Northeastern

CS 7140:
ADVANCED MACHINE LEARNING

Some Recent HeadLines

NEWS · 11 JULY 2019

No limit: AI poker bot is first to beat professionals at multiplayer game

Triumph over five hur
to solving complicated

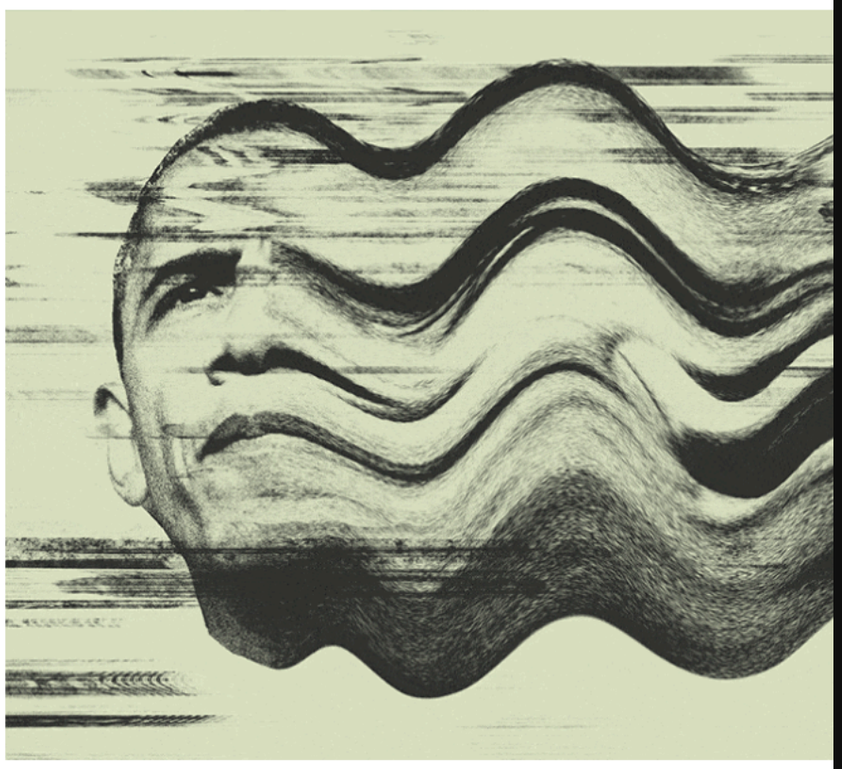
Douglas Heaven

Some Recent HeadLines

Technology

You thought fake news was bad? Deep fakes are where truth goes to die

Technology can make it look as if anyone has said anything. Is it the next wave of (mis)information

