IEMS 304 Lecture 1: Introduction to Statistical Learning

Yiping Lu
yiping.lu@northwestern.edu

Industrial Engineering & Management Sciences
Northwestern University

Logistics

Textbook: James G, Witten D, Hastie T, et al. *An introduction to statistical learning.*

CS 229 Lecture Note: https://cs229.stanford.edu/main_notes.pdf

Time and Location: Monday, Wednesday and Friday, 9.00 A.M.- 9.50 A.M. Tech L251

Office Hour: Friday: 1 P.M. Tech M237

TA Office Hour:

Pre-requisite and Pre-test

This is a mathematically intense course. But that's why it's exciting and rewarding!

Pre-requisite: A previous course in statistics at the level of IEMS 303 plus a course in matrix analysis. Comfort with programming (we will be programming in R) is also necessary.

Pre-test: Passing the pretest is worth 3% of your final course grade. You must achieve a passing score of 70% or higher by Monday, Apr 15th at 11:59 p.m. This deadline will be firmly enforced.

Honor Code

Do's

- ❒ form study groups (with arbitrary number of people); discuss and work on homework problems in groups
- ❒ write down the solutions independently
- ❒ write down the names of people with whom you've discussed the homework
- use ChatGPT as a TA

Don'ts

- ❒ It is an honor code violation to copy, refer to, or look at written or code solutions from a previous year, including but not limited to: official solutions from a previous year, solutions posted online, solutions you or someone else may have written up in a previous year, and solutions for related problems.
- ❒ Directly copy the answer from ChatGPT/Claude/Any GenAI

Lab Session

Homework

Publish on course website, due Friday (except pretests/midterm weeks)

Submit on Gradescope

Campusewire

Exams

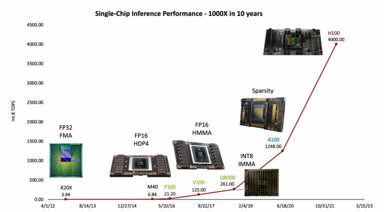
Let's Start

Massive Data

Massive complex data : Images, Acoustic signals, Text, ...

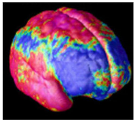
- Wikipedia pages: 13 millions (2014), 57 million (2022)
- Facebook users: 800 million (2014), 2.96 billion (2022)
- Flickr photos: 6 billion (2014), 10 billion (2022)
- Twitter tweets/day: 340 million (2014), 500 million (2022)
- Youtube video/min: 24 hours (2014), 500 hours (2022)
- Google pages: $\geq$ 1 trillion (2014), $\geq$ 130 trillions (2016)

