Syllabus for ORIE 4740: Statistical Data Mining I

Damek Davis

Basic Info

Lectures, schedule, and course staff

Online Lectures: TR 1:00PM–2:15

Tentative Schedule: Available on canvas (check for updates regularly).

Instructor: Damek Davis (dsd95 at cornell dot edu, Rhodes 218)

TAs:

1. **Lead TA**: Tao Jiang (tj293)
2. Kevin Jiang (kcj42)
3. Tonghua Tian (tt543)
4. Duanduan Zhu (dz223)

Office hours: See Canvas Module

Topics

“[Data mining is] the process of discovering meaningful correlations, patterns, and trends by sifting through large amounts of data… It employs pattern recognition technologies, as well as statistical and mathematical techniques.” (The Gartner Group).

Data mining often involves datasets with many records and many variables. Frequently little is known about the distribution of any particular variable, or about the relationships between variables. Desirable approaches have few assumptions or are robust to the violation of those assumptions. They also must be computationally tractable on large data sets. By the end of this course, you will be able to take a large commercial or governmental data set, decide on data mining techniques to answer our question of interest, apply those techniques, compare them, and draw conclusions. In order to cement your understanding you will implement some techniques, and modify or apply implementations of some more complex techniques.

We will cover most of the chapters in *ISLR*, including (tentatively): Linear Regression, Classification, Dimensionality Reduction/PCA, Clustering, Nonlinear Methods, Decision Trees and Random Forests, Support Vector Machines, Model Validation/Selection and Regularization.

Email Policy

Because this is a large class, I will typically not respond to emails from students. Instead, direct all your questions regarding the course to our class Ed discussions forum (ED); see canvas for a link. If you need to discuss a personal matter, please see me during my office hours. If you’re having trouble reaching me, please contact your TA–preferably in person–and they can tell you whether it is necessary to meet with me in person.

Prerequisites

- **ORIE 2700 and 3500 (statistics and probability) or equivalent:** Marginal probability, joint probability, conditional probability, Bayes’ theorem, multivariate Normal distributions, mean and variance. Point and interval estimation, hypothesis testing, p-values. Simple linear regression.

- **Math 2940 (linear algebra) or equivalent:** Matrix/vector notation and operations, eigenvalues and eigenvectors, eigen and singular value decompositions, inverse, trace, norms.
- **Programming experience** in R, Python, Matlab, C or Java.
- Strongly recommended: Background in multiple linear regression and logistic regression (this will be taught, but prior knowledge would help).

Textbooks

- **Required:** *An Introduction to Statistical Learning (ISLR)* by James, Witten, Hastie and Tibshirani. A pdf of the book is available for free from the authors' web page.
- **Required: i>clicker or a REEF polling compatible device:** You must sign up for an i>clicker student app subscription at https://www.iclicker.com/school/cornell-university. Participation in the polling will count towards your participation points.
- Optional:
  - *Data Mining for Business Intelligence*: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner by Shmueli, Patel, and Bruce. Second Ed., 2010. (book web page)
  - *The Elements of Statistical Learning*: Data Mining, Inference, and Prediction by Hastie, Tibshirani and Friedman. Second Ed., 2009. This book is the advanced version of ISLR. Freely available here.

Websites

- **Canvas:** We use Canvas for all course materials and communication. You should be automatically given access to the course Canvas site when you enroll in the course
- **Gradescope:** We use Gradescope for all course grading. You should automatically have access to gradescope when you enroll in the course.
- **ED:** We will have a class ED forum where students can post and answer questions about the content. ED participation will count towards your participation points. You can find the link to ED on the course Canvas site.
  - The instructor and TAs will monitor the forum, but it is primarily the responsibility of other students to help each other.
  - The goal for such a forum is to encourage learning through peers; knowing how your peers are struggling with a problem can be useful in your learning. Answering your peers' questions can help identify any gaps in your understanding. This will not be achieved if the questions/comments are only seen by the instructor/TAs. If you believe your question or a comment will be useful to the entire class, make it public. This should generally be the default post type. I may make a question public if I think it's the more appropriate option.
  - **Only rarely would the private option be appropriate.** An example where a private note would make sense is in a situation where you are far along answering an assignment question and are unsure about some aspect of your approach.
  - You can keep yourself anonymous from your peers if you want; I will still see your identity.

