MATH 390.4 (soon to be 342W) Spring 2020 Course Syllabus

Adam Kapelner, Ph.D.

Queens College, City University of New York

document last updated Monday 27th January, 2020 3:07pm

Instructor	Professor Adam Kapelner
Contact	`kapelner@qc.cuny.edu`
Lecture Time / Loc	Tues and Thurs 11AM - 12:50PM / KY 283
Required Lab Time / Loc	Friday 9-10:50AM KY 061
My Office Hours / Loc	Tues and Thurs 12:50-1:30PM / Kiely 283
TA Office Hours / Loc	TBA / Kiely 5th fl lounge
Course Homepage	https://github.com/kapelner/QC_Math_390.4_Spring_2020

Course Overview

MATH 342W. Data Science via Machine Learning and Statistical Modeling. 6 hr.; 4 cr. Prereq.: MATH 241 (intro to probablity and statistics), MATH 231 (intro to linear algebra), CSCI 111 (intro to programming) or equivalents. Recommended: ECON 382 (intro to econometrics) or equivalent, MATH 341 (Bayesian modeling), MATH 633 (statistical inference) or equivalent. Philosophy of modeling and learning using data. Prediction via the ordinary linear model including orthogonal projections, sum of squares identity, R^2 and RMSE. Polynomial and interaction regressions. Prediction with machine learning including neural nets (the perceptron), support vector machines and the tree methods CART, bagged trees and Random Forests. Probability estimation using logistic regression, asymmetric cost classifiers and the ROC / DET performance curves. Underfitting vs. overfitting and the bias-variance decomposition / tradeoff. Model validation including out of sample techniques such as cross validation and bootstrap validation. Correlation vs. causation, causal models, lurking variables and interpretations of linear model coefficients. Extrapolation. The `R` language will be taught formally from the ground and up (its use will be a substantial part of the

homework) as well as visualization using the `ggplot` library and manipulation using the `dplyr` and `data.table` libraries.

You should be familiar with the following before entering the class:

- Basic Probability Theory: axioms, conditional probability, in/dependence
- Modeling with discrete random variables: Bernoulli, Hypergeometric, Binomial, Poisson, Geometric, Negative Binomial, Uniform Discrete and others
- Expectation and variance
- Modeling with continuous random variables: Exponential, Uniform and Normal
- Frequentist confidence intervals and hypothesis testing for one-sample proportions
- Basic visualization of data: plots, histograms, bar charts
- Linear algebra: Vectors, matrices, rank, transpose
- Programming: basic data types, vectors, arrays, control flow (for, while, if, else), functions

We will review the above *throughout the semester* when needed and we will do so rapidly.

This is not your typical mathematics course. This course will do lots of modeling of real-world situations using data via the `R` statistical language.

The 650.4 section

You are the students taking this course as part of a masters degree in mathematics. Thus, there will be extra homework problems for you and you will be graded on a separate curve.

Course Materials

We will be using many reference texts and three popular books which you will read portions from. However the main materials are the course notes. You should always supplement concepts from class by reading up on them online; wikipedia I find the best for this.

Theory Reference: It is not necessary to have these two books, but it is recommended. The first is "Learning from Data: A Short Course" by Abu-Mostafa, Magdon-Ismael and Lin which can be purchased used on Amazon. We will also be using portions from "Deep Learning" by Goodfellow, Bengio and Courville that can be purchased on Amazon and read for free at `http://www.deeplearningbook.org/`.

Popular Books: We will also be reading the non-fiction novel "The Signal and the Noise" by Nate Silver which can also be purchased on Amazon. This is *required* — you will have homework questions directly from this book. We will also be reading "Predictive Analytics, Data Mining and Big Data" by Steven Finlay that can be purchased on Amazon and it is also available online from the Queens College library system.

Computer Software: You need your own personal computer, laptop preferred. We will be using `R` which is a free, open source statistical programming language and console available for all operating systems. Please download the latest version from: `http://cran.mirrors.hoobly.com/`. You will be expected to do programming. I recommend the IDE `RStudio` available for free at `https://www.rstudio.com/products/rstudio/download/`.

Source Control: You will be expected to use `git` and have a `github.com` account with a repository named `QC_MATH_342`. You will use this repository to submit coding homework assignments (and theory assignments if you use LaTeX).

Book on `R`: We will be making use of "R for Data Science" by Wickham and Grolemund which can be purchased on Amazon or read online at `http://r4ds.had.co.nz/`.

Announcements

Announcements will be made via email. I am known to send a couple emails per week on important issues. Thus, I will need the email address that you reliably check. The default is whatever is in CUNYfirst which many of you do not check. (See Homework #0 for more information).

Classes

There are 42 scheduled meetings. Of these, 25 will be lectures, 13 will be labs, 2 will be midterm exams which are in class and 2 will be review periods (the meeting before the exams). The exam schedule is given on page 8.

