MAT 758 Statistical Machine Learning

Wei Li

Syracuse University

Spring 2024

OVERVIEW

What is Machine Learning?

The Driving Forces

What is "Special" about this Course?

Types of Learning Problems

What is Machine Learning?

What is Machine Learning?

Data Mining: the process of identifying valid, novel, potentially informative, and ultimately useful patterns in data.

Machine Learning: a field of study of the methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under computational constraints.

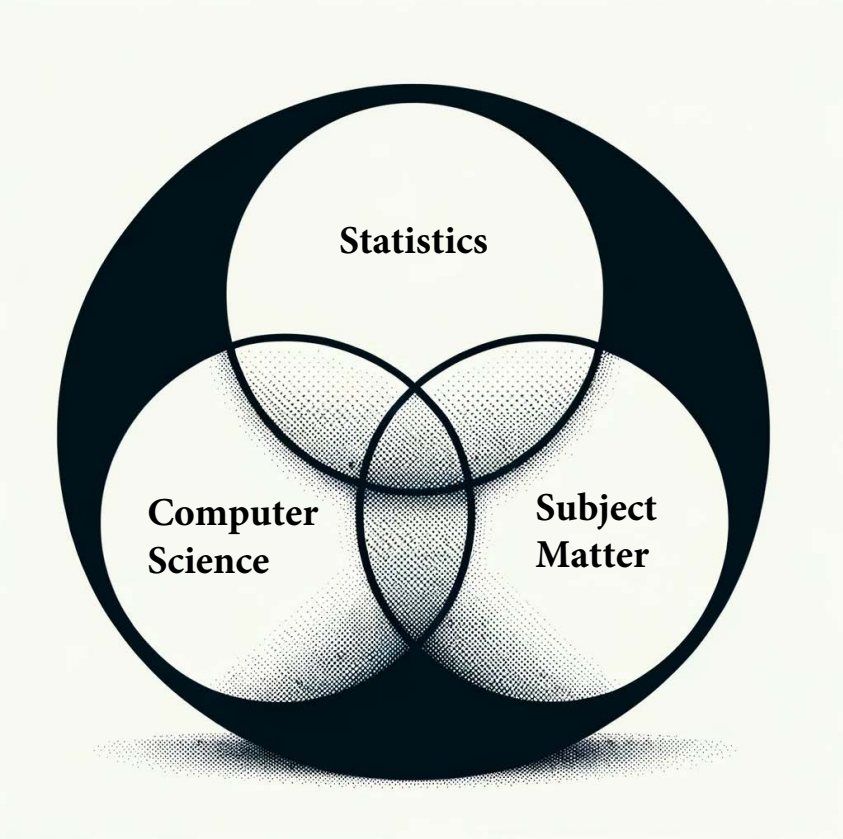
- prediction: its success is to be judged on the future data (unseen data)
- automation: its performance at certain tasks can improve with its experience with the tasks done in the past
- computation: how well (stable) the algorithm performs? does it scale up?

Statistics

Statistics: Statistics is a mathematical body of science that pertains to the collection, analysis, interpretation or explanation, and presentation of data, or as a branch of mathematics (Wiki). While many scientific investigations make use of data, statistics is concerned with the use of data and the properties of the methods in the context of sampling uncertainty.

- provide theoretical foundations for learning algorithms
- give useful tools to analyze an algorithm's statistical properties and performance guarantee
- help researchers gain deeper understanding of the approaches, design better algorithms, and select appropriate methods for a given problem

Statistics and other disciplines



Intersections: computational physics, analytical chemistry, econometrics, business analytics, sports analytics, cliometrics, medical radiology/pathology, genetics...

ML/AI are becoming prevalent tools...