Academic Integrity

Each student in this course is expected to abide by the Cornell University Code of Academic Integrity. Any work submitted by a student in this course for academic credit will be the student's own work. See above for the policy regrading homework. The Code is available here.

Deliverables and Grading

Grading

- **Homework & labs:** 40%
- **Exams**: 30%
- **Project**: 22%
- **Participation**: 3%.
- **Quizzes**: 5%.

These weights are approximate; we reserve the right to change them later.

Grades for each homework/lab/exam/quiz etc will be posted on Canvas and gradescope. Regrade requests must be submitted via Gradescope within one week of when the grade is posted. Questions on the exam grades must be submitted in writing to **Tao** within one week of when the grade is posted.

Exams

Requests for special accommodation must be made at least 2 weeks prior to each exam. No final exam.

Dates

- **Prelim 1: Tuesday March 22nd (evening)**
- **Prelim 2: Thursday April 28th (evening)**

Labs

The discussion sessions will be a combination of recitations and computer labs on using R. They will be held in person, weekly unless notified otherwise. The room is equipped with computers, but you are encouraged to bring your own laptop. Each student should register for one of the discussions.

You will need to submit your work for each lab similarly to a homework assignment.

TAs will be responsible for holding the labs. Lab participation is crucial to prepare you for the final project. Questions are best addressed during office hours and labs, or on ED (instead of email).

R is freely available here. You may consider using the RStudio environment.

Homework

There will be about 9 labs and homework assignments in total. **Homework/lab is due at 11:59pm on Friday a week after it is given out (unless specified otherwise)**, and must be submitted electronically through Gradescope, NOT to the homework dropboxes or by email, under door, etc.

Out of all your homework/lab grades, the 2 lowest ones will be dropped; this accommodates sickness, family emergency, religious holiday or other circumstances without a formal process. If you miss an assignment for these reasons then it must count as the dropped assignment.

On Canvas, you can find detailed submission instructions and policies on dropped/late/regrading homework/labs.

You may discuss the content of the course with other students in your 4740 class, but you must complete your homework/lab independently and individually.

Participation

Students are expected to submit and answer questions on ED and fill out the course evaluation. In addition, students must participate synchronously or asynchronously in the following fashions.

- **Synchronous** participation points are awarded to students who attend the lecture and answer at least 75% of the iClicker questions asked during the class; responses are graded on completion. Make sure that you have purchased an iClicker subscription and have joined the course via the iClicker app – I suggest downloading the app to your phone so that you do not have to navigate away from zoom to answer questions.
- **Asynchronous** participation points may be awarded if you fill out this participation form before the beginning of the following lecture. In contrast to the synchronous option, the questions on the asynchronous form will be graded based on completion.

Note: you need only use one of the above options. In addition, **if you fill out the course survey at the end of the class, then we allow you to miss up to 4 days without losing participation.**

Guidelines for receiving full credit via asynchronous participation

1. **Submit on time:** Submit your participation form before the beginning of the next class! Responses are time stamped and answers will not be graded if the submission is late.

2. **Graded on Correctness:** We grade based on correctness, though multiple answers may be correct. Use only comma-separated letters A,B, etc depending on the question. Do not include any other symbol or text. Make sure that all the questions are answered. (Note that there's no need to use the iClicker app if you're submitting the async form.)

3. **Summarize well:** A good summary captures the key takeaways from the lecture and **should be more than just an outline of the lecture.** We ask you to summarize to help you learn and review the material. Poor summaries will receive poor grades.

Here is an example of a good summary:

> This lecture provided a broad overview of the course. We introduced a basic definition of the term “data mining” and went over several relevant applications of data mining in business, medicine, and information technology. Then we discussed two different types of learning: supervised and unsupervised learning. We answered several iclicker questions that elucidated the difference between supervised and unsupervised learning. Then we finished the lecture with an application of supervised learning in heart disease detection.

4. **Take the comment/question seriously:** One word comments (like “great”) will not receive points. Some helpful questions to get you thinking: What did you find exciting? What do you want to explore further? Did you find anything confusing? What more do you want to learn about the topics covered?

