Maymester MATH347 Linear Algebra for Applications in Data Science

An exploration of the mathematics and art of data science

Course syllabus

Times	Daily, 5/19 - 6/3, 11:30AM-2:45PM
Office hours	TuTh 10:00-11:00AM (Zoom)
Instructor	Sorin Mitran

Motivation

Artificial intelligence, *data science*, *machine learning* are buzzwords arising in academia, business, politics. What do they really mean? What is the knowledge foundation of all the excitement? This is where mathematics comes in. The entrepreneurial mindset is to resolutely transform an idea into action that changes society. The mathematical mindset is to rigorously distill observation, intuition, and aesthetics into an idea. The mathematical idea explored in this course is simple to state:

“*What can I build by simple combination of some multidimensional objects?*”.

After choosing n objects each characterized by m numbers, $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n \in \mathbb{R}^m$, the simple combination studied is to resize each object by scale factors $x_1, x_2, \ldots, x_n$ and then add them together,

$$\boldsymbol{b} = x_1 \boldsymbol{a}_1 + x_2 \boldsymbol{a}_2 + \cdots + x_n \boldsymbol{a}_n.$$

The above *linear combination* formula is about as complicated as the mathematics gets to be in the course, but leads to a treasure trove of applications: balancing a chemical reaction, determining market equilibrium, finding genetic inheritence, analyzing social interactions, identifying faces in a crowd. Solutions to all these problems are found by linear combinations, and linear algebra provides the rigorous framework to determine answers to questions such as:

- can all objects of interest be reached by linear combination?
- what type objects cannot be reached by linear combination?

- if an object cannot be reached by linear combination, what's closest to it?
- are there special objects that do not get significantly changed by linear combination?
- can objects be more insightfully be described by a different linear combination?

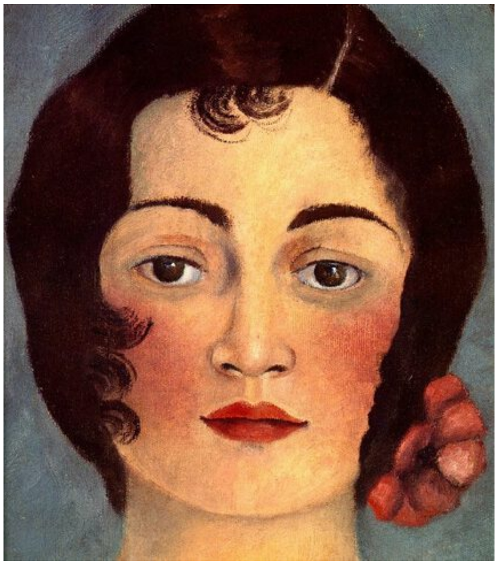
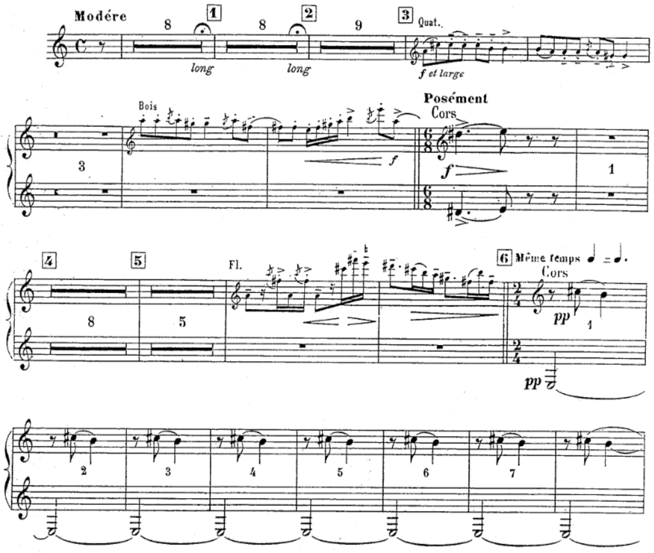
Mathematical concepts are introduced to precisely frame each question above (range, null space, least squares, eigenvectors, change of basis), but the technical terms should not cloud the essential simplicity of the questions.

A common feature of applications is that very many numbers are required to describe an object, such that m is very large, while n, the number of objects we wish to combine is small. Once the limits of linear combinations in such cases are determined in *linear algebra*, the natural question is to ask if some other way of combining objects, formally described by some yet unspecified non-linear function $\boldsymbol{f}$ as

$$\boldsymbol{b} = \boldsymbol{f}(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n),$$

is more powerful. This is the main question within *data science*. It turns out that the relevant mathematics is more complicated and incomplete, and linear algebra is again a useful guide, as exemplified by deep neural networks that serially link several linear combinations by simple nonlinear functions.