The Driving Forces

The Driving Forces

- explosive growth of data in a great variety of fields
  - revolution in biotechniques (DNA microarray, GWAS, next generation sequencing)
  - internet, network, search engines, digital images, multi-media information
- rapidly increasing computer power
  - cheaper storage devices with higher capacity
- faster communications; better database management systems

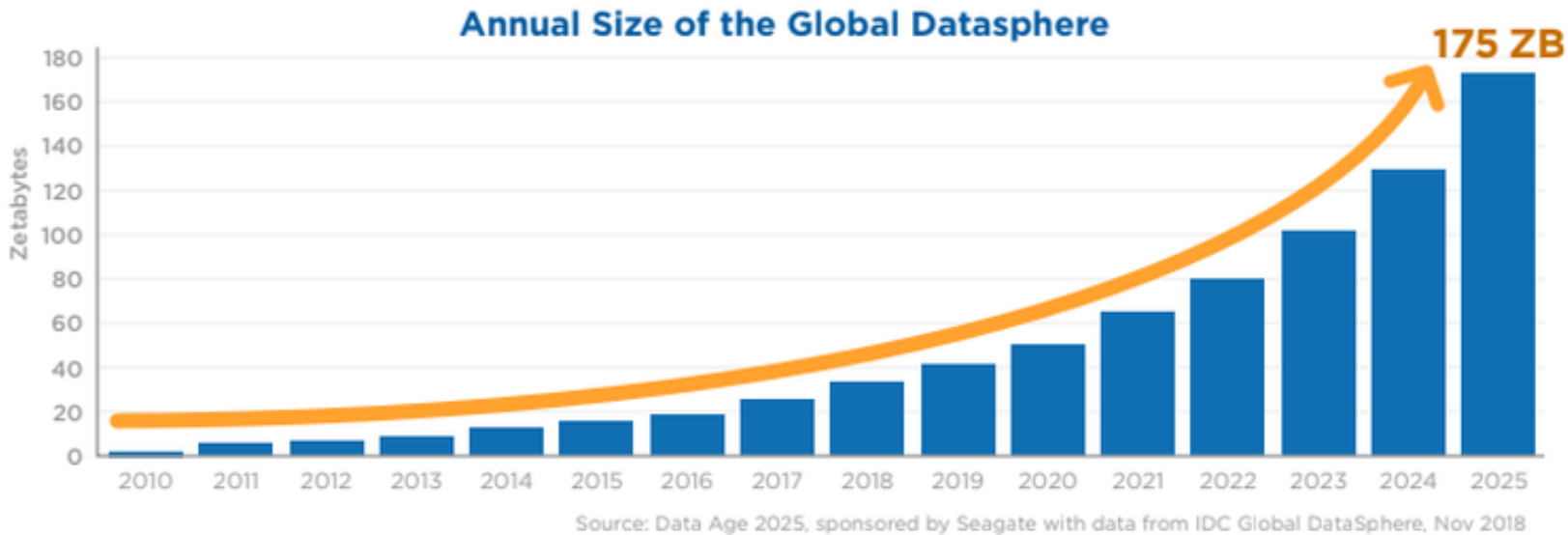
Big data

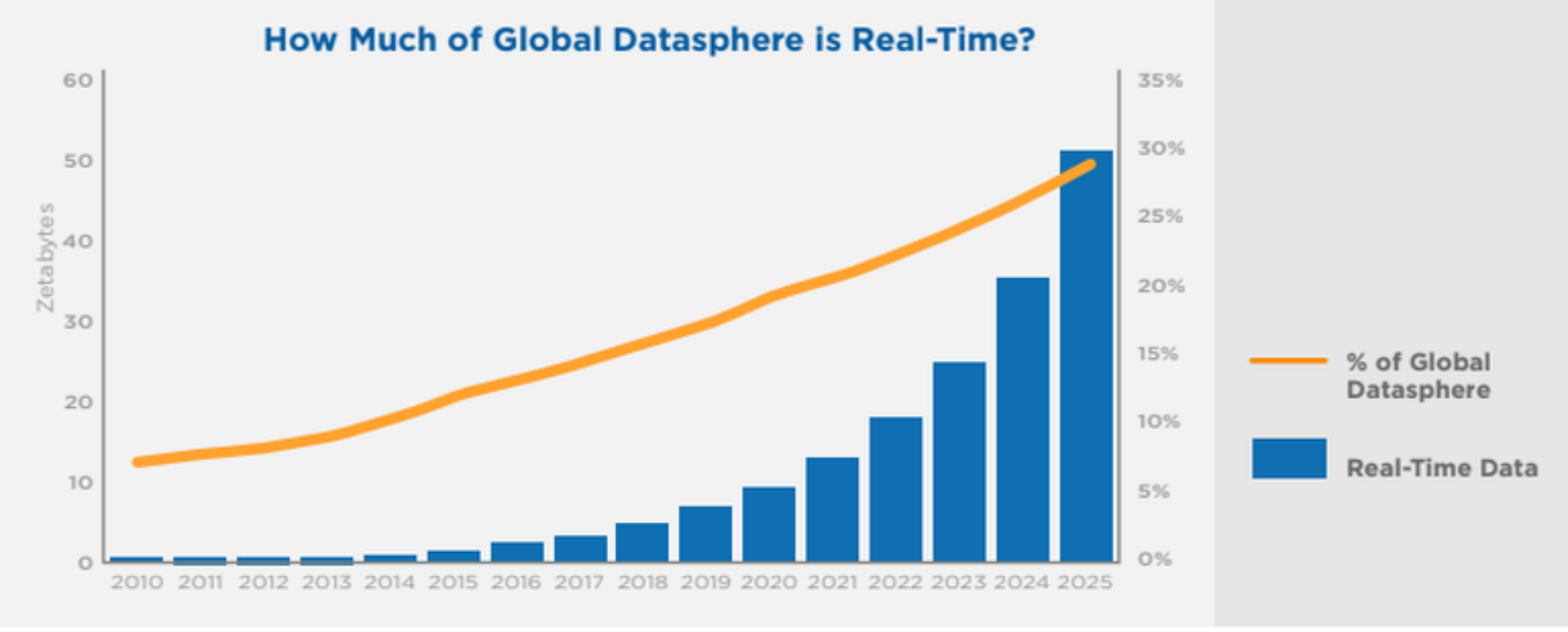
The sizes of modern datasets are increasing faster than ever:

Bytes = 8 bits, kilobyte = 10^3 bytes, Megabyte = 10^6 bytes, Gigabyte = 10^9 bytes, Terabyte = 10^{12} bytes, Petabyte = 10^{15} bytes, Exabyte = 10^{18} bytes, Zettabyte = 10^{21} bytes, Yottabyte = 10^{24} bytes, ...

- 1 Megabytes: A short novel
- 5 Megabytes: A high resolution photograph
- 500 Gigabytes: a common storage volume of a USB drive
- 24 Terabytes: Amount of video data uploaded to YouTube per day (2016)
- 11 Petabytes: How much e-mail information per year
- 5 Exabytes: All words ever spoken by human beings

Figure 1 – Annual Size of the Global Datasphere





Source of image: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

Examples of big data

- Wal-Mart made > 20 million transactions daily, and constructed an 11 terabyte database of customer transactions
- AT&T had 100 million customers and carried on the order of 300 million calls a day on its long distance network
- molecular data: DNA copy-number alteration, mRNA expression, protein expression
- images, network data, tweet data, ...
- there are about 1 trillion web pages ...

Examples of learning problems

- Predict whether a patient hospitalized due to a heart attack will have a second heart attack, based on demographic, diet, and clinical measurements. (Regression and Classification)
- Predict the price of a stock in 6 months, based on company performance measures and economic data. (Regression)
- Identify a handwritten ZIP code from a digitized image. (Classification)
- Identify risk factors for prostate cancer based on clinical, genetic, and demographic variables. (Feature Selection)
- Identify groups of genes with similar functions from DNA microarray data. (High-Dimensional Clustering)
- In fraud detection, determine which covariates are useful in building a model to predict the probability of an order being fraudulent. (Classification Problem)

What is "Special" about this Course?

