

Computational Data Analysis

Machine Learning

Yao Xie, Ph.D.

Associate Professor

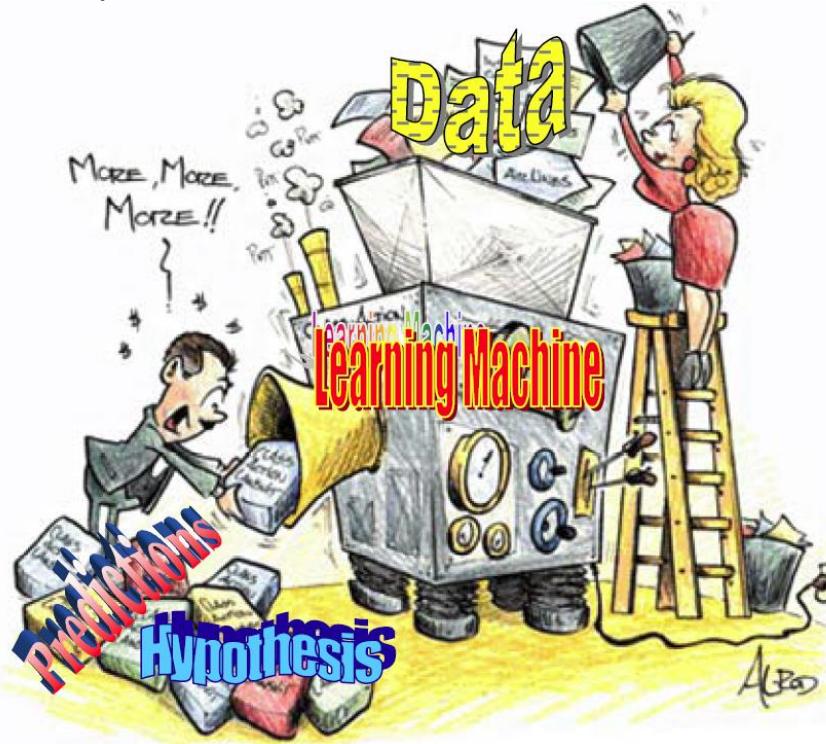
Harold R. and Mary Anne Nash Early Career Professor
H. Milton Stewart School of Industrial and Systems
Engineering

Introduction



What is machine learning (ML)

- Study of algorithms that improve their performance at some task with experience



