MAT245: Mathematical Methods in Data Science

Contents

1 Content of this Course

This is an introduction to the mathematical methods behind scientific techniques developed for extracting information from large data sets encountered in industrial applications. The course provides an introduction and overview of the field which stays motivated by real-world examples. The demands imposed by example applications will lead us to develop theoretical techniques requiring mathematical expertise. Many of the topics are not included in standard mathematical curriculum (singular value decomposition, Cholesky factorization,...) despite their ubiquity in the field and their relative simplicity and naturalness. Often the topics encountered in the analysis of data only superficially cover technical details necessary for making advances in refining existing techniques. This course aims at refocussing attention on the underlying mathematical concepts needed to fully understand the methods used by modern data scientists.

Technical Details

Make sure you are aware of the following University details regarding this course.

Instructor: Nicholas Hoell – nicholashoell@gmail.com. Office: PG200A. Office Hours: TBD.
Catalogue Description: Elementary probability density functions, conditional expectation, inverse problems, regularization, dimension reduction, gradient methods, singular value decomposition and its applications, stability, diffusion maps. Examples from applications in data science and big data.
Prerequisites: MAT137Y1/MAT157Y1, MAT223H1/MAT240H1, MAT224H1/MAT247H1
Corequisites: MAT237Y1/MAT257Y1
Distribution Requirement Status: Science
Breadth Requirement: The Physical and Mathematical Universes (5)

2 Who is This Course For?

First and foremost **this is a course on mathematics**, not a course on Data Science. What that means is that we will be emphasizing mathematical concepts underlying tools used by practitioners with a heavy slant towards the "theorem/proof" style common to specialist courses in the mathematics department. The exercises are challenging and the results are rigourous. I should say that we will be covering a lot of topics throughout the term and each topic requires thought and attention, so it's very important to not fall behind. Adding to the challenge will be the fact that we are **not using a textbook** so attending lectures will be of even more importance than usual.

This is a course principally aimed at mathematics specialists. The ideal student will have taken all of the prerequisites, and ideally, the specialist track. If that's not you, that's alright just be aware that the course covers things at a fairly high level of *mathematical maturity* - which means there will be ϵ's and δ's and somewhat involved proofs at points. The broad learning expectations are to introduce you to an important part of applied mathematics focussed on modern modelling practices. The ideal student has an interest in learning how mathematics is used in practice and how mathematical ideas can be brought to bear on real-world industrial problems.

If you find yourself enjoying the material we cover this semester, there are a lot of course offerings in Statistics (Probability, Time Series, Machine Learning, Data Analysis, etc.) and Computer Science (Machine Learning, Numerical Analysis, etc.) which you may also wish to pursue since they will cover various topics in more depth than we can this term.

3 Textbook

This is a unique course, bringing in mathematical tools not normally presented in a single course, for the purposes of solving problems arising in different fields related to the analysis of data sets. We will be pulling material from what may look like a daunting number of sources. Your *primary resource therefore should be the lectures!* No single textbook exists which really meets the needs of a course like this so my suggestion is to stay close to the lectures and follow the material closely in the main required texts. Beyond that, when you encounter an interesting idea you wish to study more, you will find a lot of help in the listed optional texts. They are arranged, *roughly*, according to topics of emphasis for us in the course, but many of them cross the genres I've indicated.

3.1 Main Optional Texts

I will pull some material from the following daunting list of texts, you may wish to purchase them to get a deeper dive into the material presented in lectures.

Numerical and/or Advanced Linear Algebra

1. Deuflhard & Hohmann's *"Numerical Analysis in Modern Scientific Computing"*, published by Springer ISBN: 978-0-387-95410-4. Don't let the name fool you - this is a theorem/proof style text which can be quite challenging with difficult end-of-chapter mathematics problems. We will follow their treatment of conditioning, orthogonalization methods, and generalized inverses.

2. Golub & Van Loan's *"Matrix Computations"*, by Hopkins ISBN-13: 978-0801854149. This is a book on numerical linear algebra which has a nice presentation of SVD, conditioning, least squares and pseudo-inverses useful for this course.

3. Press, Teukolsky, Vetterling, & Flannery *"Numerical Recipes: the Art of Scientific Computing"* published by Cambridge University Press, ISBN: 978-0-521-88068-8. This classic tome covers algorithms with great explanations and derivations. We will follow its style in our discussion of FFT and HMM.

4. Horn & Johnson's *"Matrix Analysis"* published by Cambridge University Press, ISBN: 978-0-521-54823-6. Should be titled "Everything you wanted to know about matrices but were afraid to ask". It's a comprehensive, serious mathematical work on topics related to matrices. I will base some of my lectures on QR and SVD factorizations on material in here. It's an excellent reference for material on norms.

5. Rao's *"Linear Statistical Inference and its Applications"* published by Wiley ISBN: 978-0-471-21875-3. This is a classic from one of the major figures in statistics. Chock-full of advanced but clear topics. Not modern at all but still a classic.