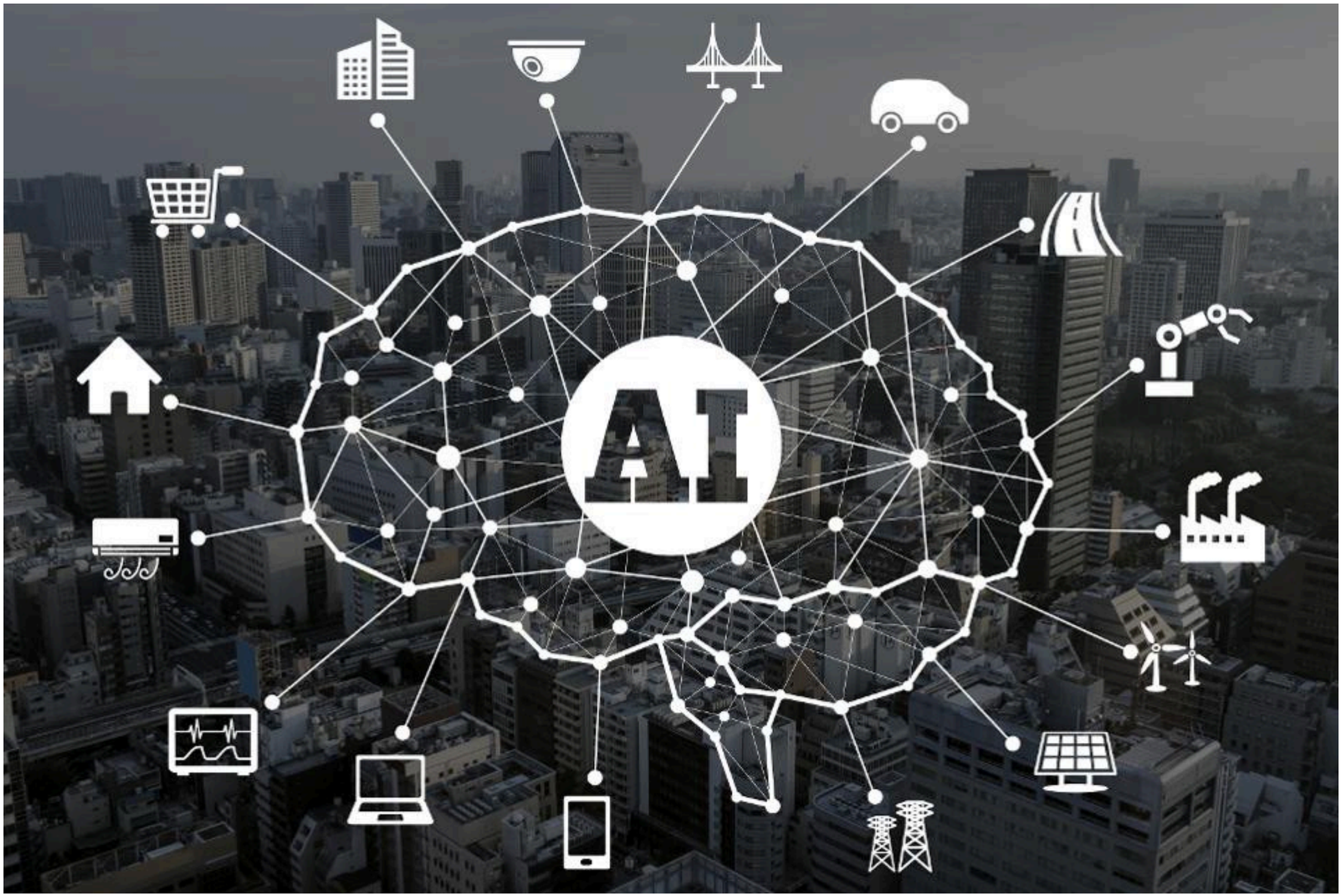


▲ 'When nothing is true then the dishonest person will thrive by saying what

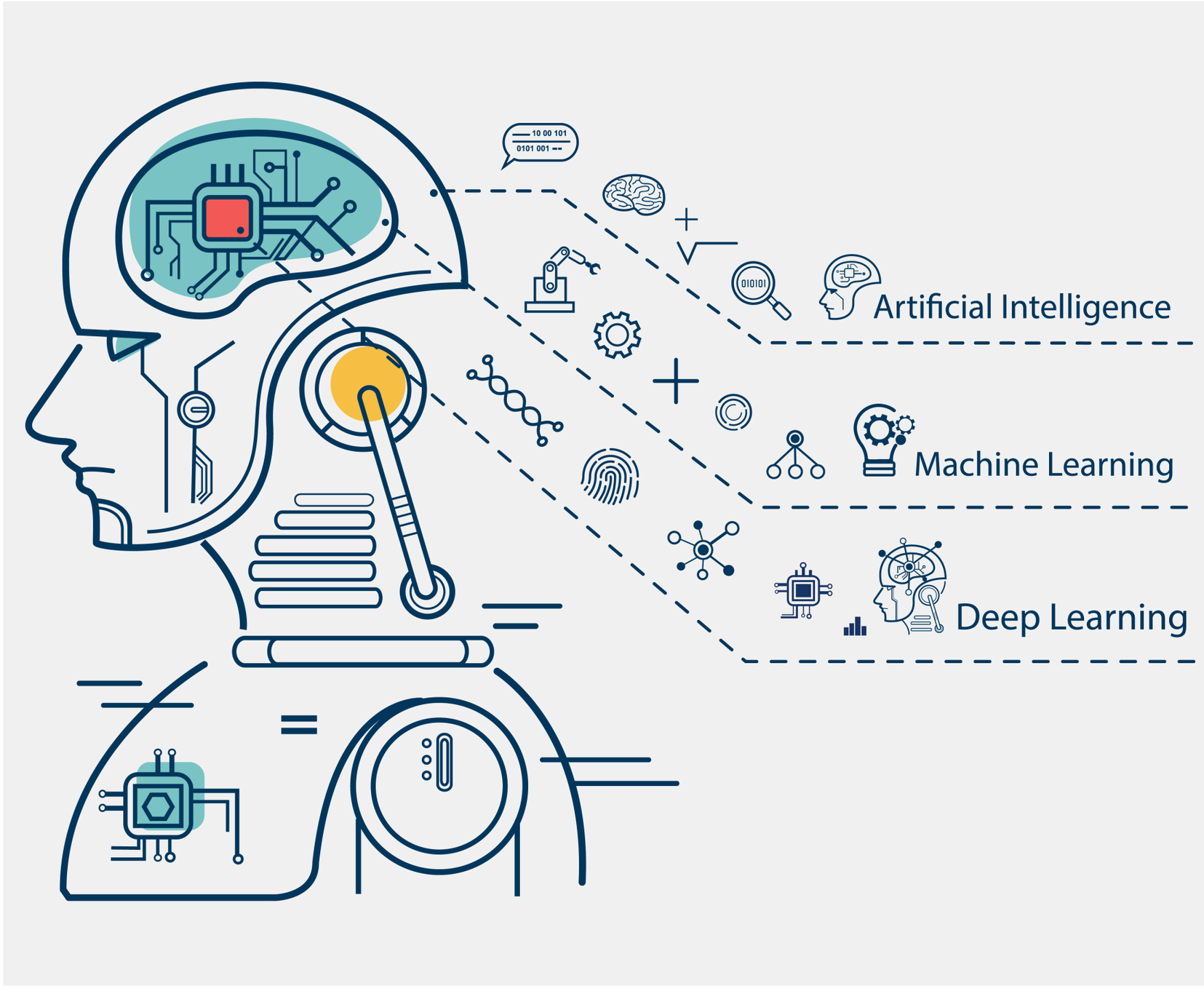
World's first AI presenter unveiled in China – video

China's Xinhua state news agency has introduced the newest members of its newsroom: AI anchors who will report 'tirelessly' all day, every day, from anywhere in the country

Artificial Intelligence



Machine Learning



ADMINISTRATIVE

Staff

Instructor: Rose Yu

roseyu@northeastern.edu

roseyu.com @yuqirose

TA: Clara De Paolis

clara@ccs.neu.edu

Students

Course Info

- 11:45 am - 1:25 pm Monday, Thursday
  - 11:45 - 12:30
  - 12:40 - 1:25
- Location: Forsyth Building 202
- Office Hours: by appointment

Why take this course?

- Learn the backbone of Machine Learning: Probabilistic Graphical Model (PGM)

- Learn the state of the art of Deep Learning

- Understand the connections between PGM and Deep Learning

- Hands on experience in conducting Machine Learning research projects

Course Materials

- Website: https://sites.google.com/view/cs7140spring2020/home
  - Syllabus
  - Reading materials
- Piazza: piazza.com/northeastern/spring2020/cs7140
  - Announcements
  - Discussion
  - When in doubt, post on Piazza

Course Materials

- Required Reading
  - [1] Information Theory, Inference, and Learning Algorithms
  - [2] Probabilistic Graphical Models: Principles and Techniques
  - [3] Deep Learning Book
- Optional Reading
  - [4] Graphical Models, Exponential Families and Variational Inference
  - [5] The Elements of Statistical Learning
  - [6] Graphical Model Course

Grades Breakdown

- 40% Homework (10 % x 4)
- 40% Project
  - Proposal 5%
  - Milestone 10%
  - Final Report (Presentation) 15%
- 15% Paper Discussion
- 5% Lecture Scribe

Research Project

- Complete a research project in groups (2-3)
  - Based on the papers you read
  - Meet with me to discuss and get approval
  - Conduct experiments and write report
  - Details http://roseyu.com/CS7140/final_project_suggested_topics.pdf

Course Policy

- Late policy
  - 1 late waiver (up to one week)
  - 1 point deduction per day
  - Email TA to use the waiver
- Plagiarism
  - Do NOT do it!!!