Common to industrial scale problems



13 million wikipedia pages



800 million users



6 billion photos



340 million tweets per day

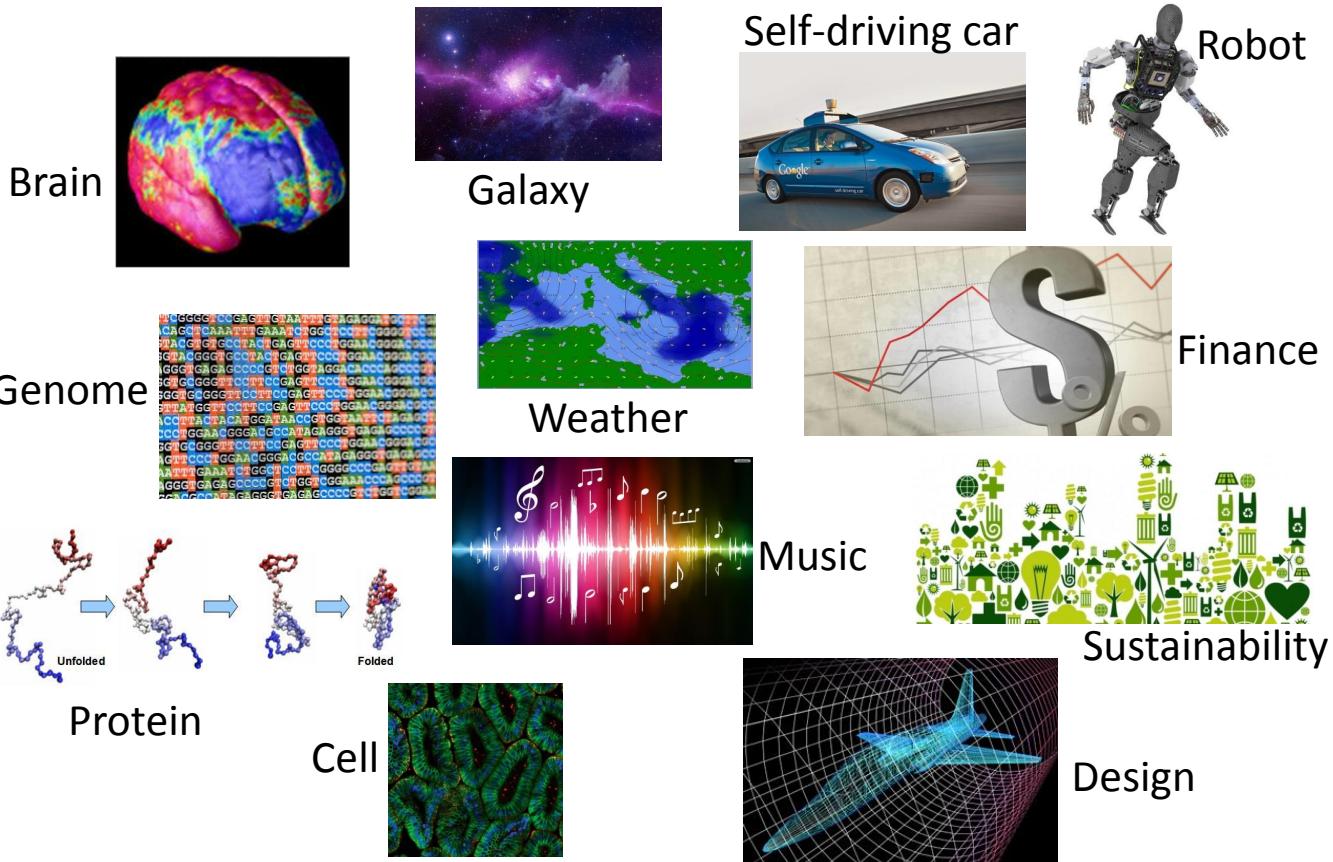


24 hours video uploaded per minutes

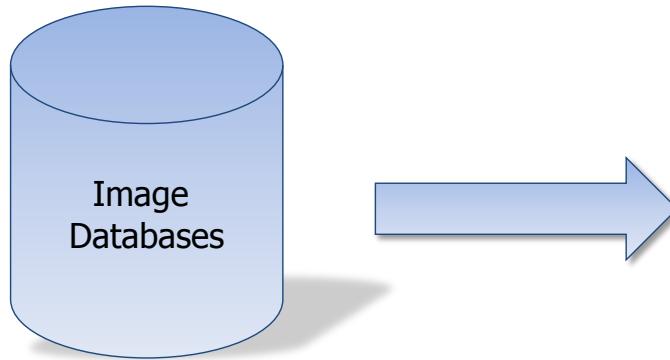


> 1 trillion webpages

Increasingly relevant to science problems



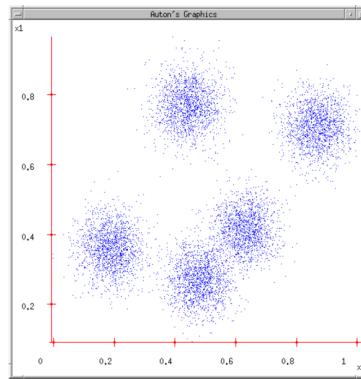
Organizing Images



What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?