Quizzes

Students are expected to complete an online quiz once a week. The quiz will be posted to our Canvas site. It is primarily a tool for you to test your understanding of the course concepts.

Final Project

In the final project, the techniques taught in the class are used to analyze a large dataset. Students work in teams of **2-4 students**. Each team finds the necessary data, carries out the project, and writes a project report.

Detailed projection information can be found in Section 12.

Important distinction between 5740 and 4740. Final projects completed by graduate students enrolled in ORIE 5740 should be business-oriented, using techniques from class to address a clear business problem. In addition to the other requirements for ORIE 4740, graduate students must also submit a final (video) presentation.

Detailed Information on Final Project

In the final project, the techniques taught in the class are used to analyze a large dataset chosen by the students. Students work in teams of **2-4 students**. Each team finds the necessary data, carries out the project, and writes a project report.

Due dates

- **Team and dataset**: Once you form your group and decide on which dataset to use, **email Tao (tj293) ASAP** with the names and NetID of your group members as well as the source of the dataset(s). The general rule is that **no two groups may use the same dataset.** In case of a conflict, the first group that emails Tao will have priority. On ED, Tao will maintain a list of datasets that have been chosen.
- (22% of final grade)
  - *Project proposal*: **April 1st (Friday) 11:59PM.**
  - *Peer project proposal evaluation*: **April 15th (Friday) 11:59PM.**
  - *Project report and peer team evaluation form*: **May 6th (Friday) 11:59PM.**
  - *Peer project evaluation*: **May 10th (Tuesday) 11:59PM.**
- The submission location will be announced at a later date.

Project Teams

You should work in a team of 2 to 4 students. Please try to form a team yourself; if you have trouble finding teammates then let me know and I will help you find a team. You may not work alone.

You may search for teammates on ED.

Project Assignment

You will apply tools that you have learned in 4740 to a dataset of your choice.

1. You may NOT use CMU Statlib
2. You may NOT use the UC Irvine Machine Learning Repository
3. Simulated (artificially generated) datasets are not allowed
4. You may NOT use a dataset used before in HW/labs, nor any of the datasets from ISLR.

You may obtain a dataset from a company, for instance if you have had an internship in a company and they are willing to provide you with such a dataset for this purpose. You may use a dataset from a research project at this university or another university, with permission. **I highly encourage you to look around for a dataset on a topic that particularly interests you**, rather than using generic datasets from data mining websites. Example: say I am interested in doing a project related to beer. A web search on “beer dataset” brings up a dataset with 1 million + beer reviews, from BeerAdvocate. You are allowed to use datasets from other textbooks, but you cannot do an analysis similar to that done in the textbook on the same dataset.

Here are a few data sources:

- Data.gov
- Stanford Large Network Dataset Collection
- Yelp dataset challenge
- Kaggle
- quandl
- “The 50 Best Public Datasets for Machine Learning”
- Call center data
- Marine environment data
- Datasets from the KDD Cup
- Various datasets from Yahoo!

Note that some datasets in the above links are **not** acceptable as per rules 1-4 above. Regardless of the source of the data, this source must be referenced in your report. If the dataset is not in the public domain, then you must obtain permission for its use in this class project. No two groups may use the same dataset; if two groups propose the same dataset by chance, the one that emails Tao first will have priority.

What is required? Each team must write a project proposal, find the necessary data, carry out the data analysis, and write a project report. The analysis should be motivated by one or two particular scientific/commercial goals, such as (for a fictional dataset consisting of veteran’s data): “We will predict two response variables. First we predict whether or not an individual will contribute donations to a veterans’ organization. Then, if they do contribute, we predict the amount of the contribution. We will make these predictions based on two sources of information: demographic characteristics and contribution history. This prediction can be used to choose which individuals receive solicitations, or to estimate the total expected contributions in order to guide the organization’s financial planning.”

The data analysis that you perform needs to be more than a direct mapping of one of our lab analyses to another dataset. You will need to use more than one of the approaches that we have learned in the class. An example of a data analysis with sufficient scope is:

> For a data set like the veteran’s data that has a continuous outcome variable and a binary outcome variable: Applying linear regression to predict the continuous outcome and applying logistic regression and decision trees to predict the binary outcome, while handling missing data. Comparing the results from logistic regression and decision trees, and recommending which should be used.

