materials, but there is no substitute for the face-to-face interactions that happen in class.

   I have allocated 3 hours once per week to our in-person lectures. This is a *long* lecture period, but this is intentional - the goal of each lecture session is to only spend 1.5-2 hours actually covering new materials while treating the rest of the period as a lab session. This lab session will see students working together on various problems related to that week's topics. The weekly lab session problems **will be the weekly homework**! I've set up the class in this way to promote collaboration and to provide an avenue for guided practice and implementation of the course's materials. My hope is that the additional time needed outside of class to complete the assignments will only be an hour or two, on average.

2. **Problem Sets** - As stated above, there will be frequent problem sets **that will account for 30% of you final grade**. In total, I expect that there will be 6-8 problem sets (this is given as a range because the number may change depending on the flow of the class and how long we're remote). Problem sets will contain a few exercises that are related to the previous week's content. Problem sets may contain derivations and proofs, coding exercises, or exercises that apply methods to real world data - most will contain a combination of all three.

   Markdown or a similar typesetting method should be used for problem set solutions submissions. In `R`, RMarkdown is the best option. A short introduction to RMarkdown can be found here. For `Python`, Rmarkdown can be used with the `reticulate` package from RStudio or an equivalent markdown style submission can be made with a `.ipynb` file using Jupyter Notebooks.

   All homework submissions should include, at a minimum, a `.Rmd` or `.ipynb` file and a rendered HTML (recommended) or PDF version of the markdown file.[1] If any external libraries are used that aren't explicitly outlined at the top of each problem set, be sure to include the package `.tar.gz.` file in a `.zip` archive that is included with your problem set submission. Any non-base packages should be explicitly specified in a code block at the top of your write-up. There will be a times throughout the semester where we will explicitly use packages and implementations of algorithms, but the majority of your assignments will be completed with unique self-written code - you will learn much more coding it yourself than using a pre-built package.

   Each student must submit a solutions document using the above format. However, **students may complete any homework assignment (unless otherwise stated) in up to groups of 3. At the top of your solutions, please be sure to include the names of all students who worked together on the assignment.** If I find that there are identical solutions to any or all of the problems that are turned in and your collaborators are not specified at the top of the document, I will treat the submission as fraudulent and each student will receive a 0 for the assignment. You can work with the same students all semester (recommended), different groups, or complete the assignments individually. If you would like to work with other students but find yourself having a hard time finding others to work with, let me know and I'll try to place you with other students.

3. **Prerequisite Refresher/Problem Set #1** - The first problem set will be posted the first day of class and will be due 1.5 weeks later. Unlike other problem sets, **this assignment will be worth 10% of your final grade and must be completed on your own!** This problem set is included as a different item on this syllabus because all of the problems in this assignment should be solvable by a student with the appropriate prerequisite coursework and background in statistical programming. While not everything may be immediately obvious, a student who is well-prepared for this class should find that no single problem is too difficult or too difficult to troubleshoot. This problem set will cover topics ranging from calculus, matrix manipulations/calculus, probability and statistics, and basic statistical programming.

[1] Instruction for converting `.ipynb` files to HTML and PDF are outlined here.

The purpose of this initial assignment is not to scare you/weed out students/decrease class size. Rather, it is intended to provide you and I with a credible commitment mechanism for assessing whether you have the necessary prerequisite knowledge to thrive in this class. While I would love to teach and work with each and every student who comes through the QTM department, I feel as if it would be a disservice to both you and the other students in the class if you begin well behind the level of knowledge and experience that I expect is necessary to learn from and enjoy this class. My expectation is that each and every one of you will excel on this assignment and receive full credit!

4. **Midterm Applied Data Exercises** - Over the course of the semester, there will be two larger scale applied data exercises **will account for 30% (15% each) of your final grade**. These exercises - one focusing on predictive models for continuous outcomes and one focusing on predictive classification problems - will contain novel data sets and questions that you (and your team, should you so choose) will try to answer using the methods discussed in class. These assignments will have less guidance than a standard problem set and will require you to try multiple methods to determine the best approach to answering the question. **There will be no single correct answer for these assignments!** Much like real-world data analysis projects, there are pros and cons to any choices. Therefore, these assignments will see you make a decision as to the best approach and provide a write-up that discusses what you did, why you chose that approach, and what weaknesses your chosen approach has.

