Math 313 Statistics for Data Science

Welcome to Fall 2023!

Guangliang Chen

Associate Professor

cheng@hope.edu

Hope College, Fall 2023

Presentation Overview

1. Introductions
   - Know your professor
   - Please tell us who you are
2. Syllabus information
3. What is statistical learning
4. Takeaways

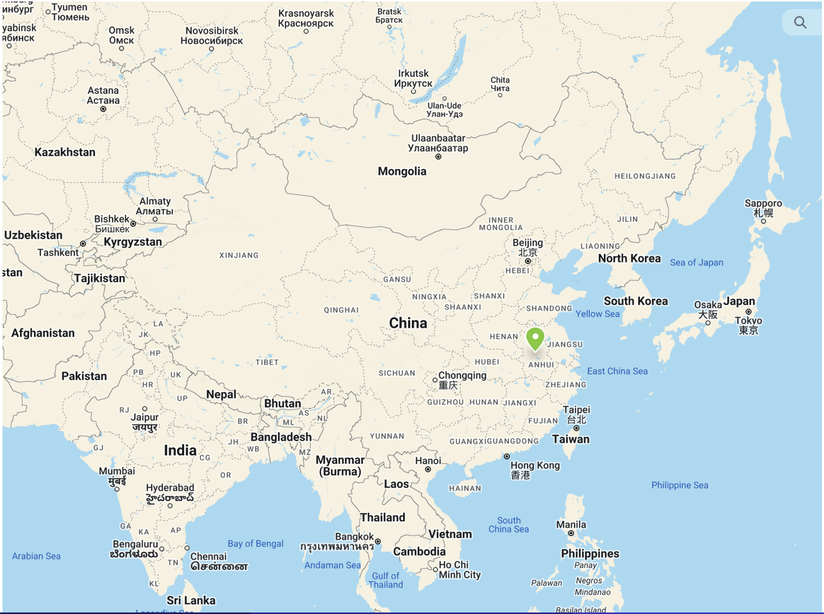
One page summary

Name: Guangliang Chen, PhD

Current title: Associate Professor of Statistics and Data Science, Hope College, 07/2023 –

- Born and grew up in China (till finishing college)
- Attended graduate school and worked in the U.S. (since 2003)
- Born into an atheist family; became a Christian in the U.S. (faith has been an important part of my life)
- A father of three: Mina (11), Zachary (9), Noah (7)
- Biggest passion is teaching, with 14 years of experience in applied math, statistics, and data science
- Research interest in machine learning (algorithms, computing, and applications)