What is "Special" about this Course?

To combine the "art" of designing good learning algorithms and the "science" of analyzing statistical properties and performance of the approaches

- emphasize "statistical" principles behind the approaches and algorithms
- learn how to formulate a learning problem in a "statistical" framework
- understand existing techniques from a "statistical" perspective; how can one quantify the uncertainty related to the solutions? what are the limitations and strengths for different methods?

The goals of this course

- learn basic statistical concepts and principles
  - statistical inference on uncertainty, distribution, loss, risk
  - model building, evaluation, selection, prediction, and optimality; bias-variance trade-off
  - hyperparameter tuning, training error, test error, cross-validation
- learn statistical and machine learning methods for big data
  - lasso, boosting, tree, random forest, neural network...
  - causality
- learn using Python and related modules to analyze data

Types of Learning Problems

Types of Learning Problems

Typically, we collect data $(X_i, Y_i), i = 1, \cdots, n$.

- Y_i is the outcome or response variable.
- X_i is the input or prediction variables.

Various Learning Problems

- Supervised learning (Y observed)
  - Regression (Y quantitative, a numerical quantity)
  - Classification (Y qualitative, a class label)
- Unsupervised learning (Y unobserved)
  - Density estimation (no Y)
  - clustering (no Y)
- Semi-supervised learning (Y partially available)
- Reinforcement learning (maximizing certain cumulative reward)

	Supervised Learning	Unsupervised Learning
Response Y	observed	unobserved
Major Goal	predict Y , given the observed input X	find interesting patterns in data
Examples	linear regression nonparametric regression classification	clustering density estimation dimension reduction

Example 1: email or spam (ESL)

- Training data: 4,601 email messages with known email type.
  - Outcome Y : -1 = email , $+1$ = spam
  - Input X : the relative frequencies of 57 of the most commonly occurring words and punctuation marks in the email message.

TABLE 1.1. *Average percentage of words or characters in an email message equal to the indicated word or character. We have chosen the words and characters showing the largest difference between* `spam` *and* `email`.

	george	you	your	hp	free	hpl	!	our	re	edu	remove
spam	0.00	2.26	1.38	0.02	0.52	0.01	0.51	0.51	0.13	0.01	0.28
email	1.27	1.27	0.44	0.90	0.07	0.43	0.11	0.18	0.42	0.29	0.01

Possible classification rules:

- if (george < 0.6)&(you > 1.5), then spam; otherwise email.
- if (0.2 you -0.3 george) > 0, then spam; otherwise email.

Two types of decision errors:

(1) false positive: classify email to spam (filter out email)
(2) false negative: classify spam to email (email box jammed)

Example 2: prostate cancer (ESL)

To examine the correlation between the level of prostate specific antigen (PSA) and a number of clinical measures

- Training data: 97 male patient with different stages of prostate cancer.
  - Outcome Y : the log of the level of prostate specific antigen (lpsa)
  - Input X : eight clinical measures: log-cancer volume (lcavol), log prostate weight (lweight), age, and the other five.
    - log cancer volume (lcavol)
    - log prostate weight (lweight)
    - age
    - log bengin prostatic hyperplasia (lbph)
    - seminal vesicle invasion (svi)
    - log of capsular penetration (lcp)
    - Gleason score (gleason)
    - percent of Gleason scores 4 or 5 (pgg45)

Questions of interest:

- What is the relationship between PSA and clinical measures?
- Is the linear model sufficient? Nonlinear effect? Interactions?
- Which clinical measures are more relevant to the prediction?

Example 3: handwritten digit recognition (ESL)

- Goal: identify single digits $0 \sim 9$ based on the images.
- Raw Data: images that are scaled segments from five digit Z IP codes.
  - Each digit has an image of 16×16 eight-bit gray scale maps
  - Pixel intensities (grayscale values) range from 0 (black) to 255 (white)
  - Images are normalized to have approximately the same size and orientation
- Output $Y \in \{0, 1, \ldots, 9\}$
- Input X is a 16×16 matrix, or a 256 dimensional vector.