I am **canceling** Friday, April 17 (the Friday after the break which ends the day before) because many people will be away for the whole week returning the Monday after. We will do a makeup 2hr session at the end of the semester possibly during finals week to help with the projects.

Lectures

Lectures will be in KY258 and will besplit into two periods: theory and practice. The first is a standard chalkboard lectue where we learn concepts and the second will be

using the computer/projector to see the concepts in action in the `R` language. I have a no computer / tablet / phone policy during the theory component of the lectures (only pen / pencil and paper) but you are highly recommended to have the laptop during the second part.

Labs

Labs will be in KY061 (the computer lab) duiring the Friday morning meetings. Sometimes we will spend some of the two hours doing a practice lecture but the majority of this time will be your time. You will take turns "driving" the coding in front of the class, working on exercises that you will finish for homework. Thus we will spend most a lot of time talking through problem solving skills in data science.

Lecture Upload

As many previous students have noted, my handwritten notes are useful to me and not to many others. Thus, I will be rewarding students for taking notes, scanning them in and sending them to me. You will be rewarded in two ways: (1) if you do this for more than 10 lectures, you will be given the automatic 5 points (see grading policy on page 9) for your classroom participation grade and (2) you have the option for me to say your name publicly on the course homepage. Make sure you follow these instructions:

- You have *one week only* from the time of the lecture to email me lecture notes.
- There must be *one* file and it must be in PDF format only.
- The file must be <2MB. No exceptions.

If you send me your notes, you are (1) agreeing to the MIT license which means someone can freely copy your notes and even make money off of it (and not owe you a cent!) and (2) since github is mirrored, once your upload is on the web, it is there indeliby forever. **If you are not comfortable with these two points, do not send me your notes.**

Homework

Homework will be split into *theory* and *practice* (called "labs"). This course will be the "writing in the major course" next year. Thus, a portion of each theory and practice homework will involve writing *English* and being graded on *English*.

Theory Homework

There will be 4-7 theory homework assignments. Homeworks will be assigned and placed on the course homepage and will usually be due a week later in class. Homework will be **graded** out of 100 with extra credit getting scores possibly > 100. I will be doing the grading and will grade an *arbitrary subset of the assignment* which is determined after the homework is handed in. Homework must be printed, neat and stapled (**it cannot be emailed to me**) but if you use LaTeX it can be pushed to your own repository (see "Homework LaTeX bonus points" section below) and I will grade it digitally. Homework can be given to me in class or delivered under my office door (KY 604).

Graded homework will be returned in class. Regrades are handled during office hours or right after class is over. Scores for homeworks are finalized one week after the graded copies are handed back. Thereafter there will be no changes and no re-grading. Do not delay checking your graded homeworks. I am not perfect and I do make mistakes. It is your obligation to find these mistakes and report them.

You are encouraged to seek help from me if you have questions. After class and office hours are good times. **You are highly recommended to work with each other and help each other. You must, however, submit your own solutions, *with your own write-up* and in *your own words*. There can be no collaboration on the actual *writing*. Failure to comply will result in severe penalties.** The university honor code is something I take seriously and I send people to the Dean every semester for violations.

Practice Homework (Labs)

These will almost exclusively consist of short and medium coding exercises in `R`. Most of the assignment will be done for you and your peers during the Friday lab session.

Philosophy of Homework and Labs

Homework is the *most* important part of this course.[1] Success in Statistics and Mathematics courses comes from experience in working with and thinking about the concepts. It's kind of like weightlifting; you have to lift weights to build muscles. My job as an instructor is to provide assistance through your zone of proximal development. With me, you can grow more than you can alone. To this effect, homework problems are color coded **green** for easy, **yellow** for harder, **red** for challenging and **purple** for extra credit. You need to know how to do all the greens by yourself. If you've been to class and took notes, they are a joke. Yellows and reds: feel free to work with others. Only do extra credits if you have already finished the assignment. The "[Optional]" problems are for extra practice — highly recommended for exam study.

[1] In one student's observation, I give a "mind-blowing homework" every week.

Time Spent on Homework

This is a four credit course. Thus, the amount of work outside of the 4hr in-class time per week is 8-12 hours. I will aim for 10hr of homework per week on average. However, doing the homework well is your sole responsibility since I will make sure that by doing the homework you will study and understand the concepts in the lectures and you won't have all that much to do when the exams roll around.

Late Homework

Late homework will be penalized 10 points per business day (Monday–Friday save holidays) for a maximum of five days. *Do not ask for extensions*; just hand in the homework late. After five days, **you can hand it in whenever you want** until May 15, 2019 noon. As far as I know, this is one of the most lenient and flexible homework policies in college. I realize things come up. Do not abuse this policy; you will fall far, far behind.

LaTeX Homework Bonus Points

Part of good mathematics is its beautiful presentation. Thus, **there will be a 1–7 point bonus** added to your theory homework grade for typing up your homework using the LaTeX typesetting system based on the elegance of your presentation. The bonus is arbitrarily determined by me.