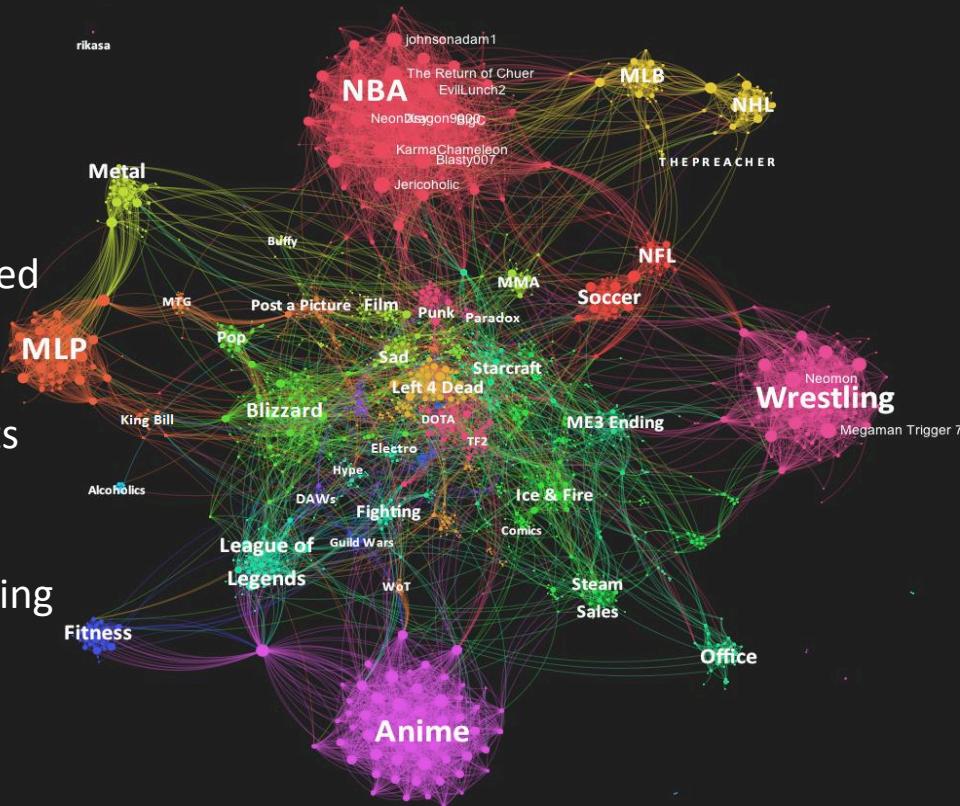


Find community in social networks

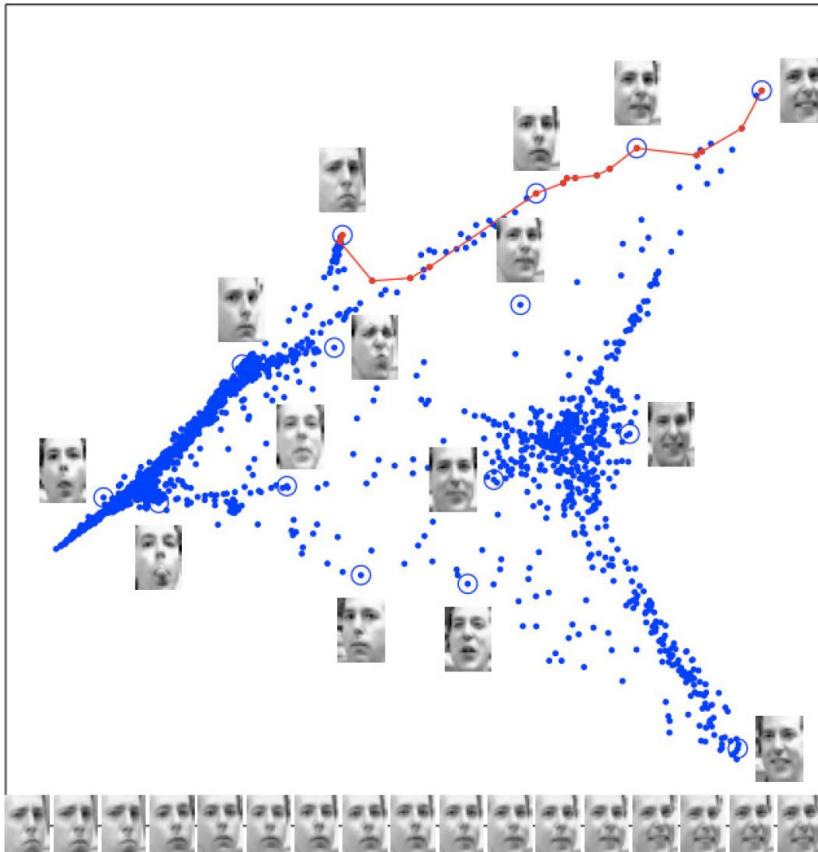
What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?