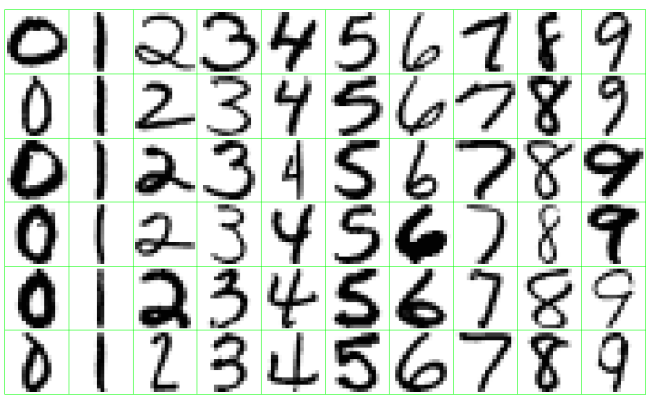


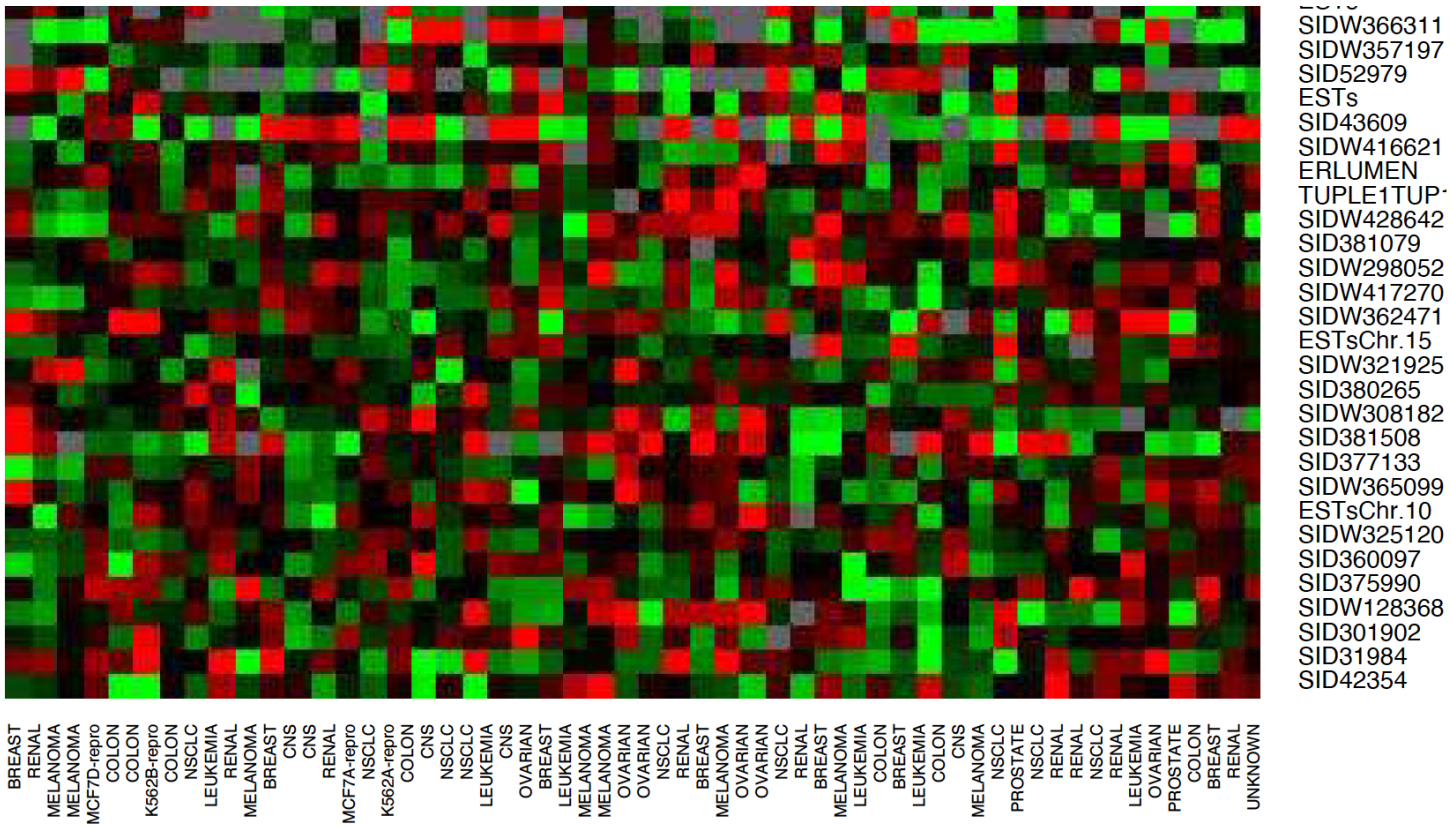
FIGURE 1.2. *Examples of handwritten digits from U.S. postal envelopes.*

