San José State University, Spring 2023

Math 250: Mathematical Data Visualization

Course Introduction and Overview

Dr. Guangliang Chen

Agenda

1. Introductions
2. Course overview
3. Course policies

Homework 0 (due 2/2, Thursday, 11:59pm)

Lecture 1: Matrix algebra (if time permits)

Introductions

Instructor: Guangliang Chen, Associate Prof. of Statistics & Data Science, 2014-present

Education:

- BS Math, Univ. of Sci.& Tech. of China, Hefei, 2003
- PhD Applied Math, University of Minnesota, July 2009

Employment history:

- Duke University: Visiting Assistant Professor, 2009-2013
- Claremont McKenna College: Visiting Assist. Prof., 2013-2014

Now it is your turn

Please tell us

- **your name**,
- **program of study**,
- **academic year**, and
- **anything else about you that you want us to know**.

What is this course about?

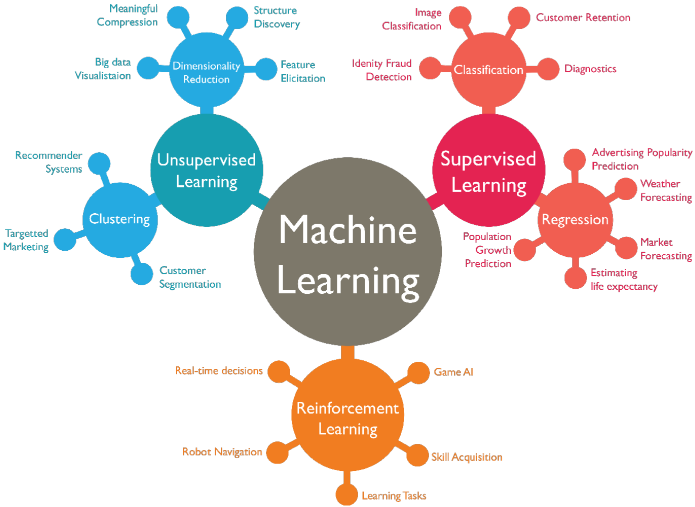
Context: Modern data sets often have hundreds, thousands, or even millions of features (or attributes). ⟵ large dimension

Math 250 course introduction and overview, Spring 2023

This course focuses on the machine learning task of dimension reduction, also called dimensionality reduction, which is the process of reducing the number of input variables of a data set under consideration, for the following benefits:

- It reduces the **running time** and **storage space**.
- Removal of **multi-collinearity** improves the interpretation of the parameters of the statistical / machine learning model.
- It can also clean up the data by reducing the **noise**.
- It becomes easier to **visualize the data** when reduced to very low dimensions such as 2D or 3D.

Math 250 course introduction and overview, Spring 2023



There are two different kinds of dimension reduction approaches:

- Feature selection approaches try to find a subset of the original features variables.

  Examples: *subset selection*, *stepwise selection*, *Ridge and Lasso regression*. ⟵ Covered in Math 261A

- Feature extraction transforms data in a high-dimensional space into another space of fewer dimensions. ⟵ Focus of this course

  Examples: *principal component analysis (PCA)*, *ISOmap*, and *linear discriminant analysis (LDA)*.

Use of dimension reduction

Dimensionality reduction can greatly help with the following statistical and machine learning tasks:

- **Regression** (Math 261A)
- **Classification** (Math 251)
- **Clustering** (Math 252)
- **Dimensionality reduction (and visualization)** ⟵ this course

We'll focus on data visualization in this course as the main application and motivation of dimension reduction.

What is data visualization and why is it important?

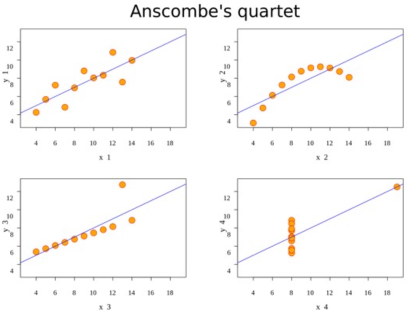
Data visualization is the graphic representation of data. According to Friedman (2008) "*the main goal of data visualization is to communicate information clearly and effectively through graphical means*."

Why it is important: A picture is worth a thousand words – especially when you are trying to understand trends, outliers, and patterns in data sets that include thousands or even millions of variables.

Data visualization can provide insight that descriptive statistics cannot, see an example on next slide.

This following example highlights why it's important to visualize data and not just rely on descriptive statistics.

Example: Anscombe's Quartet (Francis Anscombe, 1973): The 4 datasets have almost identical mean, variance, correlation between X and Y coordinates, and linear regression lines.