My hometown in China

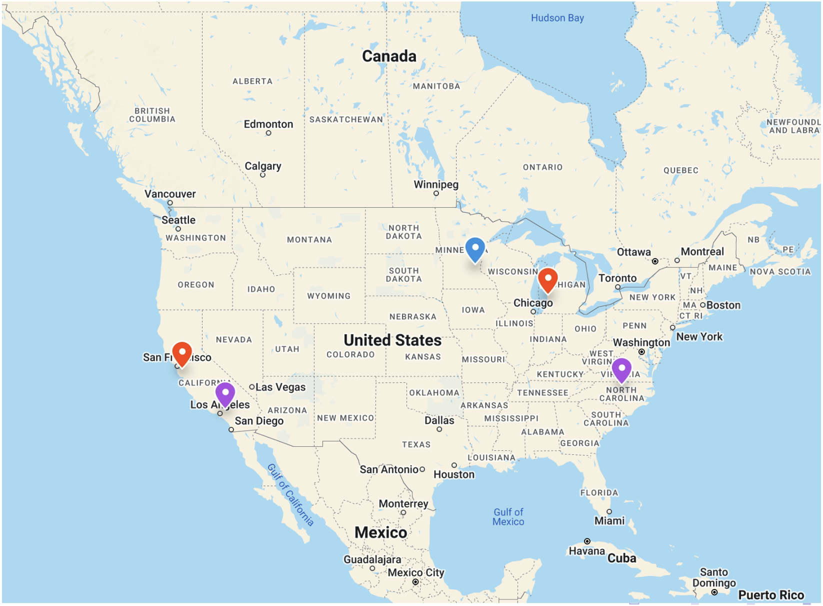


My academic journey

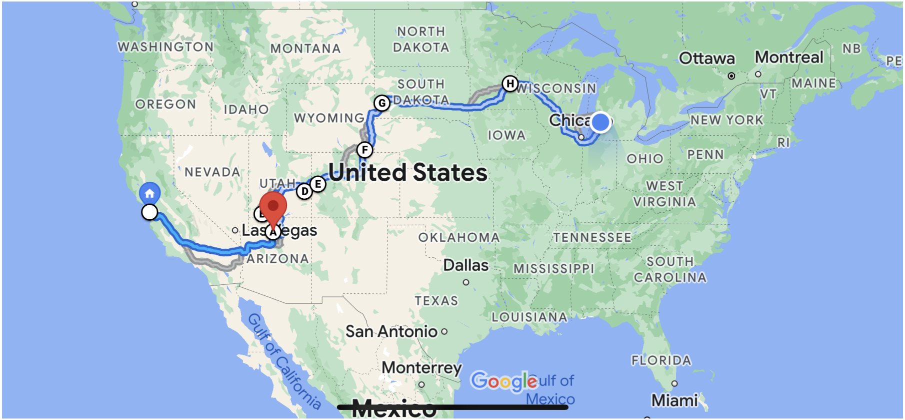
Degrees and previously held positions:

- **B.S. Math**, University of Science and Technology of China, Hefei, Anhui, 2003
- **Ph.D. Applied Math**, University of Minnesota, Minneapolis, 2009
- **Visiting Assistant Professor of Mathematics**, Duke University, 2009–2013
- **Visiting Assistant Professor of Mathematics**, Claremont McKenna College, 20013–2014
- **Associate Professor of Statistics (with tenure)**, San Jose State University, California, 2014–2023

Stops in the U.S.



I have come a really long way



My teaching experience

This semester at **Hope**, I am teaching

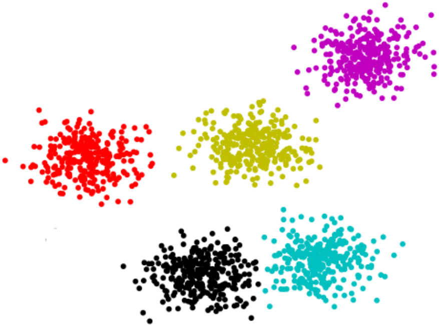
- *Math 245: Linear Algebra and Applications*
- *Math 313: Statistics for Data Science*

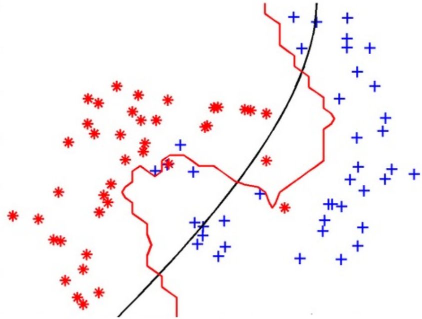
In the past,

- **SJSU**: linear algebra, discrete math, probability theory, mathematical statistics, regression, stochastic processes, mathematical data visualization, statistical and machine learning classification
- **Claremont McKenna**: Calculus, statistics
- **Duke**: calculus, differential equations, linear algebra

My research areas

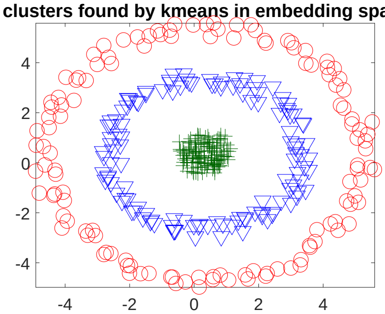
I work in multiple areas of machine learning, such as **clustering** and **classification**, with applications to image processing and documents analysis.