PROBABILISTIC GRAPHICAL MODEL

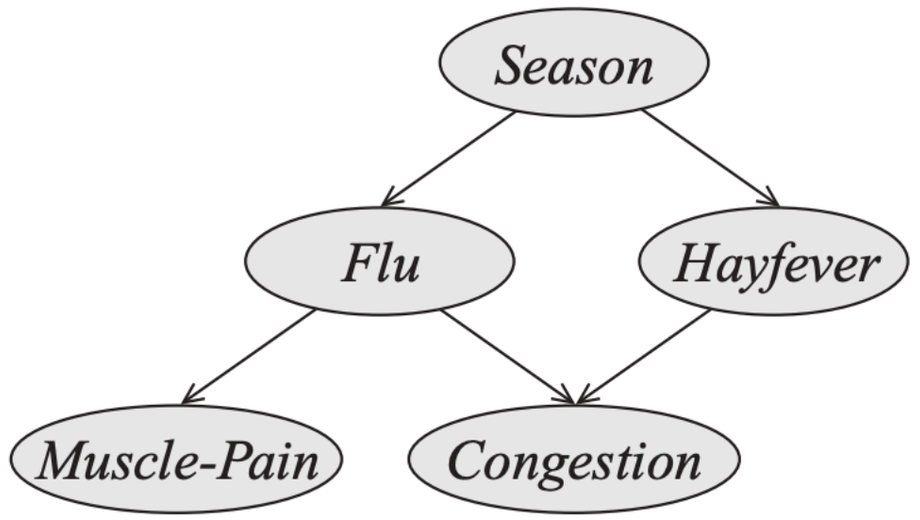
Reason with Uncertainty

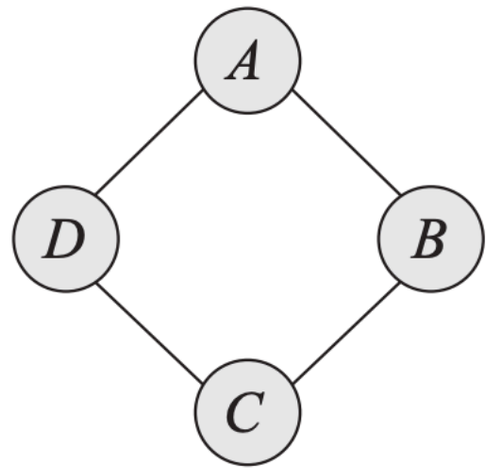
- Complex systems involve a significant amount of uncertainty

- **Probabilistic Graphical Model**: Graph representation of complex joint distributions

Probability

Graph Representation





Independencies

$$(F \perp H \mid S)$$
$$(C \perp S \mid F, H)$$
$$(M \perp H, C \mid F)$$
$$(M \perp C \mid F)$$

$$(A \perp C \mid B, D)$$
$$(B \perp D \mid A, C)$$

Factorization

$$P(S, F, H, C, M) = P(S)P(F \mid S)$$
$$P(H \mid S)P(C \mid F, H)P(M \mid F)$$

$$P(A, B, C, D) = \frac{1}{Z}\phi_1(A, B)$$
$$\phi_2(B, C)\phi_3(C, D)\phi_4(A, D)$$