   As with the problem sets, each student must turn in a copy of the final write-up. Similarly, students may complete this assignment in up to groups of 3. The same rules that apply to problem sets will apply for these assignments.

5. **Final project** - Each student will also be tasked with completing a final project that **will account for the other 30% of your final grade**. As with the problem sets, each student must turn in a copy of the final write-up. Similarly, students may complete this assignment in up to groups of 3. The same rules that apply to problem sets will apply for the final project.

   This project should apply the tools covered in this course to a novel problem that is of interest to the student (or group). This can be analysis of a meaningful data set, a comparison of methods, development of a new method (methodological or computational), writing a package, etc. The only rule is that the topic must be approved by the instructor (i.e. me). You/your group should be thinking about this project early and often. In order to create a number of meaningful checkpoints, the final project has 3 parts:

   (a) **Initial Project Discussion (5%):** Before March 25th, 2022, each student should select a project group and meet with me outside of class at least once to discuss their initial project ideas. This meeting can be used to get assistance with ideas or to let me know your final project topic. The goal of this meeting (or meetings, if need be) is to get approval on your final project topic.

   (b) **One-page project progress report (5%):** Before April 15th, 2022, each student should submit a one-page project progress report that outlines your final project progress. This proposal should include your main question, data being used, methods, etc. It should also include, at a minimum, evidence that you have begun the main portion of your project (e.g. preliminary analysis, a proof, etc.). I will provide comments as needed for each submission. At this stage, there should be no projects that do not meet course standards.

   (c) **Final project report (20%):** Before May 3rd, 2022 at 5:00 PM EST, each student should submit a final project report. **This deadline is as late as possible in the semester and cannot be further extended!** Depending on the final project, the final deliverable may differ. However, I expect that most final reports should be written like an academic paper and report the question/motivation, methods, and findings. Any references should be properly cited. Along with the report, replication files for any computational work should be submitted. These files should allow a researcher that does not have your data to reproduce your main findings (figures, tables, important quantities, etc.). Replication files can be submitted as a `.zip` archive or they can be uploaded to a Github project page. Final projects will be graded based on the quality of

the work rather than the novelty of the findings - a strong attempt with null results is preferred to a low-quality paper with "surprising" results.

Late Assignments: Over the course of the semester, each student has 5 "late days" that can be used to turn in problem sets or applied data exercises after the due date. These late days are no questions asked, no notification needed, no grade penalty days to give students a grace period for late work over the semester. Late days are rounded up to the nearest day (an assignment due at 11:59 PM turned in at 12:01 AM on the next day counts as one late day). Assignments that are turned in late with multiple group members count as late days for all students in the group.

Should the situation arise where all late days are used by a student, then any assignment turned in less than 7 days after the due date late will receive 50% credit. No assignments will be accepted more than 1 week after the original due date. However, I recognize that there is a lot going on in the world right now and that each of us has our own unique set of things to deal with. If you find that you are unable to complete the course assignments on time, please reach out to me so we can come up something that is both fair and amenable to your unique situation.

Note that late days cannot be used for the final project and related checkpoints. I have provided ample time for each group to meet these deadlines, so all late final project assignments will be granted 50% credit.

Final Grades: As stated above (and stated again for emphasis), your final grade is made up of 4 components: problem sets (30%), Problem Set #1 (10%), Midterm Data Analysis exercises (30%) and the final project (30%). Final grades will be determined using the following (estimated) grade ranges:

- **A**: $93\% - 100\%$
- **A-**: $90\% - 92\%$
- **B+**: $87\% - 89\%$
- **B**: $83\% - 86\%$
- **B-**: $80\% - 82\%$
- **C+ and below**: $< 80\%$

As assignment scores are computed and I get more of a feel for the average grades on each assignment, the grading rubric may be updated. However, the grade scale will only be loosened from this initial rubric (e.g. if your final average is 93%, then you will get an A for the course regardless of rubric changes). Final grade percentages will be rounded to the nearest integer. There is no curve for final course grades - if everyone in the class scores high enough to get an A, then everyone will get an A. Hence, there is no "competition" for grades.