Visualize Image Relations



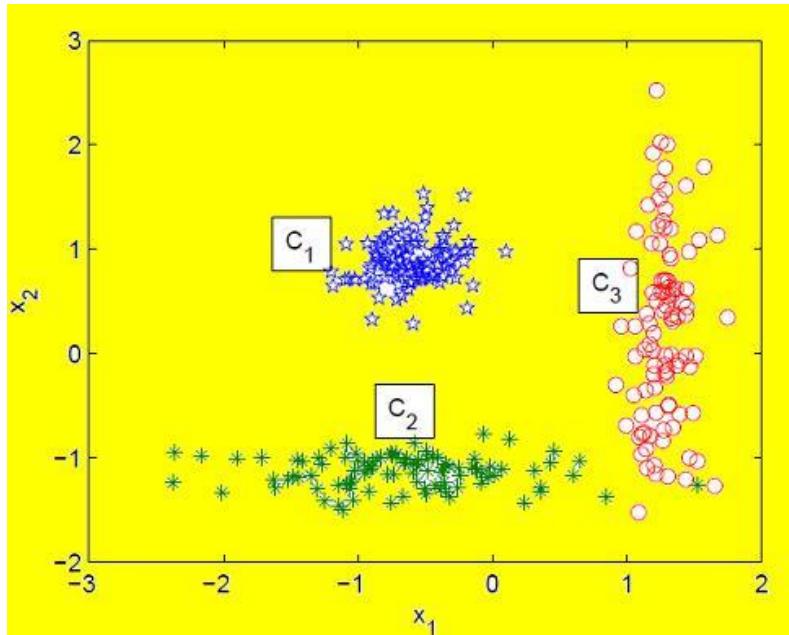
Each image has thousands or millions of pixels.

What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?

Feature selection

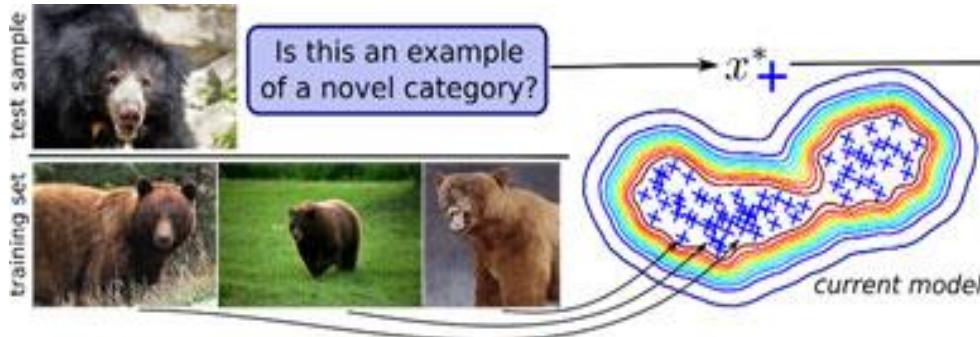


What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?

Novelty/abnormality detection



Find
abnormal
object



What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?