Massive Computing : Huang's Law



Broad Applications in Science and Engineering

Brain

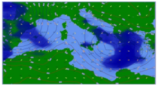


Galaxy

Self-driving car

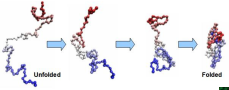
Robot

Genome



Weather

Finance

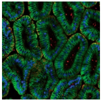


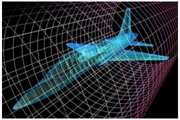
Protein

Music

Sustainability

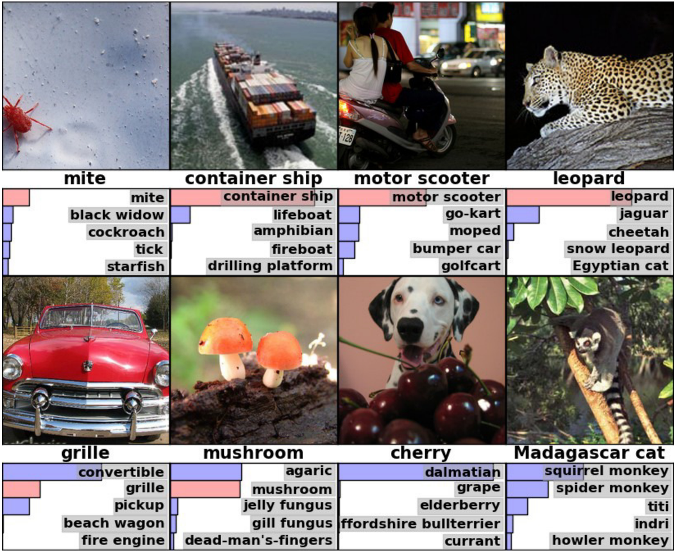
Cell



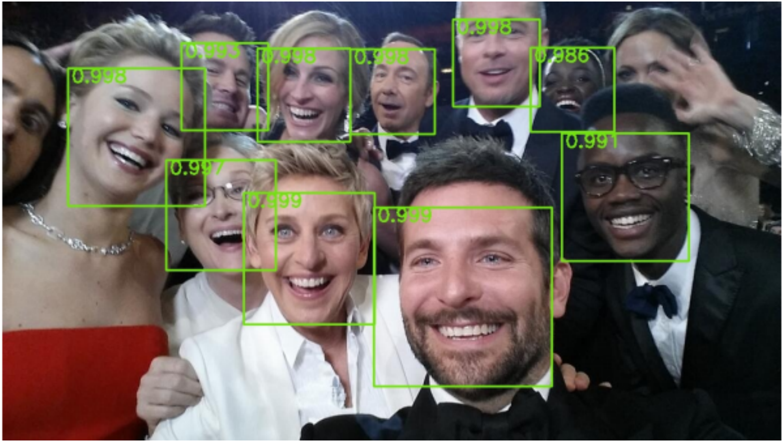


Design

Image Classification



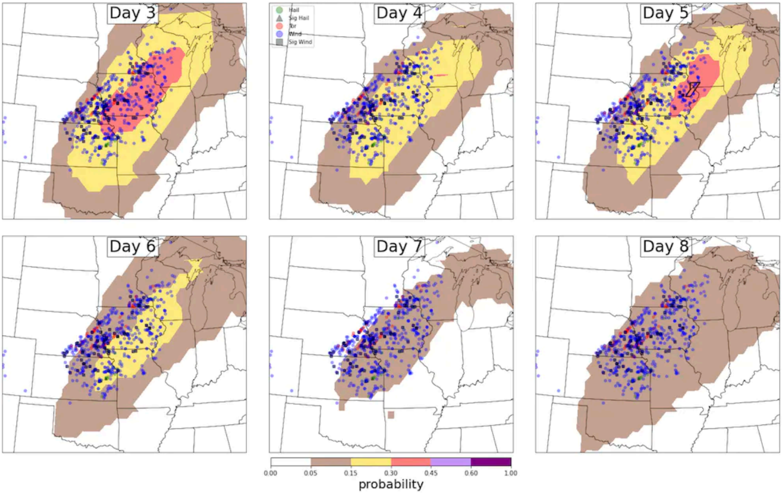
Face Detection



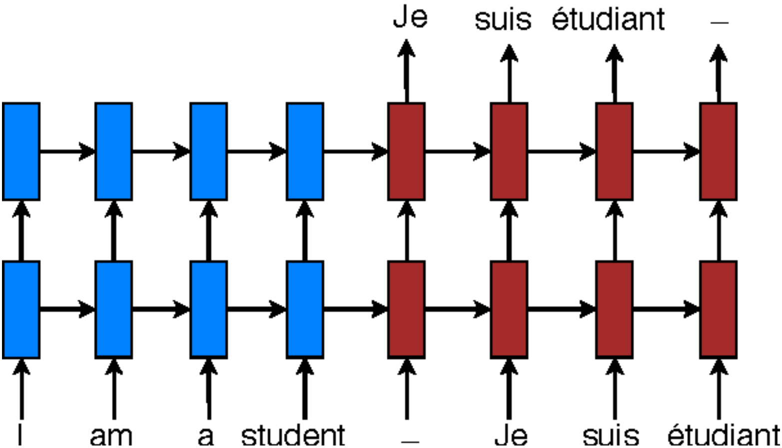
Spam Detection

Subject	Correspondents	Date
URGENT RFQ	AL WALEED EQUIPMENTS	03/13/2017 06:55
	starsescorts@gmail.com	03/15/2017 01:27
New Order Attached **KINDLY SEND INVOICE	Amr Hassan	03/15/2017 19:30
We're sad to let you know that our delivery was unusuccessful....	FedEx Expedited Express	03/16/2017 02:53
47929 username2	pkeith@gejlaw.com	03/16/2017 05:29
Delivery Status Notification	webmaster@stroy-exp...	03/16/2017 05:47
	vowsbyjudy@shaw.ca	03/16/2017 14:38
Formal Inquiry	"Anaïs VANACKER"<Va...	03/16/2017 21:16
We have delivery problems with your parcel #7104543	webmaster@whfarm2....	03/17/2017 00:57
INQUIRY	Saigon Offshore	03/17/2017 03:47
	dava@ac-lyon.fr	03/17/2017 14:25
54343 username	juanro5554@hotmail.c...	03/17/2017 14:48
Item Delivery Notification	alifeof8@server.alifeofj...	00:34
UPS courier can not deliver parcel #004287245 to you	webmaster@stroy-exp...	06:23
Parcel Delivery Notification	abidjanbateau@vps286...	06:52
Visa Card Award	info@visa.com	07:21
Problems with item delivery, n.4930349	Apache	09:54
Package Delivery Notification	Apache	10:06
Delivery Status Notification	contrav8@box980.blue...	17:05