However, the patterns are very different when plotted on a graph.

Design of this course

This course covers the following (to prepare you for Math 251):

- Central topic: **dimension reduction**
- Main motivation and application: **Data visualization**
- Supporting tools
  - **Advanced linear algebra**
  - **Matrix computing** and **3D data plotting** in Matlab

Overall, this course is 70% theory + 30% programming.
(Math 251 will be the opposite)

Dimension reduction methods to be covered in this course:

- **Linear projection methods**:
  - PCA (for unlabeled data),
  - LDA (for labeled data)
- **Nonlinear embedding methods**:
  - Multidimensional scaling (MDS), ISOmap
  - Locally linear embedding (LLE)
  - Laplacian eigenmaps

Use of the course

This course is used in the following ways:

- A prerequisite to **Math 251 Statistical and Machine Learning Classification**
- An elective for the regular **MS Statistics degree**
- A required course by the **MS Statistics Machine Learning Specialization** (along with Math 251)
- A required course by the **MS Data Science degree** (joint with CS)

Prerequisites of the course

- Math 32 multivariable calculus**
- Math 39 linear algebra**
- Math 163 probability theory*

Though the course has no programming prerequisite, there is a significant computing component (and you will have the chance to learn a new technical language - MATLAB).

**Requires a B or better grade, *Requires a C or better grade

Textbook

Required: None, but instructor's draft book chapters will be provided.

Recommended Readings (at more advanced levels):

- "Foundations of Data Science", Avrim Blum, John Hopcroft, and Ravindran Kannan, Cambridge University Press, March 2020.[1]
- "Probabilistic Machine Learning: An Introduction", by Kevin Patrick Murphy. MIT Press, March 2022.[2]

[1] https://www.cs.cornell.edu/jeh/book.pdf

[2] https://probml.github.io/pml-book/book1.html

Technology requirements

- Access to a scanner (physical or cell phone app)
- The MATLAB software (see next slide)

Computing

This course will use MATLAB as the main programming language due to its advantages in matrix computing and data plotting.

San Jose State has purchased a campus-wide MATLAB license for everyone to use for free.[3].

This course only needs the Statistics and Machine Learning Toolbox[4] (besides the MATLAB main platform).

[3] https://www.mathworks.com/academia/tah-portal/san-jose-state-university-31511582.html

[4] https://www.mathworks.com/products/statistics.html

Data sets to be used in this course

We will use the following data for learning and practice:

- **MNIST Handwritten Digits**[5]: 70,000 digital images of size 28x28 of handwritten digits $0 \ldots 9$ collected from about 250 people
- **Fashion-MNIST**[6]: Same size and format with MNIST, but the images contain clothes instead

[5] `http://yann.lecun.com/exdb/mnist/`

[6] `https://github.com/zalandoresearch/fashion-mnist`

- **USPS Zip Code Data**[7]: 9,300 size 16x16 grayscale images of handwritten digits scanned from envelops
- **20 Newsgroups Data**[8]: about 19,000 text documents that are divided into 20 groups (according to their topics)

Smaller data sets such as the **Wine Quality Data Set**[9] from the *UCI Machine Learning Repository*[10] will also be used for teaching demonstration and homework assignments.

[7] `http://statweb.stanford.edu/~tibs/ElemStatLearn/data.html`

[8] `http://qwone.com/~jason/20Newsgroups/`

[9] `https://archive.ics.uci.edu/ml/datasets/wine+quality`

[10] `http://archive.ics.uci.edu/ml/`

Requirements of this course

- **Homework** (20%): Assigned roughly weekly (12 in total)
- **Midterm 1** (30%): Monday, March 20, regular class time
- ***Midterm 2** (35%): Wednesday, April 26, regular class time
- **Final project** (15%): Due Wednesday, May 17
  - Project presentations: 9:45am - 12pm (slides due 9:30am)
  - Project reports (and codes): 11:59pm

*comprehensive

Grade cutoffs

...will be determined by combining the following **percentages**:

- A+: 98%, A: 93%, A-: 90%
- B+: 86%, B: 80%, B-: 76%
- C+: 73%, C: 68%, C-: 65%
- D+: 63%, D: 58%, D-: 56%

and **the actual distribution of the class** at the end of the semester.

Homework policy

Homework assignments will typically contain both theory and coding questions.

- You must submit homework directly to Canvas and on time in order to receive full credit (late submission within 24 hours will still be accepted but there is a penalty of 10% of the total number of points).
- You may collaborate on homework but must write everything on your own.
- For theory questions, your answers must have necessary supporting steps.
- For programming questions, results must be supported by codes and presented in a concise, meaningful manner, e.g., by using figures or tables.
- The lowest homework score will be dropped.