6. Damelin & Miller's *"The Mathematics of Signal Processing"* by Cambridge University Press ISBN-13: 978-1107601048. This is a more mathematically advanced book covering a lot of advanced (and modern) topics useful for practitioners in data-driven fields, particularly around imaging and time series.

Probability

1. Ross' *"A First Course in Probability"*, published by Pearson, ISBN-13: 978-0321794772. In full disclosure I haven't used it but it seems to be good (and popular).

2. Hoel, Port, & Stone's *"Introduction to Probability Theory"*, published by Brooks Cole, ISBN-13: 978-0395046364. This is the first book in a well-written trilogy which covers all of the standard stuff but very well.

3. Jaynes' *"Probability Theory: The Logic of Science"*, published by Cambridge University Press, ISBN-13: 978-0521592710. This is a good overview with nice emphasis on the Bayesian viewpoint.

4. Feller's *"An introduction to Probability Theory and Its Applications"* by Wiley ISBN-13: 978-0471257080. This is a classic. Probably the clearest, most elegant in its class, although it can get quite advanced.

Machine Learning

1. Christopher Bishop's *"Pattern Recognition and Machine Learning"* published by Springer, ISBN: 978-0-387-31073-2. This is a very readable, excellent textbook which covers a lot more material than we'll need but which will serve you well in the future. It's mathematical but written more in the style of statistical sciences.

2. James, Witten, Hastie & Tibshirani's *"An Introduction to Statistical Learning"* published by Springer, ISBN-13: 978-1461471370. The authors have made this book freely available in a pdf form on the website

   http://www-bcf.usc.edu/~gareth/ISL/

   The book gives a more gentle lay of the land than Bishop's book and has nice presentation of the elementary topics.

Miscellaneous

1. Kaipio & Somersalo's *"Statistical and Computational Inverse Problems"* published by Springer, ISBN: 978-0-387-27132-3. This is a great book which focuses on applications important in biomedical research. It is in the traditional theorem/proof style you should be familiar with by now. The proofs are clean but somewhat tricky. It's more mathematically advanced than Bishop's book is. Most of the material covers **continuous problems** which generalize what we cover in this course so it's better for advanced study.

3.2 Online Texts

The following are excellent books which are available for free.

1. Blum, Hopcroft & Kannan's *"Foundations of Data Science"* available free online at

   https://www.cs.cornell.edu/jeh/book2016June9.pdf

   This has great material on the Curse of Dimensionality and various topics in machine learning. Theorem/proof style but with a well-chosen list of topics and clear explanations.

2. Bandeira's *"Ten Lectures and Forty-Two Open Problems in the Mathematics of Data Science"* available free online at

   http://www.cims.nyu.edu/~bandeira/TenLecturesFortyTwoProblems.pdf

   This is more advanced and more in the form of lecture notes rather than a text. One of its main attractions is the list of (advanced) open problems which, while technical and advanced, may be worth your attention when the course is over.

Other Materials

In addition to the textbook, we will occasionally post additional problems and references on the course website as the semester progresses.

4 Homework

Homework will be collected throughout the course. It's fine to discuss ideas with classmates but the write-up must **be entirely your own**, and anyone caught violating this will be dealt with in the same way as if you were found to be cheating on a test (for which you should read the section on Academic Integrity). **No late work will be accepted** without written documentation of medical issue (see section on medical notes).

5 Teaching Assistants

This course has one part-time and two full-time teaching assistants:

1. Matt Sourisseau – Email: sourisse@math.utoronto.ca.
2. Tristan Milne – Email: tristan.milne@mail.utoronto.ca

Both are extremely competent and knowledgable. Matt is your main point-of-contact for programming/python issues throughout the semester and will run the required computer labs.

6 Computer Labs

This course has weekly computer labs scheduled for

1. Mondays at 3-5pm in SS561 lab with TA Danny Luo. Email: danny.luo@mail.utoronto.ca
2. Fridays at 9-11am in SS561 lab with TA Gideon Providence. Email: g@math.toronto.edu

You **must be enrolled in one of these lab times**. The room SS561 is on the lower level on Sidney Smith. The labs will introduce you to the Python programming language. In case you have never written computer programs before, the first couple labs will be spent introducing you to the Python interpreter and going over fundamental principles used in the design of simple programs in Python, a standard language used in industry. Throughout the term the labs will expose you to the implementation of concepts covered in more theoretical detail during lectures. You will be given some programming assignments and time to execute it during the lab time. As well, you will be introduced to useful standards of practice and helpful packages.

7 Grading

Grades will be based according to some homework given throughout the term, work done in labs, one midterm exam and a final exam. Your final grade in the course will be determined by the following

- *Labs: 10%*
- *Homework: 20%*
- *Midterm: 30%*
- *Final Exam: 40%*

8 Midterm

A midterm will be given in class on **Thursday, October 18**. There are no make-up tests.