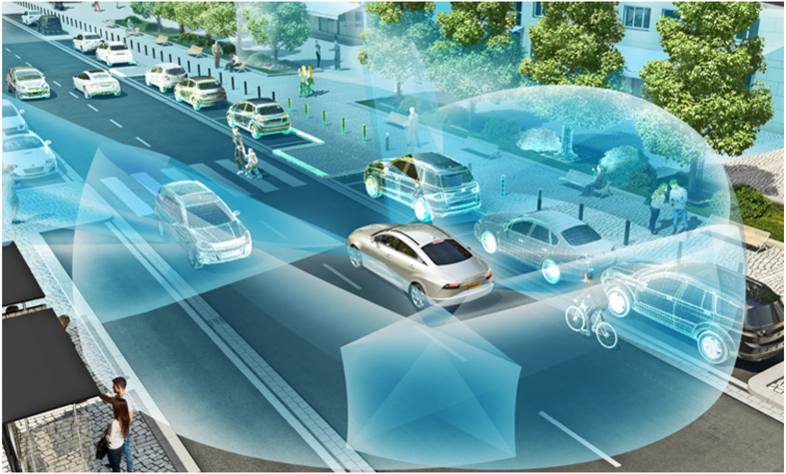
Weather Forecasting



Machine Translation

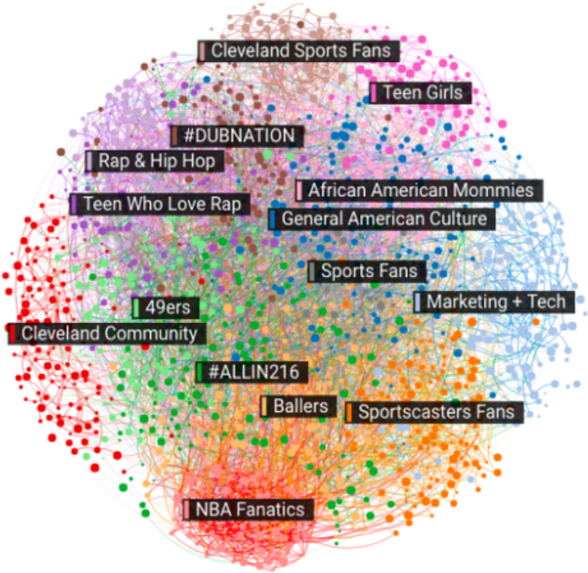


Autonomous Driving



Do We Always have the input output pair

Community Detection



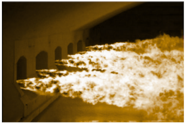
Anomaly Detection

Glass furnace

Product line

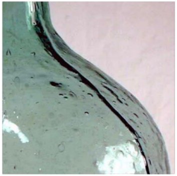
After 24 hours

Something goes wrong



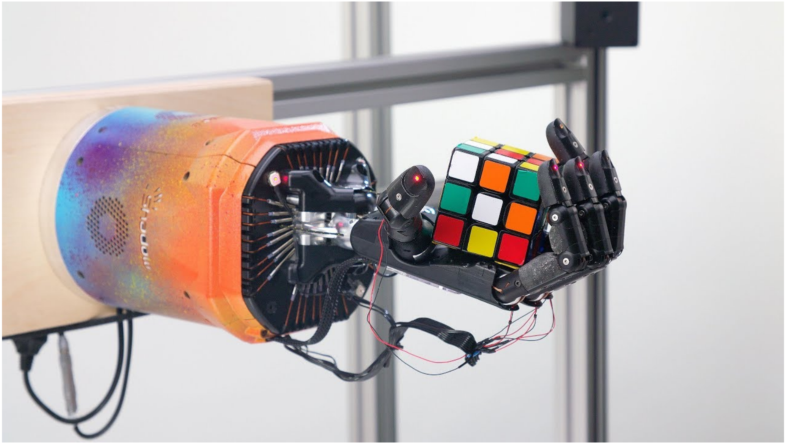
Anomaly propagates

Blisters in glass



Movie Recommendation

Robotics



Chatbot

Art Making

Introduction: Machine Learning

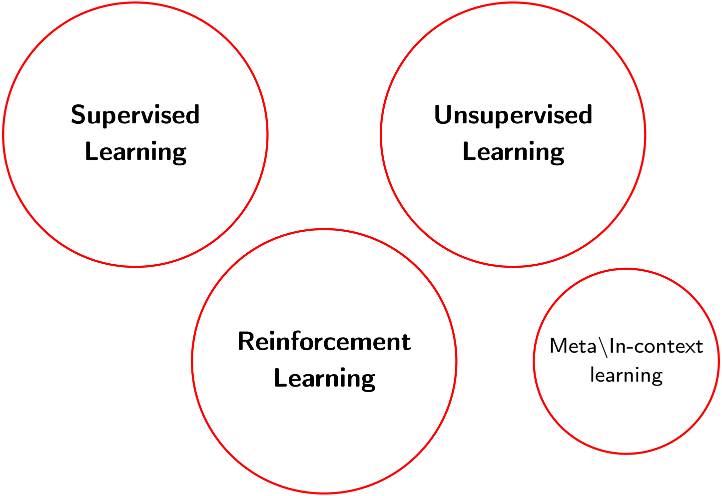
Tom Mitchell (1998): a computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

- ❒ Experience (data): games played by the program (with itself)
- ❒ Performance measure: winning rate

- ☹ We want to provide clear, interpretable models. These models allow you to understand the direct influence of each predictor on the outcome, which is essential in fields where insight into relationships (rather than just prediction) is needed.
- ☹ No confidence interval estimation
- ☹ In cases where data is scarce, simpler parametric models used in statistical learning can perform better. (Why?)