The data analyses that you perform should be appropriate for the goal(s) that you have stated. You should choose one or several data sets that are appropriate to address the goal(s) you have stated. If you have more than one data set or more than one goal, the project should form a coherent whole, rather than being two or three unrelated data analyses. For instance, you may have a single scientific goal, and use two data sets to address this goal. Or, you might have a single data set, with which you address two related scientific questions.

If you analyze a single data set then it should have a reasonably large number of observations **(at least 1000)**; otherwise, two smaller data sets suffice but they should each contain **at least 500 observations**. One of your data sets should have **at least five predictors**.

Sometimes a data analysis yields negative or inconclusive results. For instance, perhaps none of the predictors were significant in the model, even though they seemed like reasonable predictors. Perhaps the predictions were poor, and the methods chosen, although they were a reasonable choice and had good promise, turned out to not work well. These are acceptable results, as long as all of the analyses and conclusions are correct. You might in this case suggest alternative approaches in your conclusion.

Work on the project is to be done entirely by the project group; communication between groups regarding project work is not allowed. You may not apply a technique that has been previously applied to the same data set in a published or unpublished work, if you are aware or could reasonably be expected to be aware of the existence of their work. You should cite in the bibliography any and all published or unpublished written works or spoken communications that have influenced your analysis.

You should employ at least one technique covered in class/ISLR, but are free to use any additional methods beyond class. You are encouraged to use **R**, but using another language (such as Python) is allowed.

Project Proposal

The proposal should be one page (double-spaced, 11+ pt font) and include:

- The proposed scientific/commercial goal(s) of the analysis;
- The proposed data set(s) that will be used, including their source, number of variables and data points;
- The proposed data analyses to be performed;
- What figures or tables you might include;
- Why you expect the data set(s) and analysis methods to successfully address your goal;
- Any other details at your discretion

Final Report

The report should be no more than 8 pages + 1 page for bibliography (double-spaced, 11+ pt font) and should contain:

- Title page with authors and abstract
- Introduction telling what the project is about, what your team has accomplished, and a brief statement of results and conclusions.
- One or more sections describing the project
- Conclusions
- Bibliography

Tables and figures can be interspersed in the text or at the end of the report. All tables and figures should be numbered and referred to by number. The report should not contain raw computer output. Rather, any computer output should be in a table or figure, with explanation in the main text. Do not hand in the code (R, Python, etc.) for your analysis, but the instructor reserves the right to ask for your code if he deems it necessary.

If your report has more than 8 pages, then there is no guarantee that the extra pages will be read by the instructor or graders.

Peer team evaluation form

Each student is asked to fill out a peer evaluation form, which has been uploaded to gradescope and will assess each individual’s contribution to the group. This form is due the same day as the final project report.

Peer project and proposal evaluation

You will be required to provide an evaluation of another team’s project proposal and their report. The exact instructions will be uploaded to canvas, towards the middle of the semester.

Submission

We will upload instructions to canvas, towards the middle of the semester, that describe how to submit your project proposal, the final reports, and the peer project/proposal evaluations.

The final peer team evaluations, on the other hand, should be submitted electronically on Gradescope.

Each student will submit their own peer evaluation form. In contrast, only one member of each team needs to submit the project report.

Grading

Grades will be based on:

- Validity of the goal(s)
- Whether the data set(s) and data analyses selected are appropriate to address that goal
- Sufficient scope of the data analysis
- Comprehensiveness and validity of the conclusions
- Creativity
- Clarity and conciseness of the report. A wordy report will get a lower grade than one saying the same amount in less space.
- Number of students. Projects done by larger teams are expected to be more extensive.
- Individual’s contribution to the group, as assessed by peer team evaluations

Sample Project Reports

Can be found on Canvas. Please do not circulate these reports outside this class.

Note: These reports are not necessarily among the ones that received the highest grades in previous years, and may even contain errors and flaws. They simply give you a sense of the scope and structures of the project, as well as the possibility of techniques and outcomes.