8.1 Missed Tests

There are no makeup tests. A student presenting proof of a **valid** reason for missing a test (see the section on Missed Term Tests in the Rules and Regulations section of the Faculty of Arts and Science 2017-2018 Calendar as well as the following section of this syllabus) will have their grading scheme adjusted to the following

- *Labs: 10%*
- *Homework: 20%*
- *Final Exam: 70%*

It is strongly advised that you write the term test as 70% for the final exam puts quite a lot of pressure on a single test.

8.2 Medical Notes

In the case of a legitimate medical issue **medical notes will be accepted ONLY from MDs with a valid CPSO number**. You must present your section Instructor with a University of Toronto Verification of Student Illness or Injury form available at http://www.illnessverification.utoronto.ca/getattachment/index/Verification-of-Illness-or-Injury-form-Jan-22-2013.pdf.aspx .
Some important remarks about these notes.

- These forms **must be submitted to your course instructor within 3 days of the missed test** for the absence to not be penalized. Failure to submit proper, valid and timely documentation will result in a grade of 0 on your missed test.

- The form must have all required fields filled properly and legibly.
- The form must give the doctor's OHIP number.
- The form must be original.
- The form is only considered valid if **completed by a qualifed medical doctor - not an acupuncturist, chiropractor, naturopath or other health care professional.**
- Upon submission of the documentation review of the medical note will be done before it is accepted as valid. The review **may include following up with your doctor, your college registrar, other departmental advisors.**

Presenting a false medical excuse is a severe offence and will be dealt with through the Office of the Dean of the Faculty of Arts and Science.

8.3 Athletic Absences

If you are a member of a University of Toronto sports team which has an event scheduled on the date of one of our tests and you wish to not miss your event then you **must get a letter on University letterhead** from your coach in order for this to count as an excused absence. The same grading policy towards excused medical absences applies in this instance. The only difference is that **you must have the letter sent to us prior to the week of the midterm you plan on missing.**

9 Contact

9.1 Email Etiquette

It is University policy that **instructors need only reply to emails sent from University email accounts**. Acceptable emails are of the form student@utoronto.ca, topstudent@math.toronto.edu, etc. I **do not reply** to non-University email addresses (those from addresses like, say, studentwhoseemaildidntgetreturned@hotmail.com or studentwhodidntreadthesyllabus@gmail.com). Any email should have the words "MAT245" somewhere in the subject line.

As for email etiquette, sending technical mathematics questions to me is okay if they are **very short** and worded very precisely. Save longer questions for the beginning Q&A part of the following lecture, or ask during office hours. You may also use the online Piazza forum for discussing questions with each other (which we monitor). You should **not address me** as "Hey", "Yo", or other highly informal salutation in an email. If you do you won't get a reply. Keep in mind I get a high volume of emails each semester and it may take time before you hear back from me (if at all). Your chances of having your email read and being replied to, increase *dramatically* if you can follow the instructions above.

10 Academic Integrity (Important)

Cheating (including plagiarism) is very serious and, consequently, will be taken **very** seriously. Cheating can result in failure **or worse**. Don't do it! I caution you, I am *extremely* diligent in pushing for the maximum possible penalties for those found cheating. Collaboration on the homework problem sets is fine, in fact, we encourage it since discussing problems with your peers helps bolster your problem solving abilities. But any collusion during test situations will be thoroughly punished. This includes talking (or making other extraneous noises of any kind) during a test. We don't tolerate any kind of chatter during tests.

One other thing. There are students for whom the statement "The test is now over, please put your pens and pencils down while we collect the tests" seems to not entirely register. We consider egregious dismissals of our requests to stop writing to be a form of academic integrity violations which we enforce with the same stringency as talking during a test. **It's not worth the risk!**

11 The Menu

We will keep, *very roughly*, to the following schedule.

- Week 1
  - **Topics:** Vector norms and unit balls, FTLA and SVD.
- Week 2
  - **Topics:** Matrix norms, the Eckart/Young/Mirsky theorem. Dimension reduction via PCA.
- Week 3
  - **Topics:** Some numerical issues: overflow/underflow, conditioning/stability. Condition number.
- Week 4
  - **Topics:** Linear least squares. Normal equations and gradient descent. The Penrose axioms.
- Week 5
  - **Topics:** More on linear least squares: Pseudo-inverse. QR factorization and Householder reflections.
- Week 6
  - **Topics:** Probability overview: Random variables, expected value, probability density and mass functions.
- Week 7
  - **Topics:** Probability continued: Conditional probability, Bayes formula. Likelihood functions.
- Week 8
  - **Topics:** Common distributions. CLT. Curse of Dimensionality: Law of Large Numbers and the unit ball in high dimensions.
- Week 9
  - **Topics:** Bayesian regularization. A second look at PCA. Clustering via K-means.
- Week 10
  - **Topics:** Decision boundaries and discriminants. Logistic regression.
- Week 11
  - **Topics:** Perceptron and online perceptron. Block's theorem. Neural networks.
- Week 12
  - **Topics:** More on neural nets. Backpropagation.