Regression: Predict the Unknown

Taxonomy of Machine Learning



Supervised Learning (Regression)

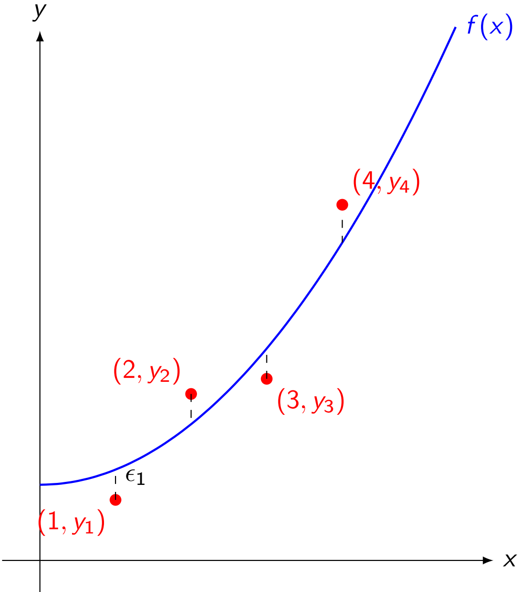
Supervised Learning: a set of observed data points $\{(x_i, y_i)\}_{i=1}^n$, where x_i represents the predictor (or vector of predictors) and y_i represents the response variable. Regression is the process of modeling the relationship between x and y by assuming:

$$y_i = f(x_i) + \epsilon_i,$$

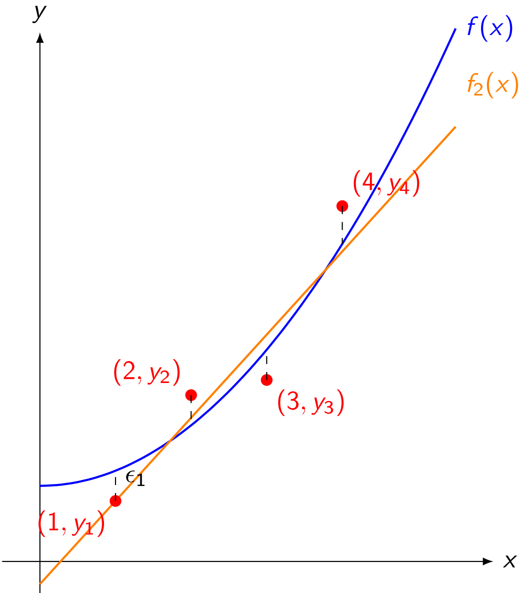
where:

- $f(x_i)$ is an unknown function that describes the systematic component of the relationship
- ϵ_i is a random error term.

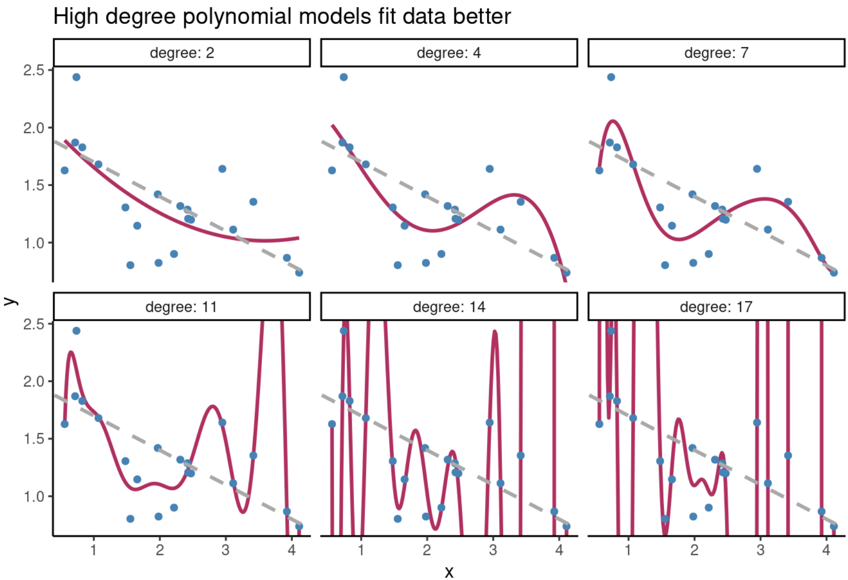
Regression



Regression



Runge Phenomenon

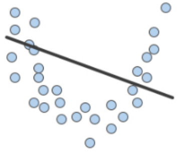
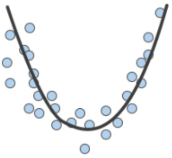
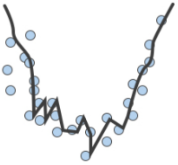
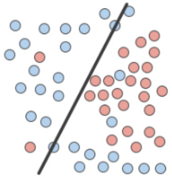
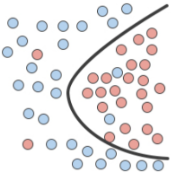
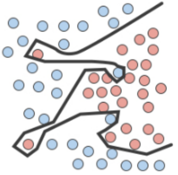


Bias and Variance Trade-off

$$\mathbb{E}\left[(y - \hat{f}(x))^2\right] = \underbrace{\left(f(x) - \mathbb{E}\left[\hat{f}(x)\right]\right)^2}_{\text{Bias}^2} + \underbrace{\mathbb{E}\left[\left(\hat{f}(x) - \mathbb{E}\left[\hat{f}(x)\right]\right)^2\right]}_{\text{Variance}} + \underbrace{\sigma^2}_{\text{Irreducible}}$$

- ☹ An unbiased estimator could still make systematic mistakes – for example, if it overestimates 99% of the time, and underestimates 1% of the time *by a lot*, in expectation it could be unbiased.
- ☹ An unbiased estimator is **not** necessarily better than a biased estimator, because the total error depends on both the bias and variance of the estimator.