The applications of linear algebra and its role as a foundation for data science arguably make the subject of greater relevance to today's society than topics such as calculus. There is also an art, a certain aesthetic to statement of linear algebra problems captured by a symbiosis of notation, definitions and understanding of concepts. The course reinforces this link with many examples from the world of art, hopefully leading to an appreciation of the essential unity of the three examples of human ingenuity presented in the table below.

	$\min_{\boldsymbol{b}} \lVert \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} \rVert$	 Modére long long Quat. f et large Bois Posément Cors Fl. Même temps Cors pp pp
Portrait of a Woman in White, 1930	The least squares problem	*Romanian Rhapsody No. 1*, 1901
Frida Kahlo	David Hilbert, ~1909	George Enescu

Though it might not be readily apparent all the above involve linear combinations!

Course goals

This Maymester course is intended as a rapid introduction to concepts from algebra that are most useful to data science. Upon course completion students:

- will understand what can and cannot be obtained by a linear combination of objects;
- will recognize the principal problems within linear algebra, i.e.,
    - change of basis (solving a linear system),
    - best approximation (least squares),
    - invariant directions (eigenvectors);
- will become proficient in organization of vector and matrix manipulations by hand;
- understand the role of the most useful matrix factorizations ($\boldsymbol{LU}$, $\boldsymbol{QR}$, SVD, eigendecomposition);
- be exposed to topics within calculus with close links to linear algebra and data science;
- gain the basic practical coding skills in Matlab/Octave needed to solve linear algebra problems;
- be exposed to applications of linear algebra outside of the realm of the physical sciences, with an emphasis on examples from art, biology and medicine, and the social sciences.
- understand the role of linear algebra within the wider topics of algebraic structures and data science.

Honor Code

Unless explicitly stated otherwise, all work is individual. You may discuss various approaches to homework problems with students, instructors, but must draft your answers by yourself.

Course policies and organization

• Class attendance is required. Students must bring a laptop that conforms to the minimal CCI requirement to each class and the final examination.

• Homework is assigned every two lessons. The last third of each daily meeting during this Maymester course is used to start the homework with assistance of Instructor. Completion of the homework will require no more than two additional hours outside of class time. Each homework will consist of 6 theoretical questions (1 point each), and an application to realistic data (4 points)

• Homework is to be submitted electronically through Sakai. Late homework is not accepted.

• Three fifteen-minute quizzes are given on days 4,7,10 and test basic comprehension of definitions and simple operations.

• The final examination will consist of a first, closed-book part with questions similar to those on quizzes that test understanding of basic concepts, followed by a two-hour, open-book part in which students will use course concepts and laptops to solve a practical problem of complexity similar to a homework application.

Grading

Required work

• Homework: 4 assignments, 2 x 10 + 2 x 20 points = 60 points.

• In-class quizzes: 3 tests x 5 points = 15 points.

• Final examination (solution): 25 points = 15 True/False quiz questions, 2 x 5 point application questions

Mapping of point scores to letter grades

Grade	Points	Grade	Points	Grade	Points	Grade	Points
A+	101-112	B+	86-90	C+	71-75	D+	56-60
A	96-100	B	81-85	C	66-70	D-	50-55
A-	91-95	B-	76-80	C-	61-65	F	0-49

Course materials

Course topics

The course is organized around six basic questions, each discussed in two days of the Maymester course schedule. Each question leads to specific mathematical concepts listed below. In each leasson the mathematical concepts are applied to realistic data, chosen from a variety of fields.

COM. *What tools are needed to work with linear combinations?*. Vectors, matrices, matrix operations, norm, scalar product. Images, electroencephalograms, musical phrases as vectors and their transformation.

VEC. *What is the mathematical framework for questions about linear combinations?* Algebraic structures, vector spaces and subspaces, vector set span, range and null spaces, linear dependence, matrix rank, orthogonal matrix. Data redudancy and the facial recognition problem.

THM. *Can we classify objects as reachable or unreachable by linear combination?* Fundamental theorem of linear algebra, rank-nullity theorem, singular value decomposition. Painter style and motifs, bases for a large dimensional space.

LSQ. *How close can we get to an object by linear combination?* Gram-Schmidt algorithm, projection, least squares, data fitting. Data compression, simplification of complex models from structural engineering (reduced-order systems).

BAS. *What happens if we change the objects we combine?* Linear systems, coordinates, change of basis, Gauss elimination, LU-factorization, determinants. Construction of bases by greedy data approximation to distinguish painter style.

EIG. *Are some objects left essentially unchanged by linear combination?* Eigenvalues, eigenvectors, characteristic polynomial, repeated eigenvalues (algebraic and geometric multiplicities), the Schur decomposition, spectral expansion, rank-1 expansions. Musical phrases, mechanical vibrations.