Grade Appeals: If you believe that your grade on any assignment is incorrect or unfair, you should submit your concerns, in writing, to me. The written appeal should fully summarize what you believe the problems are and why. Unless the appeal regards a simple addition error, please wait 48 hours before submitting a written appeal.

Accessibility and Accommodations: As the instructor of this course, I endeavor to provide an inclusive learning environment. I want every student to succeed. The Department of Accessibility Services (DAS) works with students who have disabilities to provide reasonable accommodations. It is your responsibility to request accommodations. In order to receive consideration for reasonable accommodations, you must register with the DAS at http://accessibility.emory.edu/students/. Accommodations cannot be retroactively

applied so you need to contact DAS as early as possible and contact me as early as possible in the semester to discuss the plan for implementation of your accommodations. For additional information about accessibility and accommodations, please contact the Department of Accessibility Services at (404) 727-9877 or accessibility@emory.edu.

Tentative Course Schedule:

Note: The current schedule is to post assignments on Wednesday before class and have them be due the next Friday. Midterm data analysis exercises will be given an extra week for completion.

Week 1: January 12th

Topics: Introductory Comments; Syllabus Review; The role of statistics in machine learning; What is statistical learning?

Assignments: Problem Set 1 Posted

Week 2: January 19th

Topics: Review of linear regression; loss functions; in-sample vs. out-of-sample predictive ability

Week 3: January 26th

Topics: Resampling methods - bootstrapping and cross-validation; LOOCV; k-fold cross validation

Assignments: Problem Set 2 Posted

Week 4: February 2nd

Topics: Large-p regressions and variable selection; regularization approaches (Ridge, LASSO, and ElasticNet)

Assignments: Problem Set 3 Posted

Week 5: February 9th

Topics: Moving beyond linearity; k-nearest neighbors regression; polynomial regression; splines; generalized additive models (maybe?)

Assignments: Problem Set 4 Posted

Week 5: February 16th

Topics: Introduction to Generalized Linear Models (with special attention paid to logistic regression)

Assignments: Midterm Data Analysis Exercise 1 posted

Week 6: February 23rd

Topics: Intro to Classification; k-nearest neighbors classification; naive Bayes' classifiers

Week 7: March 2nd

Topics: Classification Part 2; linear discriminant analysis; logistic regression; Support Vector Machines

Assignments: Problem Set 5 Posted

Week 8: March 9th

Topics: **No Class - Spring Break**

Week 9: March 16th

Topics: Support Vector Machines, Part 2; Ensemble Learning

Assignments: Problem Set 6 Posted

Week 10: March 23rd

Topics: Tree-based approaches; random forests; classification trees; bagging; boosting

Assignments: Midterm Data Analysis Exercise 2 Posted

Final Project Checkpoint: All students must meet with me outside of class at least once to discuss project group and plan before March 25th at 5:00 PM.

Week 11: March 30th

Topics: Unsupervised Learning; clustering methods; k-means clustering; generative clustering; hierarchical clustering (maybe?)

Week 12: April 6th

Topics: Dimensionality Reduction and Latent Traits - principal components analysis; factor analyzers; principal component regression; probabilistic PCA

Assignments: Problem Set 7 Posted

Week 13: April 13th

Topics: Matrix Factorization and Covariance Decomposition - non-negative matrix factorization; multidimensional scaling; matrix completion

Final Project Checkpoint: All students must submit a 1 page final project progress report by April 15th at 5:00 PM.

Assignments: Problem Set 8 Posted

Week 14: April 20th

Topics: Semi-supervised learning and where to go from here

Week 15+: The Postseason

Topics: Each student should submit the final project deliverable before May 3rd at 5:00 PM.