Bias and Variance Trade-off

	Underfitting	Just right	Overfitting
Symptoms	• High training error • Training error close to test error • High bias	• Training error slightly lower than test error	• Very low training error • Training error much lower than test error • High variance
Regression illustration			
Classification illustration			

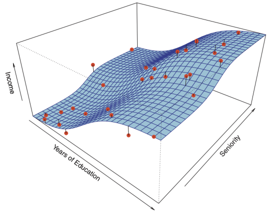
Prediction Accuracy and Model Interpretability

Why would we ever choose to use a more restrictive method instead of a very flexible approach?

High Dimensional Features

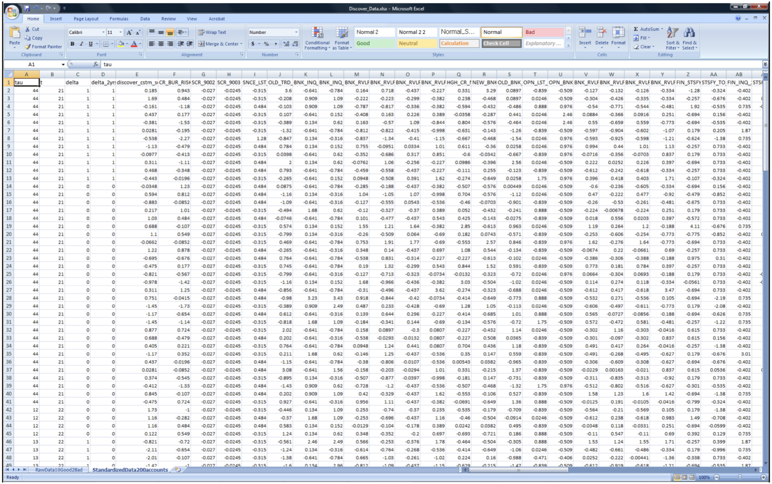
- $x \in \mathbb{R}^d$

$$x = \begin{bmatrix} x_1 & \text{— living size} \\ x_2 & \text{— lot size} \\ x_3 & \text{— \# floors} \\ \vdots & \text{— condition} \\ x_d & \text{— zip code} \end{bmatrix} \longrightarrow y \text{ — price}$$

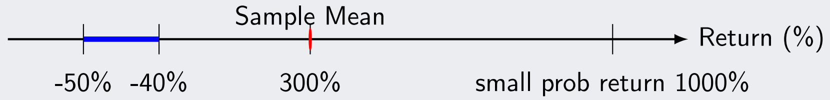


Data as a Matrix

Linear Algebra Reivew this friday!



Confidence Interval in Finance: Impact of Outliers



Why is this Important?

In this example, the confidence interval for the expected return is between -50% and -40%, indicating that most outcomes are negative. However, a very rare event pushes the sample mean to 300%, which could give the false impression of high returns. This discrepancy shows that while the sample mean may appear attractive, the confidence interval reveals the underlying risk and variability in the data, emphasizing the need to consider the full range of possible outcomes when making financial decisions.

Why Sample Mean?

Consider a dataset $x_1, x_2, \ldots, x_n$. We consider L2 loss (or squared error loss) function with respect to a constant c as the performance measure P:

$$L(c) = \sum_{i=1}^{n} (x_i - c)^2.$$

To find the minimizer, differentiate $L(c)$ with respect to c:
$\frac{dL}{dc} = \sum_{i=1}^{n} 2(x_i - c)(-1) = -2\sum_{i=1}^{n}(x_i - c)$. Setting the derivative equal to zero gives:

$$-2\sum_{i=1}^{n}(x_i - c) = 0 \quad \Longrightarrow \quad \sum_{i=1}^{n}(x_i - c) = 0.$$

Expanding the sum:

$$\sum_{i=1}^{n} x_i - nc = 0 \quad \Longrightarrow \quad c = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Thus, the minimizing value of c is the sample mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Different Predicition

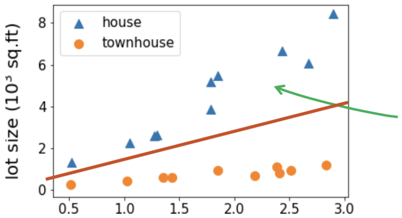
- Point Prediction: retrun $\hat{f}(x)$ since it returns a number.
- Interval Prediction, e.g., Y will be within an interval $[l, u]$ with probability $1 - \alpha$
- distributional prediction, e.g. Y will follow an $N(m, v)$ distribution.

Classification

Classification

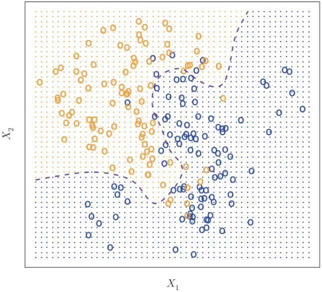
- Regression : if $y \in \mathbb{R}$ is a continuous variable
- classification : the label is a discrete variable

(size, lot size)$\rightarrow$ house or townhouse ?