Example 4: DNA expression microarray (ESL)

DNA Microarray Technique: for each sample from a tissue, the expression level (the amount of mRNA) of thousands of genes are measured.

A gene expression dataset collects together the expression values from a series of DNA microarray experiments, with each column representing an experiment. There are therefore several thousand rows representing individual genes, and tens of columns representing samples.

- Training data: $p = 6,830$ genes (rows), $n = 64$ samples (columns) (cancer tumors samples).
  - Input X : the level of expression for each gene
- Goal: discover the relationship between gene and cancer type, or find the gene signature of each cancer subtype
- Challenge: $p >> n$ (data matrix $n \times p$ "short and fat" ')



ESL Figure 1.3.

Typical questions of interest:

- Which samples are most similar to each other, in terms of their expression profiles across genes?
- Which genes are most similar to each other, in terms of their expression profiles across sample?
- Do certain genes show very high (or low) expression for certain cancer samples?

This task could be viewed as

- a supervised or an unsupervised problem
- a regression or a classification problem

Example 5: face detection and recognition (PML)

Object detection or object localization: an important special case of this is **face detection**:

- divide the image into many small overlapping patches at different locations, scales and orientations,
- classify each such patch based on whether it contains face-like texture or not. This is called a sliding window detector.
- it is important to be invariant to such details, and to just focus on the differences between faces and non-faces.

Application includes: auto-focus, face blurring-out in Google's street view

Face recognition is to identify or verify the identity of a person using their face. It goes beyond detecting the presence of a face and works to ascertain the identity of the individual.

mask detection

face-mask-12k-images-dataset@kaggle

Notations

Convention (unless otherwise noted):

A random variable or random vector:

- Y: response variable
- X: random variable or random vector
  - if a p-dim random vector, $X = (X_1, \ldots, X_p)^T$.
- Subscript i sometimes used to emphasizes for the ith observation, say the pair (Y_i, X_i), where $X_i = (X_{i1}, \cdots, X_{ip})^T$.

Observed values:

- y_i: the value of response variable for ith observation
- $\boldsymbol{x}_i$: the ith observed value of X
  - $\boldsymbol{x}_i$ could be a scalar of a vector. If a scalar, just x_i.
- $\mathbf{y}$: the n-dim response vector consisting of y_i.

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

- $\mathbf{X}$: the $n \times p$ design matrix
  - ith row is $\boldsymbol{x}_i^T$
  - jth column is $\mathbf{x}_j$

$$\mathbf{X} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1, \mathbf{x}_2 \cdots \mathbf{x}_p \end{bmatrix}$$

All vector are taken as column vectors by default. Generic capital letter or bold-face capital letter will often denote a matrix, e.g., A or $\mathbf{A}$.