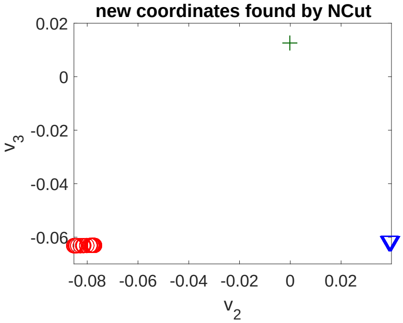




My recent (and ongoing) work

Recently I have been working on the memory and speed scalability of spectral clustering in the setting of massive data.


clusters found by kmeans in embedding space


new coordinates found by NCut

Another field I am interested in the Graph Neural Networks.

Other professional achievements I am proud of

At San Jose State,

- I developed a few clustering algorithms that are scalable to large data sets
- I created three machine learning courses:
  - *Math 185: Learning from large data*
  - *Math 250: Mathematical Data Visualization*[1]
  - *Math 251: Statistical and machine learning classification*[2]
- I designed a M.S. Data Science degree (joint with CS) and served the program in advising and admission capacities
- I was an advisor for B.S. Stats majors for many years.

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

[2] https://www.sjsu.edu/faculty/guangliang.chen/Math251.html

It is now your turn

Please tell us the following:

- **Your name**
- **major**
- **academic year**
- **hometown**
- **hobbies**, and
- **a moment you enjoyed the most during the summer**

Textbook

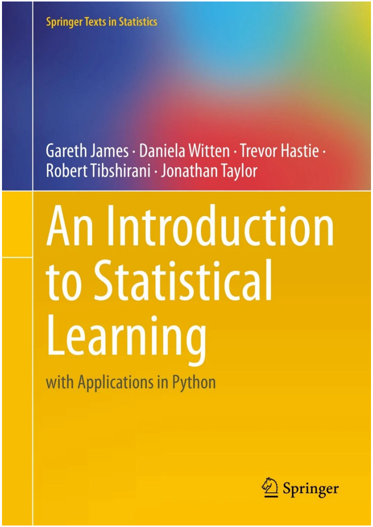
A very new, accessible, popular (and free!) book written by experts.

Title: *An Introduction to Statistical Learning with Applications in Python*[a]

Authors: Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, Jonathan Taylor

Print date: July 5, 2023

[a] https://www.statlearning.com/


Springer Texts in Statistics
Gareth James · Daniela Witten · Trevor Hastie · Robert Tibshirani · Jonathan Taylor
An Introduction to Statistical Learning
with Applications in Python
Springer

Resources accompanied by the book

There are lots of useful resources on the book website:

- Slides
- Data sets
- Figures
- Python notebook files
- ISLP package (in Python)
- Online course (with recorded lectures by the authors)

Computing

We will use Python as the programming language for the course (no prior knowledge or experience is required).

According to the TIOBE Programming Community Index,[3] Python is currently the most popular language worldwide.

Learning resources:

- The textbook
- An Informal Introduction to Python[4]
- How to Use Jupyter Notebook: A Beginner's Tutorial[5]
- Python Tutorial for Beginners (by Mosh)[6]

[3] https://www.tiobe.com/tiobe-index/

[4] https://docs.python.org/3/tutorial/introduction.html

[5] https://www.dataquest.io/blog/jupyter-notebook-tutorial/

[6] https://youtu.be/_uQrJ0TkZlc

Course requirements

The following are required components of the course:

- **Class work**
- **Labs**
- **Applied problems**
- **Midterm Exam 1 (Friday, Oct 13)**
- **Midterm Exam 2 (Wednesday, Nov 22)**
- **Project (proposal, presentation, and report)**

You are also required to attend 4+ approved math/stats events outside of class (otherwise you will lose 3% in your final grade), e.g.,

- 10/11/23 Wed 5:30 PM - Francis Su Public Lecture in Winants
- 10/12/23 Thursday 11:00AM - Student Colloquium in VWF 102

My teaching approach

I strive to make classes interactive and incorporate **active learning** whenever possible.