$y =$ house or townhouse?

Classification as Regression: Bayes Classifier



training error rate: $\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$

Here the function $I(y_i \neq \hat{y}_i)$ is an indicator variable that equals 1, if $y_i \neq \hat{y}_i$ and 0 otherwise. If $y_i \neq \hat{y}_i$, then the i-th observation was classified incorrectly; otherwise it was not misclassified.

Consider random label: $\mathbb{P}(Y = j \mid X = x_0)$. The Bayes classifier returns

$$1 - \max_j \mathbb{P}(Y = j \mid X = x_0)$$

produces the lowest possible test error rate, called the *Bayes error rate* is given by

$$\underbrace{1 - \mathbb{E}\left[\max_j \mathbb{P}(Y = j \mid X)\right]}_{\text{Irreducible}}.$$

x and y in Computer Vision

Task. Image Classification

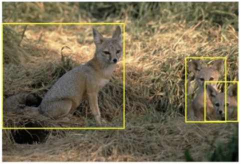
$x = ?, y = ?$

ImageNet Large Scale Visual Recognition Challenge. Russakovsky et al.'2015

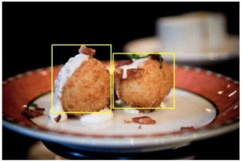
x and y in Computer Vision

Task. Object localization and detection

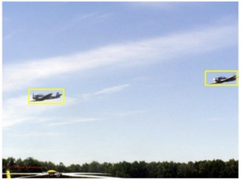
$x = ?, y = ?$



kit fox



croquette



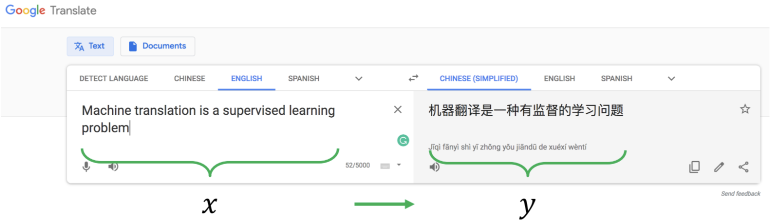
airplane

frog

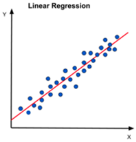
ImageNet Large Scale Visual Recognition Challenge. Russakovsky et al.'2015

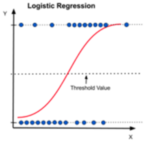
x and y in Natural Language

Task. Machine Translation d $x = {\color{red}?}, y = {\color{red}?}$



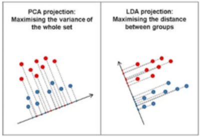
Early History

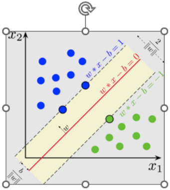




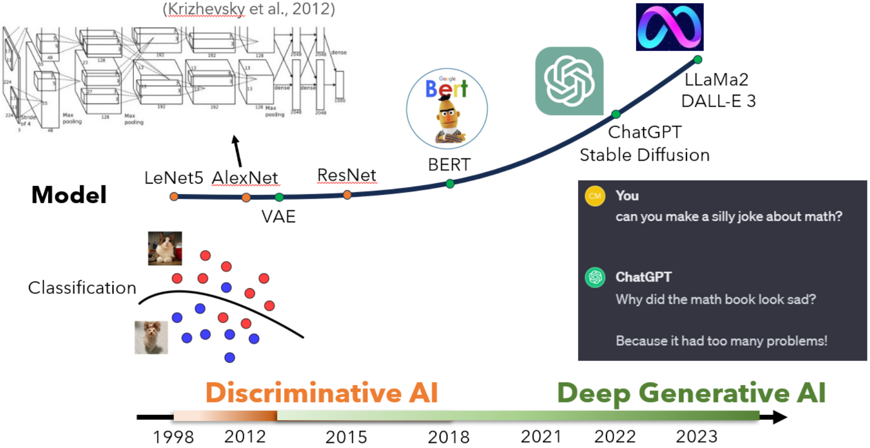
Model

- 1900: Linear regression
- 1936: Linear Discriminant Analysis (LDA)
- 1940s: Logistic regression
- 1970s: Generalized linear model
- 1980s: Classification and regression trees; Neural networks
- 1990s: Support Vector Machine (SVM)





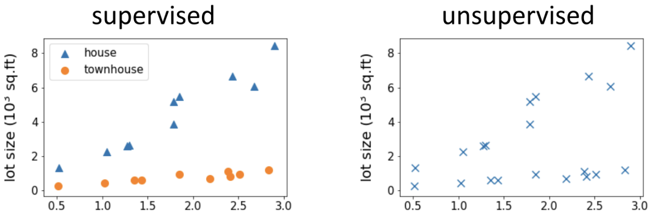
Contemporary Developments



Unsupervised Learning

Unsupervised Learning (Clustering)

