MATH 390.4 Spring 2018
Course Syllabus

Adam Kapelner, Ph.D.

Queens College, City University of New York

document last updated Sunday 29th April, 2018 6:31pm

Instructor	Professor Adam Kapelner
Contact	`kapelner@qc.cuny.edu`
Lecture Time / Loc	Monday and Wednesday 1:40-3:30 / Kiely 277
Office Hours / Loc	Monday and Wednesday 3:35-4:30PM / Kiely 604 (my office)
Course Homepage	https://github.com/kapelner/QC_Math_390.4_Spring_2018

Course Overview

MATH 390.4. Data Science via Machine Learning and Statistical Modeling with the R Language. 4 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. Polynomial and interaction regressions. Prediction with machine learning including neural nets (the perceptron), support vector machines and the tree methods CART and Random Forests. Probability estimation using logistic regression, asymmetric cost classifiers and the ROC curve. Underfitting vs. overfitting and the bias-variance tradeoff. Model validation including out of sample techniques such as cross validation and bootstrap validation. The `R` language will be taught formally from the ground and up and its use will be a substantial part of the homework. We will also cover visualization using the `ggplot` library and munging using `data.table`. Final project reports on a prediction model for a real-world dataset. *Spring*

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, Pois- son, 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 650.3-02 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. However, the course is listed as 3 credits so there won't be too much extra work since the Math 390.4 section is 4 credits.

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 "Preditive 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_390.4_Spring_2018`. 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).

Lectures

Lectures will be split 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.

There are 28 scheduled meetings. Of these, 23 will be lectures, 2 will be midterm exams which are in class and 3 will be review periods (the meeting before the exams). The exam schedule is given on page 6.

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 7) 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 (or until the end of semester 5/16/18).
- There must be *one* file and it must be in PDF format only.
- The file must be <2MB. No exceptions.

Homework

There will be 9-11 homework assignments and they will be a combination of theory and practice. 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). 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 our 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.

Philosophy of Homework

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.

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 the last day of class, Wednesday, May 16, 2018. 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.

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

LaTeX Homework Bonus Points

Part of good mathematics is its beautiful presentation. Thus, **there will be a 1–7 point bonus** added to your 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 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, Feb 2, 5PM and will receive a grade of 0 or 100 with the usual 10 point penalty for lateness.

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. There will be three exams and the schedule is below.

Exam Schedule

- Midterm examination I will be Monday, March 5 in class with the first review session on the Wednesday prior
- Midterm examination II will be Wednesday, April 18 in class with a review on the Monday prior
- The final examination will be Wednesday, May 23 8:30-10:30AM in KY277 with a review on the prior Wednesday, May 16.

I reserve the right to instead assign a final paper / report instead of the final exam.

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 one 8.5" $\times$ 11" sheet of paper (front and back). **Two 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. For the final, you are allowed to bring three 8.5" $\times$ 11" sheet of paper (front and back). **Six sheets single sided are not allowed.** I will be handing back the cheat sheets so you can reuse your midterm cheat sheets for the final if you wish.

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:

Homework	20%
Class participation	5%
Midterm Examination I	20%
Midterm Examination II	20%
Final Examination (or Paper)	35%

The semester is split into three periods :

1. From the beginning until midterm I. Midterm I covers material during this time..
2. From midterm I to midterm II. Midterm II covers material in this period only.
3. From midterm II until the final. The final (or paper) is cumulative and covers all course material.

Each of the periods is assessed evenly. Thus, each period must count the same towards your grade. Since there is 75% of the grade allotted to exams, there is 25% allotted to each period. Thus, the final is upweighted towards the material covered in the third period. In summary, the final will have 5/35 points $\approx$ 14% for the first period's material, 5/35 points $\approx$ 14% for the second period's material and 25/35 points $\approx$ 71% for the last period's material. A good strategy for the final is to just study the material after Midterm II and minimal studying for the previous material.

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 homework 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.