I recommend using overleaf to write up your homeworks (make sure you upload both the hw#.tex and the preamble.tex file). This has the advantage of (a) not having to install anything on your computer and thus not having to maintain your LaTeX installation (b) allowing easy collaboration with others (c) alway having a backup of your work since it's always on the cloud. If you insist to have LaTeX running on your computer, you can download it for Windows here and for MAC here. For editing and producing PDF's, I recommend TeXworks which can be downloaded here. Please use the LaTeX code provided on the course homepage for each homework assignment.

If you are handing in homework this way, read the comments in the code; there are two lines to comment out and you should replace my name with yours and write your section. The easiest way to use overleaf is to copy the raw text from hwxx.tex and preamble.tex into two new overleaf tex files with the same name. If you are asked to make drawings, you can take a picture of your handwritten drawing and insert them as figures or leave space using the "\vspace" command and draw them in after printing or attach them stapled.

Since this is extra credit, do not ask me for help in setting up your computer with LaTeX in class or in office hours. Also, **never share your LaTeX code with other students** — it is cheating and if you are found I will take it seriously.

Homework Extra Credit

There will be many extra credit questions sprinkled throughout the homeworks. They will be worth a variable number of points arbitrarily assigned based on my perceived difficulty of the exercise. Homework scores in the 140's are not unheard of. They are a good boost to your grade; I once had a student go from a B to and A- based on these bonuses.

Homework #0

For your first homework, you must:

(1) email me at kapelner@qc.cuny.edu from the email address you wish to be contacted at for this course (most commonly this is a gmail address),

(2) in the email, you must say "My name is <Your Full Name as appears in the registrar> and I have read and understand all the material in the course syllabus" and

(3) provide a link to your public repository on github (this means you need to sign up for github first)

This constitutes a contract — you are agreeing to this syllabus.

This assignment is due Friday, Jan 31, noon and will receive a grade of 0 or 100 with the usual 10 point penalty for lateness.

Writing Assignments

There will be two writing assingments. (1) A "philosophy of modeling" essay. Here you will coalesce the non-mathematical material that is crucial to this class. The purpose is to make you truly understand the modeling process and its limitations from start to finish. (2) A final project. Here you will use build a predictive model using a dataset. This is the capstone project for the entire data science and statistics major and it is where you will tie everything together.

This class will soon be the writing in the major course. Thus, writing is a major part of the curriculum herein.

Examinations

Examinations are solely based on homeworks (which are rooted in the lectures)! If you can do all the green and yellow problems on the homeworks, the exams should

not present any challenge. I will *never* give you exam problems on concepts which you have not seen at home on one of the weekly homework assignments.

Since the is the capstone course, there is no final exam, but a large final project. There will be two midterm exams and the schedule is below.

Exam and Major Assignment Schedule

- Midterm examination I will be Thurs, March 12 in class with the first review session on the Tuesday prior
- Midterm examination II will be Tues, May 12 in class with a review on the Friday prior
- The philosophy of modeling paper is due Tues, Mar 24 at 11:59PM
- The final project is due Sat, May 23 11:59PM by email

Exam Materials

I allow you to bring any calculator you wish but it cannot be your phone. The only other items allowed are pencil and eraser. I do not recommend using pen but it is allowed.

I also allow "cheat sheets" on examinations. For both midterms, you are allowed to bring **two** 8.5" $\times$ 11" sheet of paper (front and back). **Four sheets single sided are not allowed.** On this paper you can write anything you would like which you believe will help you on the exam.

Food is not allowed during exams but beverages are allowed.

Missing Exams

There are no make-up exams. If you miss the exam, you get a zero. If you are sick, I need documentation of your visit to a hospital or doctor. Expect me to call the doctor or hospital to verify the legitimacy of your note.

Special Services

If you are a student who takes exams at the special services center, I need to see your blue slip one week before the exam to make proper arrangements with the center.

Class Participation (and attendance)

I will be taking attendance selectively throughout the semester. Attendance counts towards the class participation portion of your grade in equal part with how often you ask and answer questions during the lecture.

Grading and Grading Policy

Your course grade will be calculated based on the percentages as follows:

Theory Homework	9%
Labs	13%
Class participation	5%
Midterm Examination I	18%
Midterm Examination II	18%
Philosophy of Modeling Paper	7%
Final Project with Writeup	30%

The second midterm is not cumulative. It only covers material beginning after midterm I.

The Grade Distribution

As this is a small and advanced class, the class curve will be quite generous. If you do your homework and demonstrate understanding on the exams, you should expect to be rewarded with an A or a B. $\leq$C's are for those who "dropped out" somewhere mid-semester or who cannot demonstrate basic understanding.

Checking your grade and class standing

You can always check your grades in real-time using the grading site. You will enter in your QC ID number (or email) and the password I will provide to you after HW #0.

Auditing

Auditors are welcome in both sections. They are encouraged to do all homework assignments. I will even grade them. Note that the university does not allow auditors to take examinations.