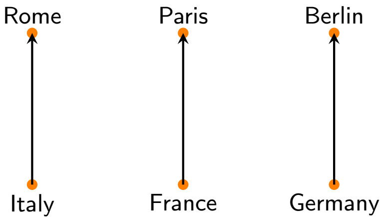
- Dataset contains no labels: $x^{(1)}, x^{(2)}, \cdots, x^{(n)}$
- Goal (vaguely-posed): to find interesting structures in the data



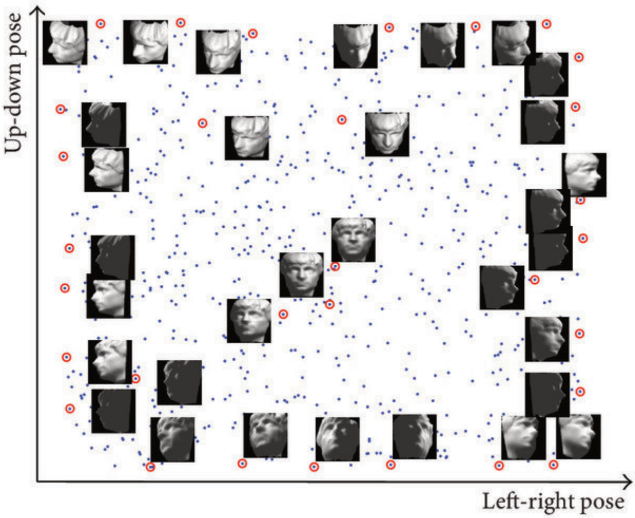
Unsupervised Learning (Feature Extraction)

- **Word** : Encode as vectors
- **Relationship** : represent as direction

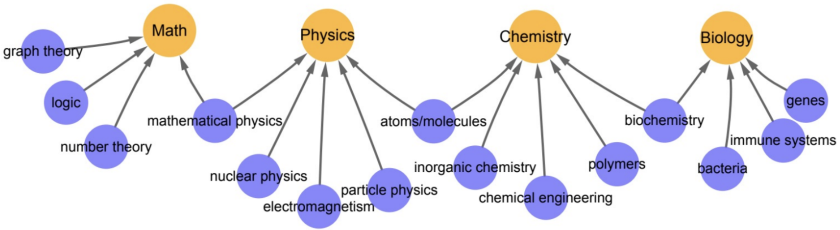
> ▷ word ⟶ encode ⟶ vector
>
> ▷ relation ⟶ encode ⟶ direction



Unsupervised Learning (Feature Extraction)

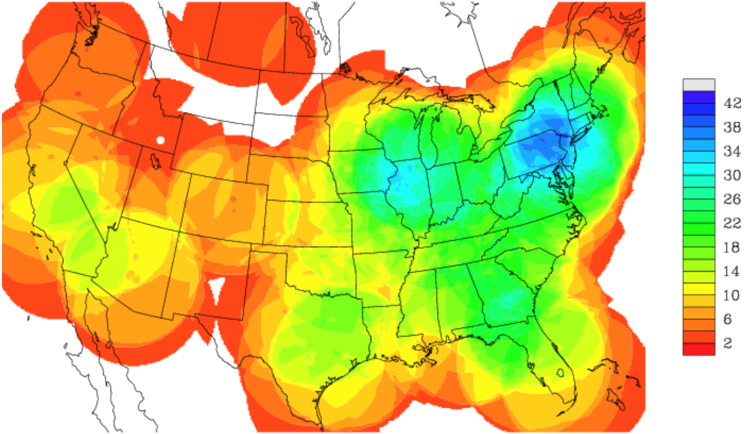


Unsupervised Learning

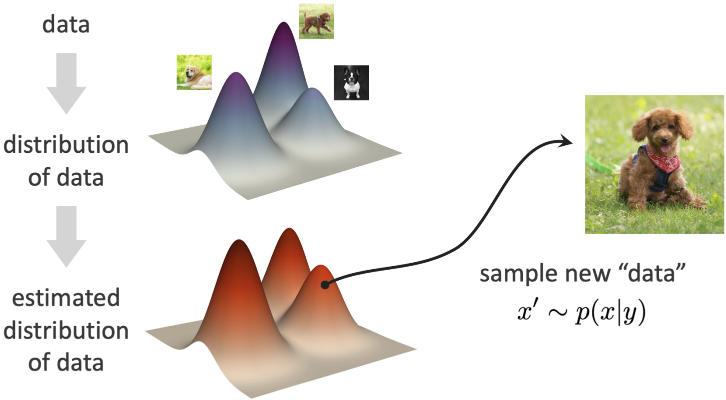


	logic deductive propositional semantics	graph subgraph bipartite vertex	boson massless particle higgs	polyester polypropylene resins epoxy	acids amino biosynthesis peptide
tag	*logic*	*graph theory*	*particle physics*	*polymer*	*biochemistry*