Image classification



mite	container ship	motor scooter	leopard
mite	container ship	motor scooter	leopard
black widow	lifeboat	go-kart	jaguar
cockroach	amphibian	moped	cheetah
tick	fireboat	bumper car	snow leopard
starfish	drilling platform	golfcart	Egyptian cat



grille	mushroom	cherry	Madagascar cat
convertible	agaric	dalmatian	squirrel monkey
grille	mushroom	grape	spider monkey
pickup	jelly fungus	elderberry	titi
beach wagon	gill fungus	ffordshire bullterrier	indri
fire engine	dead-man's-fingers	currant	howler monkey

What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?

Face Detection

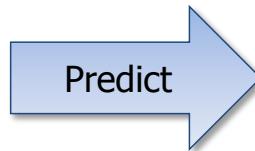


What are the desired outcomes?

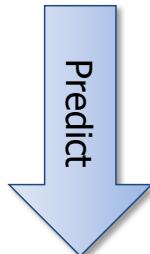
What are the inputs (data)?

What are the learning paradigms?

Weather Prediction



Numeric values:
40 F
Wind: NE at 14 km/h
Humidity: 83%



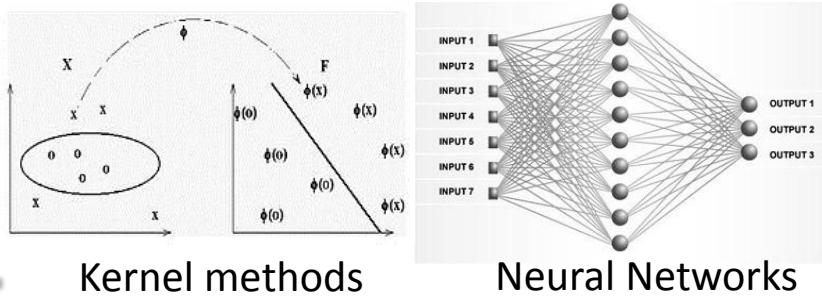
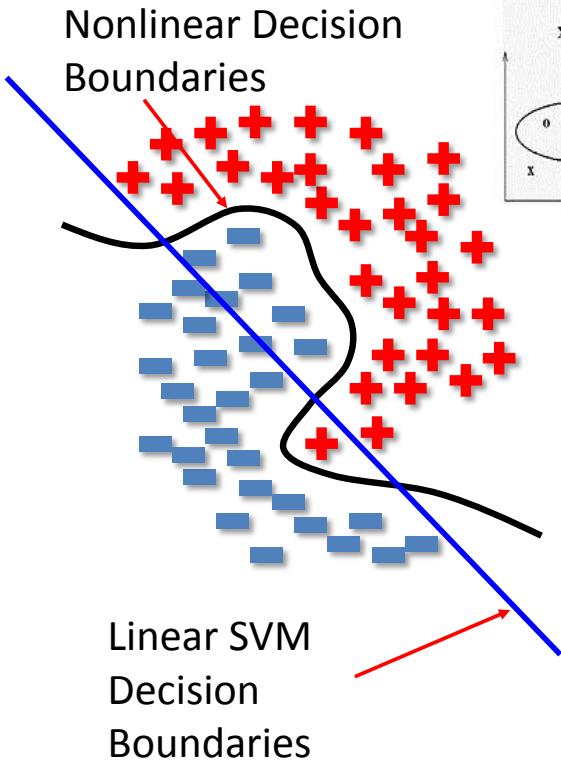
What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?