On a typical day:

- You read the required book section before class (this is where initial learning occurs)
- I will give several mini lectures (to clarify/supplement the book material, and/or demonstrate how to solve a problem)
- You will be divided into groups to discuss the topics and work on problems in class
- After class, there will be homework assigned

Note that you are expected to study at least 6 hours outside of class each week.

Course syllabus

Any questions?

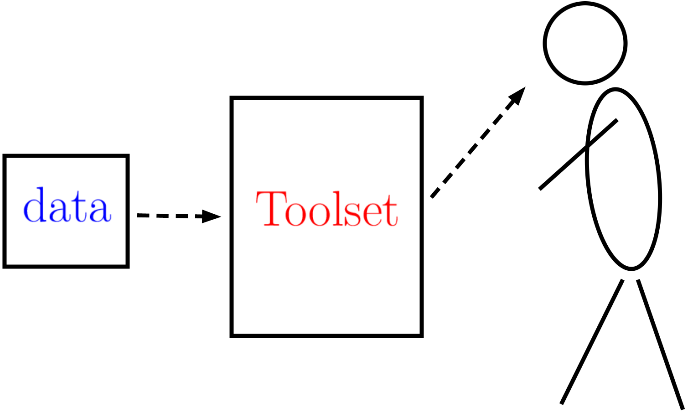
What is this course about?

This course is an introduction to statistical learning.

Okay, then what is statistical learning?

What is statistical learning?

Briefly, statistical learning refers to a diverse set of (statistical) tools for learning from data.



What does the course cover?

The course covers the following areas:

- **Regression**: Continuous (quantitative) response
- **Classification**: Categorical (qualitative, discrete) response
- **Clustering** (no response, only inputs)
- **dimensionality reduction** (with or without response)
- **Model selection** techniques

Regression and Classification are under supervised learning (which requires training data) while clustering is unsupervised.

Dimensionality reduction can be supervised or unsupervised.

Question

Do you recall any real-life example for each of the statistical learning tasks below:

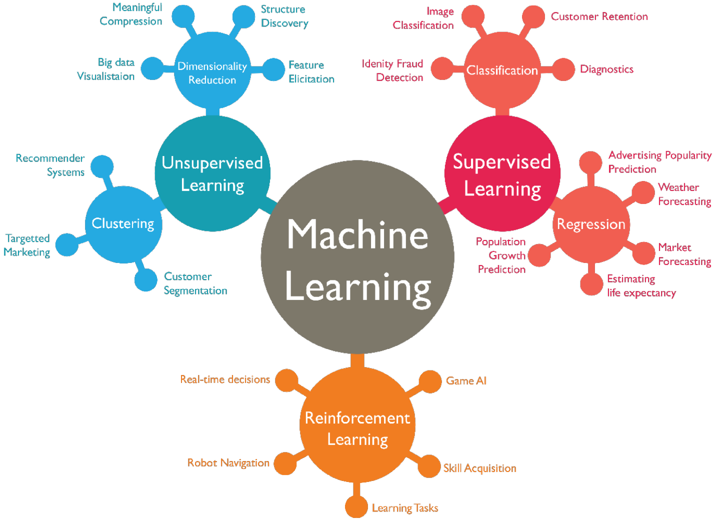
- **Regression**
- **Classification**
- **Clustering**

(more examples)[7]

[7] https://hastie.su.domains/ISLR2/Slides/Ch1_Inroduction.pdf

Machine learning

… is the vast field of using machines to learn from (large) data.