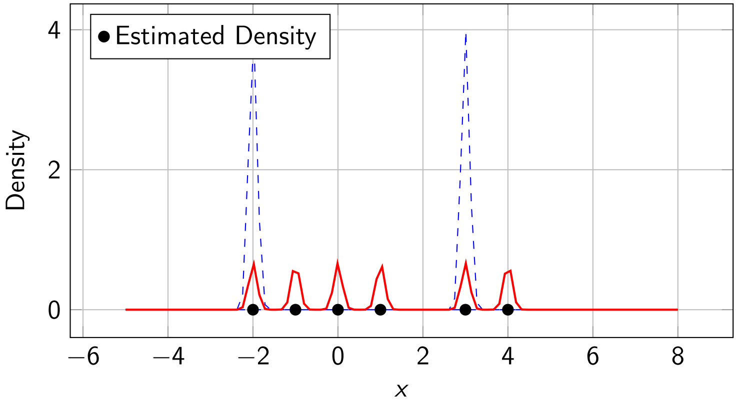
Unsupervised Learning (Density Estimation)



Generative Modeling

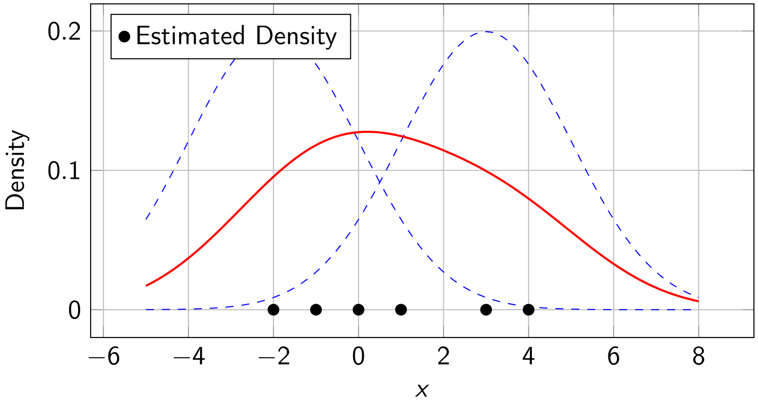


Density Estimation: Bias and Variance Trade-off



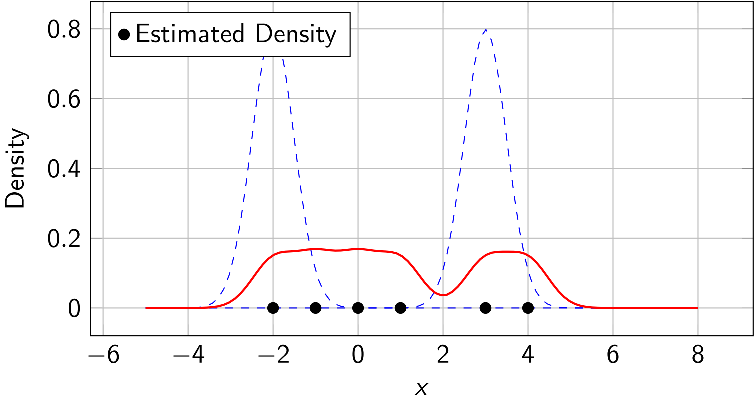
What does this mean in generating images?

Density Estimation: Bias and Variance Trade-off

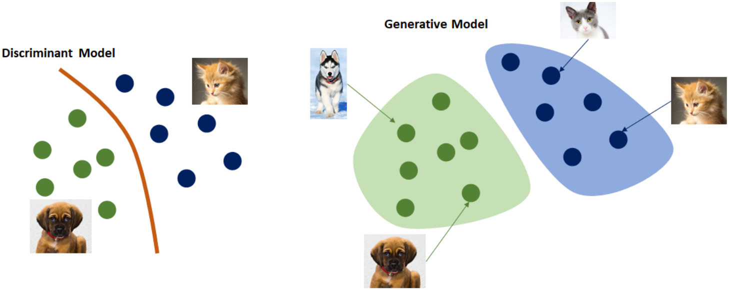


What does this mean in generating images?

Density Estimation: Bias and Variance Trade-off



Generative AI



$$p(y|x) = \frac{p(x,y)}{p(x)} = \frac{p(x|y)p(y)}{p(x)} = p(x|y)\frac{p(y)}{p(x)}$$

Generative AI Case Study: Formulate as $p(x|y)$

- **Text-to-image/video generation**

Prompt: teddy bear teaching a course, with "generative models" written on blackboard

y: text prompt

x: generated visual content

Image generated by Stable Diffusion 3 Medium

Generative AI Case Study: Formulate as $p(x|y)$

- **Text-to-3D structure generation**

Figure credit: Tang, et al. LGM: Large Multi-View Gaussian Model for High-Resolution 3D Content Creation. ECCV 2024

- **Class-conditional image generation**

"red fox" ← y: class label

← x: generated image

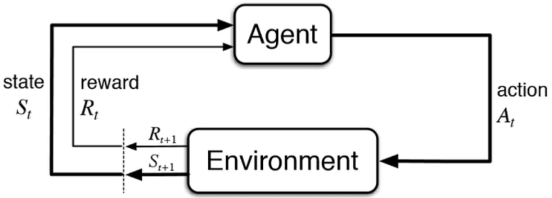
Image generated by: Li, et al. Autoregressive Image Generation without Vector Quantization, 2024

More Examples:

https://mit-6s978.github.io/schedule.html

Reinforcement Learning

Learning to make sequential decisions



mathmetical framework called: *markov decision process*

Not included in IEMS 304

Application of RL: Decision Making

What is the agent? Waht is the action? What is the state? What is the reward?

- AlphaGo
- Robotics