What is allowed when doing the homework

Collaboration is encouraged on homework, but only for the learning part. That is, you may (without needing to acknowledge your learning partners):

- Discuss homework questions together;
- Come up with a strategy/solution together;
- Help each other with certain step or line of code;
- Compare answers with each other.

However, you must write your code and/or steps individually on your own.

What is considered cheating or plagiarism

Some examples of cheating in completing a homework assignment are

- Copy other people's work partly or in full
- Use other people's products (such as plots and code) for your own submission (even with acknowledgment)
- Give your work to other people for copying or studying
- Copy solution or code found online (even with acknowledgment).[11]

[11] However, you can study it and after you fully understand it, rewrite the steps or code independently by yourself.

Midterms

The course has two midterms, to be delivered physically in class.

Both of them are closed book and closed notes.

The second midterm is comprehensive, thus equivalent to an early final.

The two exams will focus on the theory component of the course (in contrast, the programming component of the course is covered by homework and the final project).

The final project

This course ends with a data visualization project that aims to provide you with an opportunity to practice and apply the methods learned in class to large, high dimensional data sets from the internet.

The class will be divided into **groups of size two** to work on the projects.

The data sets used by different groups must be distinct. Each data set must have at least 5000 instances and 10 features, and requires advanced approval by the instructor.

The students will need to give a short oral presentation to report their findings and meanwhile write a report of 5+ pages.

Learning management system

I will use **Canvas** in various ways:

- Post homework assignments and tests
- Record homework and test scores
- Make announcements (e.g. reminders, clarifications, deadline changes)

Make sure to check your Canvas settings to receive timely notifications. Also, check if your email address in record is still good.

Piazza

This term we will be using Piazza for class discussion. The system is highly catered to getting you help fast and efficiently from classmates and the instructor.

Rather than emailing questions to me, I encourage you to post your questions on Piazza. If you have any problems or feedback for the developers, email team@piazza.com.

Find our class signup link at:
`https://piazza.com/sjsu/spring2023/math250`

Course webpage

I am maintaining a course webpage[12] for posting the following information:

- Course syllabus (and other relevant information).
- Lecture slides (and additional learning resources)

Please visit the webpage right before each class to download the latest version of the lecture slides (try refreshing your browser if you don't see them).

[12] https://www.sjsu.edu/faculty/guangliang.chen/Math250.html

Your responsibilities in learning

My duty as an instructor is to disseminate knowledge while helping you learn. The ultimate responsibility of learning is upon the student, not the instructor. Thus, you should make every effort to

- Attend all classes
- Participate in classroom discussions
- Read the textbook before and after class
- Take time to think through the concepts
- Do your homework
- ASK whenever you don't understand something!!!

Academic dishonesty

Students who are suspected of cheating in completing any assignment (homework, exam, or project) will be referred to the Student Conduct and Ethical Development office, and depending on the severity of the conduct, will receive a zero on the assignment or even a grade of F in the course.

Some final reminders

This course is

- very challenging (theory, or programming, or both)
- demanding (timewise)

A lot of hard work is required to succeed in this course.

However, the course is very useful, as it builds up the mathematical, computing, and data foundations for subsequent machine learning coursework or research.

Special accommodations

If you anticipate needing any special accommodation during the semester (e.g., you have a disability registered with SJSU's Accessible Education Center), please let me know as soon as possible.

Instructor availability

- **Office hours**: Monday 10:30 - 11:30am, Wednesday 1:30 - 3pm, and by appointment (Tuesdays and Thursdays).
- **Piazza**: `piazza.com/sjsu/spring2023/math250`.
- **Email**: `guangliang.chen@sjsu.edu`. I check my emails frequently, but you should allow a turnaround time of up to 24 hours (on weekdays) or 48 hours (during weekends).

Student feedback

I strive to teach in the best ways to facilitate your learning. To achieve this goal, it is very helpful for me to receive timely feedback from you.

You can choose to

- talk to me in person, or
- send me an email, or,
- submit your feedback anonymously through `http://goo.gl/forms/f0wUD5aZSK`.

First day assignments

1. Do Hw0 (due 2/2, Thursday, 11:59pm). Use the Math 39 webpage[13] to review any material you might have forgotten.
2. Take the MATLAB Onramp tutorial[14]

[13] https://www.sjsu.edu/faculty/guangliang.chen/Math39.html

[14] https://matlabacademy.mathworks.com/details/matlab-onramp/gettingstarted?s_cid=learn_ONRAMP_BAN

Questions?