Nonlinear classifier

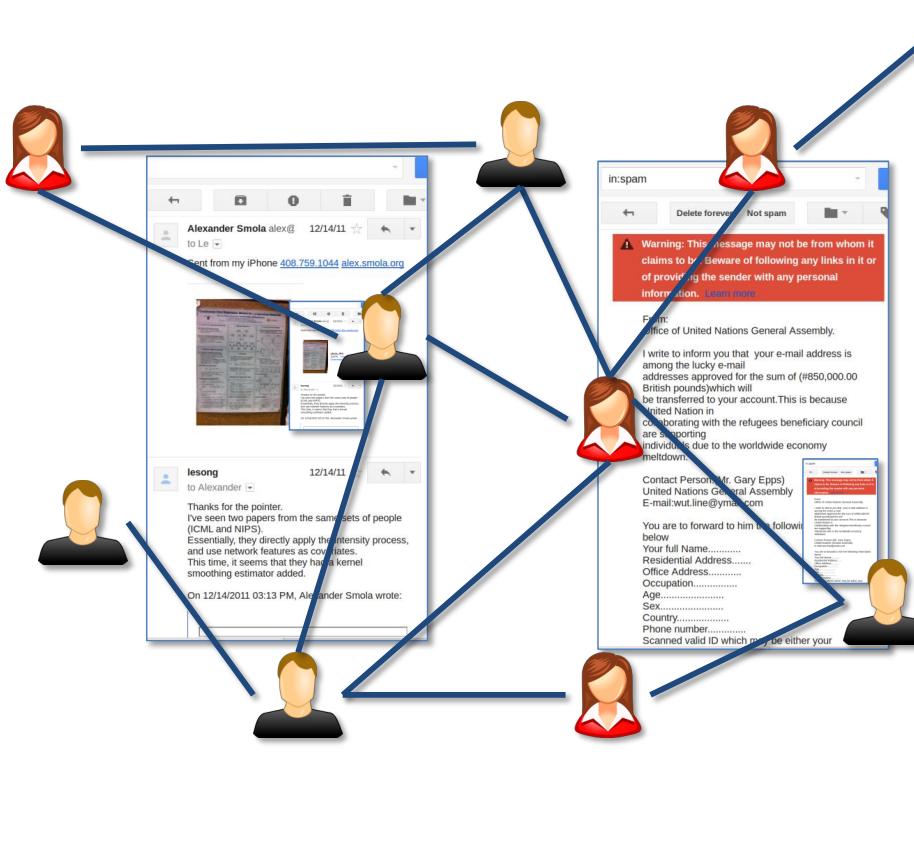


What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?

Spam Filtering



What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?

Handwritten digit recognition/text annotation

Inter-character dependency

*The unexpected
variables
Embarrass*

What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?

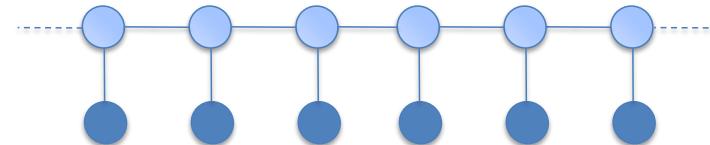
Inter-word dependency

Aoccdrnig to a sudty at Cmabrigde
Uinervtisy, it deosn't mttaer in waht
oredr the ltteers in a wrod are, the
olny iprmoetnt tihng is taht the frist
and lsat ltteer be at the rghit pclae.
The rset can be a ttoal mses and you
can stil raed it wouthit a porblm.
Tihs is bcuseae the huamn mnid
deos not raed ervey lteter by istlef,
but the wrod as a wlohe.

Speech recognition

Models

Hidden Markov Models



Text

“Machine Learning is the preferred method for speech recognition ...”

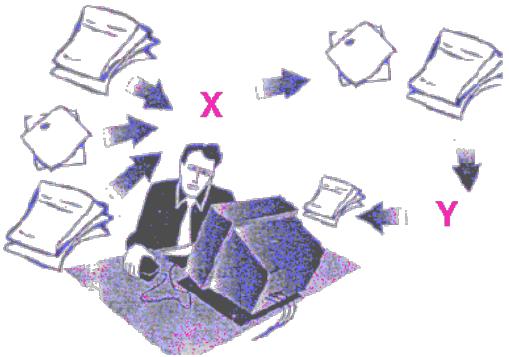


Audio signals



Organizing documents

- Reading, digesting, and categorizing a vast text database is too much for human!