SFT. Application of the above concepts requires proficiency in use of linear algebra and general scientific software. The Math@UNC environment constructed for this course provides a standard framework and is presented in Lesson00 that should be studied prior to the start of the course.

Textbook

Class notes specially drafted for this course will be provided as “live” TeXmacs documents (.tm file extension) that contain code for carrying out linear algebra operations and data analysis. The class notes are meant to be studied in the Math@UNC environment within which code examples can easily be modified by students to study new applications. Notes are posted prior to class time. Prior to semester start Lesson00, Lesson01 are posted for prospective students to form an idea on course content and approach. Notes are also available for download in Portable Document Format (.pdf file extension) for offline study and printing, as well as Extensible Markup (.xhtml file extension) for web browsing.

Lesson	Topic	LiveDoc	Web pages	Notes	Webinar	Slides
00	SFT	lesson00.tm	Lesson00.xhtml	Lesson00.pdf	Lesson00.webm	
01	COM	lesson01.tm	Lesson01.xhtml	Lesson01.pdf	Lesson01.webm	Slides01.pdf
02	COM	lesson02.tm	Lesson02.xhtml	Lesson02.pdf	Lesson02.webm	Slides02.pdf
03	VEC	lesson03.tm	Lesson03.xhtml	Lesson03.pdf	Lesson03.webm	Slides03.pdf
04	VEC	lesson04.tm	Lesson04.xhtml	Lesson04.pdf	Lesson04.webm	Slides04.pdf
05	THM	lesson05.tm	Lesson05.xhtml	Lesson05.pdf	Lesson05.webm	Slides05.pdf
06	THM	lesson06.tm	Lesson06.xhtml	Lesson06.pdf	Lesson06.webm	Slides06.pdf
07	LSQ	lesson07.tm	Lesson07.xhtml	Lesson07.pdf	Lesson07.webm	Slides07.pdf
08	LSQ	lesson08.tm	Lesson08.xhtml	Lesson08.pdf	Lesson08.webm	Slides08.pdf
09	BAS	lesson09.tm	Lesson09.xhtml	Lesson09.pdf	Lesson09.webm	Slides09.pdf
10	BAS	lesson10.tm	Lesson10.xhtml	Lesson10.pdf	Lesson10.webm	Slides10.pdf
11	EIG	lesson11.tm	Lesson11.xhtml	Lesson11.pdf	Lesson11.webm	Slides11.pdf
12	EIG	lesson12.tm	Lesson12.xhtml	Lesson12.pdf	Lesson12.webm	Slides12.pdf

The above notes are also gathered into a traditional textbook.

Homework

Homework consists of direct application of concepts discussed during each lesson, and is mostly completed during class time. Homework is drafted using the integrated mathematical editing and computation facilities of TeXmacs. A tutorial template is provided (hw00.tm) to familiriaze students with basic editing and computation procedures.

Nr.	**Topic**	**Problems**	**Solution**
00	tutorial	hw00.tm hw00.pdf	
01	COM	hw01.tm hw01.pdf	sol01.tm sol01.pdf
02	VEC	hw02.tm hw02.pdf	sol02.tm sol02.pdf
03	THM	hw03.tm hw03.pdf	sol03.tm sol03.pdf
04	LSQ	hw04.tm hw04.pdf	sol04.tm sol04.pdf

Software

Modern software systems allow efficient, productive formulation and solution of mathematical models. A key goal of the course is to familiarize students with these capabilities, by extensive use of two applications:

1. TeXmacs, a scientific editing platform, used for preparation of live lessons and to draft homework assignments;

2. Octave, an open-source numerical and graphical computation package, especially suited for linear algebra, that uses essentially the same coding language as Matlab.

These applications are provided within the Math@UNC virtual machine environment. Follow the instructions posted there to install the software on your CCI-compliant laptop.

Tutorials

Software usage is introduced gradually in each class, and class participation should be sufficient to gain enough familiarity to effectively use these tools. Some additional resources are also readily available for further study if desired:

- TeXmacs
    - http://www.texmacs.org/tmweb/help/tutorial.en.html
    - https://www.youtube.com/watch?v=mlcqGRv7xhc
- Octave

- https://en.wikibooks.org/wiki/Octave_Programming_Tutorial
- https://www.youtube.com/watch?v=sHGqwF2s-tM

Course material repository

Course materials are stored in a repository that is accessed through the subversion utility, available on all major operating systems. The URL of the material is svn://mitran-lab.amath.unc.edu/courses/MATH347DS. Under Windows, Tortoise SVN can be used to download all course materials through the subversion utility, or individual files can be downloaded from this website. Subversion is included in the Math@UNC virtual machine.