Statistical learning versus machine learning

The two fields originated from different disciplines (Stats vs CS/Engineering), but are largely the same nowadays with the following subtle distinctions:

- Statistical learning emphasizes on statistical assumptions, formulations and tasks (distributions, prediction, and inference) while machine learning focuses on prediction in settings of large, complex data, assisted by computing power

Statistical learning versus machine learning

The two fields originated from different disciplines (Stats vs CS/Engineering), but are largely the same nowadays with the following subtle distinctions:

- Statistical learning emphasizes on statistical assumptions, formulations and tasks (distributions, prediction, and inference) while machine learning focuses on prediction in settings of large, complex data, assisted by computing power
- People in the two communities speak slightly different languages, e.g., observation vs example/instance, predictor vs feature/input, response vs label/output, error vs loss, fitting vs training/learning

Statistical learning versus machine learning

The two fields originated from different disciplines (Stats vs CS/Engineering), but are largely the same nowadays with the following subtle distinctions:

- Statistical learning emphasizes on statistical assumptions, formulations and tasks (distributions, prediction, and inference) while machine learning focuses on prediction in settings of large, complex data, assisted by computing power
- People in the two communities speak slightly different languages, e.g., observation vs example/instance, predictor vs feature/input, response vs label/output, error vs loss, fitting vs training/learning
- Methodology also tends to differ: Statistical learning uses tools like mixture modeling, MLE, and Bayesian inference, while machine learning relies on optimization and gradient descent. Interestingly, they sometimes lead to the same algorithms.

Statistical learning versus machine learning

The two fields originated from different disciplines (Stats vs CS/Engineering), but are largely the same nowadays with the following subtle distinctions:

- Statistical learning emphasizes on statistical assumptions, formulations and tasks (distributions, prediction, and inference) while machine learning focuses on prediction in settings of large, complex data, assisted by computing power
- People in the two communities speak slightly different languages, e.g., observation vs example/instance, predictor vs feature/input, response vs label/output, error vs loss, fitting vs training/learning
- Methodology also tends to differ: Statistical learning uses tools like mixture modeling, MLE, and Bayesian inference, while machine learning relies on optimization and gradient descent. Interestingly, they sometimes lead to the same algorithms.
- Machine learning sounds more attractive (and easier?) than statistical learning.

A little bit of notation

Notation

Matrices are denoted by **boldface** UPPERCASE letters ($\mathbf{A}, \mathbf{B}$, etc.);
column vectors: **boldface** lowercase ($\mathbf{a}, \mathbf{x}$, etc.);
row vectors: *plain* lowercase with arrow on top ($\vec{x}_1, \vec{x}_2$, etc.);
scalars: *plain* lowercase (x_1, x_2, a, b);
random variables: *plain* UPPERCASE letters (X, Y, Z, X_1, X_2)

We say that a vector $\mathbf{a}$ with n entries has dimension n or size $n \times 1$, and denote it by $\mathbf{a} \in \mathbb{R}^n$.

We say that a matrix $\mathbf{A}$ with m rows and n columns has size $m \times n$, and denote it by $\mathbf{A} \in \mathbb{R}^{m \times n}$.

The data matrix (inputs):

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \in \mathbb{R}^{n \times p}$$

where

- n: number of observations, instances, examples, etc.
- p: number of features, predictors, variables, dimensions etc.
- Columns (features) are denoted by $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p \in \mathbb{R}^n$
- Rows (instances) are denoted by $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n \in \mathbb{R}^p$.

The response vector (outputs in the setting of regression):

$$\mathbf{y} = \begin{bmatrix} y_1 & y_2 & \ldots & y_n \end{bmatrix}^T \in \mathbb{R}^n$$

Some final words

- Statistical learning is fun and exciting (we have a whole semester devoted to it).
- We will focus primarily on the concepts, procedures, model fitting and interpretations (rather than theory).
- You will learn Python programming along with myself (useful for you to find jobs).
- A lot of hard work is expected but I will provide all sorts of guidance and support.

Any questions?

Next time

Section 2.1 – What is statistical learning

Before class please

- Read the textbook
- Go over the slides
- Watch the online lecture