- We want:

“Arts”	“Budgets”	“Children”	“Education”
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give \$1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services,” Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center’s share will be \$200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive \$400,000 each. The Juilliard School, where music and the performing arts are taught, will get \$260,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual \$100,000 donation, too.

What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?

Product Recommendation

The screenshot shows the product page for 'Pattern Recognition and Machine Learning (Information Science and Statistics) [Hardcover]' by Christopher M. Bishop. Key features include:

- Image:** Cover art for the book.
- Call-to-action:** 'Click to LOOK INSIDE!'
- Ratings:** ★★★★☆ (60 customer reviews), 74 likes.
- Price:** \$67.98 (List Price: \$94.95, You Save: \$26.97 (26%)). Includes Super Saver Shipping.
- Offer:** Special Offers Available.
- In Stock:** Ships from and sold by Amazon.com. Gift-wrap available.
- Delivery:** Want it delivered Monday, January 9? Order it in the next 21 hours and 41 minutes, and choose One-Day Shipping at checkout.
- Inventory:** 42 new from \$67.98, 23 used from \$69.97.
- Shipping:** FREE Two-Day Shipping for Students.
- Formats:** Hardcover (\$67.98, \$67.98, \$69.97).
- Book Trade-In:** Sell Back Your Copy for \$56.97.
- More Buying Choices:** 65 used & new from \$67.98.
- Customer Options:** Add to Cart, Sign In for 1-Click, Add to Wish List, Sell Back Your Copy for a \$56.97 Gift Card, Trade In, Share.
- Frequently Bought Together:** Books related to machine learning.
- Price For All Three:** \$191.05.
- Customers Who Bought This Item Also Bought:** Books like 'Machine Learning: An Algorithmic Perspective', 'Probabilistic Graphical Models', 'Data Mining', 'The Elements of Statistical Learning', and 'Pattern Classification'.

What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?

Robot Control

- Now cars can find their own ways!



What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?



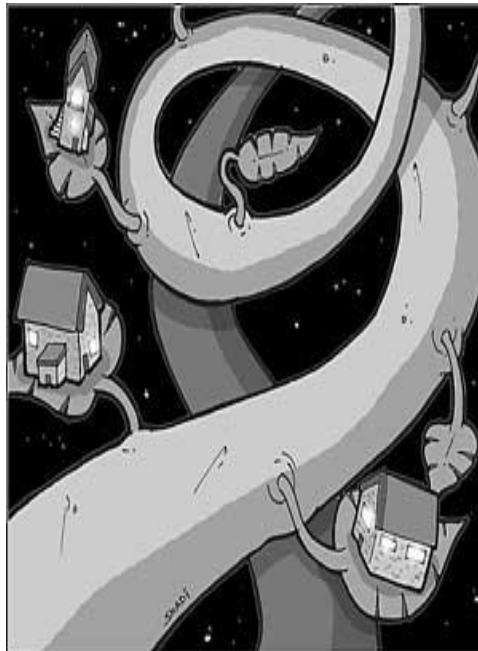
Basics/Prerequisites

- Probabilities
 - Distributions, densities, marginalization, conditioning
- Statistics
 - Mean, variance, maximum likelihood estimation
- Linear algebra
 - Vector, matrix, multiplication, inversion, eigen-decomposition
- Algorithms and Programming
 - Matlab, Basic data structures, computational complexity
- Convex optimization
 - Basics will be covered during lecture

Machine learning for apartment hunting

- Suppose you are to move to Atlanta
- And you want to find the **most reasonably priced** apartment satisfying your **needs**:

square-ft., # of bedroom, distance to campus ...



Living area (ft ²)	# bedroom	Rent (\$)
230	1	600
506	2	1000
433	2	1100
109	1	500
...		
150	1	?
270	1.5	?

Linear Regression Model

- Assume y is a linear function of x (features) plus noise ϵ

$$y = \theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n + \epsilon$$

where ϵ is an error model as Gaussian $N(0, \sigma^2)$

Probability

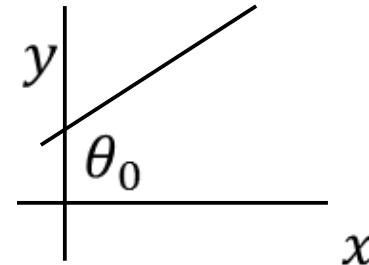
- Let $\theta = (\theta_0, \theta_1, \dots, \theta_n)^\top$, and augment data by one dimension

Linear algebra

$$x \leftarrow (1, x)^\top$$

Then $y = \theta^\top x + \epsilon$

Linear algebra



Least mean square method

- Given m data points, find θ that minimizes the mean square error

$$\hat{\theta} = \operatorname{argmin}_{\theta} L(\theta) = \frac{1}{m} \sum_{i=1}^m (y^i - \theta^\top x^i)^2$$

Optimization

Statistics

- Set gradient to 0 and find parameter

Optimization

Linear algebra

$$\frac{\partial L(\theta)}{\partial \theta} = -\frac{2}{m} \sum_{i=1}^m (y^i - \theta^\top x^i) x^i = 0$$

$$\Leftrightarrow -\frac{2}{m} \sum_{i=1}^m y^i x^i + \frac{2}{m} \sum_{i=1}^m x^i x^{i\top} \theta = 0$$

Statistics

Statistics

Matrix version of the gradient

- Define $X = (x^1, x^2, \dots, x^m), y = (y^1, y^2, \dots, y^m)^\top$, gradient becomes

Linear algebra →
$$\frac{\partial L(\theta)}{\partial \theta} = -\frac{2}{m} Xy + \frac{2}{m} XX^\top \theta$$

Linear algebra →
$$\Rightarrow \hat{\theta} = (XX^\top)^{-1}Xy$$

Algorithms
Programming

- Matrix inversion in $\hat{\theta} = (XX^\top)^{-1}Xy$ **expensive** to compute

- Gradient descent

$$\hat{\theta}^{t+1} \leftarrow \hat{\theta}^t + \frac{\alpha}{m} \sum_i^m (y^i - \hat{\theta}^{t\top} x^i) x^i$$

Optimization

Probabilistic Interpretation of LMS

- Assume y is a linear in x plus noise ϵ

$$y = \theta^\top x + \epsilon$$

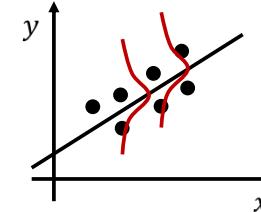
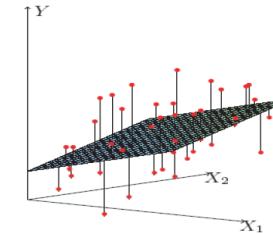
- Assume ϵ follows a Gaussian $N(0, \sigma^2)$

$$p(y^i | x^i; \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^i - \theta^\top x^i)^2}{2\sigma^2}\right)$$

- By independence assumption, likelihood is $L(\theta)$

$$= \prod_i^m p(y^i | x^i; \theta) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^m \exp\left(-\frac{\sum_i^m (y^i - \theta^\top x^i)^2}{2\sigma^2}\right)$$

Probability



Probabilistic Interpretation of LMS, cont.

- Hence the log-likelihood is:

$$\log L(\theta) = m \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_i^m (y^i - \theta^\top x^i)^2$$

- LMS is equivalent to MLE of θ !

$$LMS: \frac{1}{m} \sum_i^m (y^i - \theta^\top x^i)^2$$

Statistics

- How to make it work in real data?

Algorithms
Programming