A Brief History

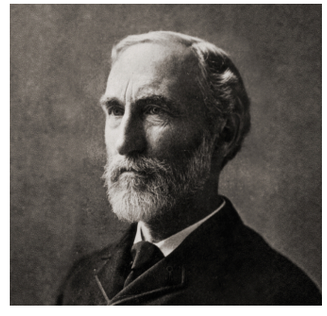
- 1763 Bayes' Theorem

  Reverend Thomas Bayes invented Bayes' theorem

Reverend Thomas Bayes

- 1902 Undirected Graph

  Statistical physicist Josiah Willard Gibbs use undirected graphs to represent interacting particles



Josiah Willard Gibbs

- 1970s Expert Systems

Judea Pearl

- 1988 Bayesian Network

  computer scientist and philosopher Judea Pearl pioneered Bayesian networks to represent probabilistic relationships

- 2002 PGM book

  computer scientist Daphne Koller wrote the text book on probabilistic graphical models

Daphne Koller

PROBABILITY WARMUP

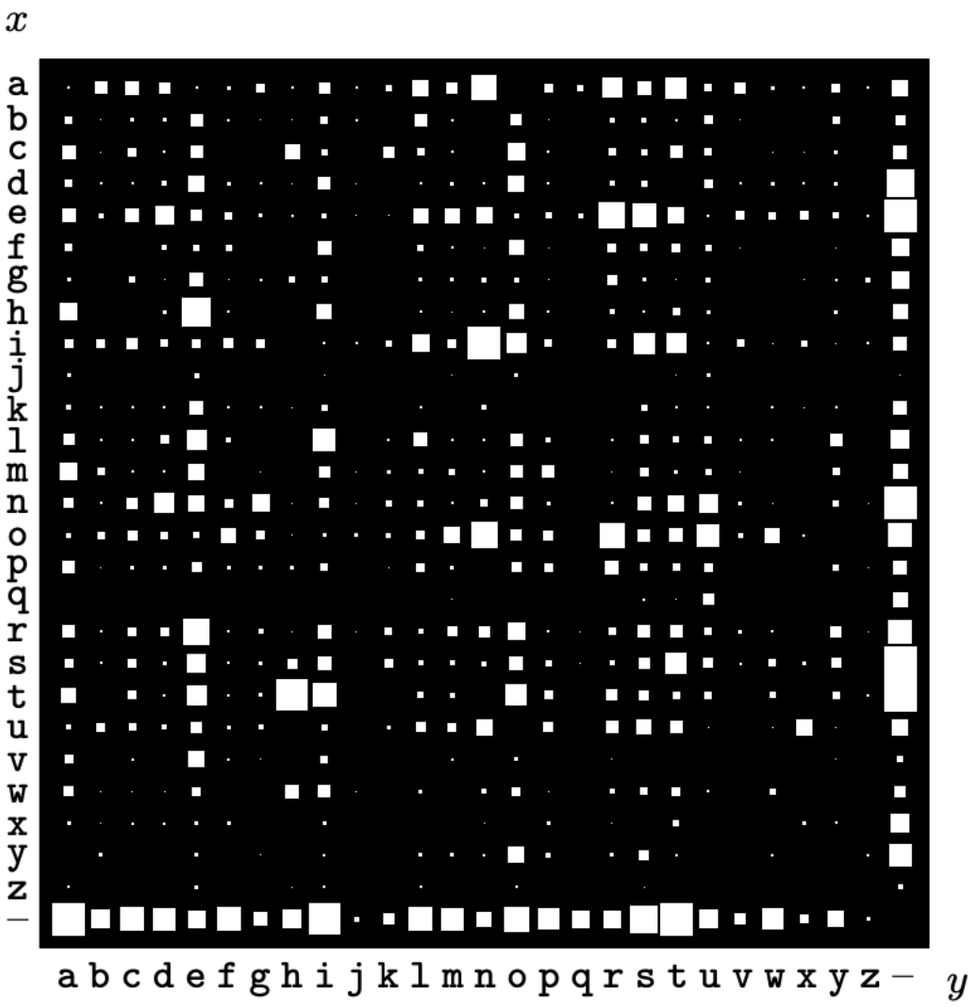
Probability

i	a_i	p_i
1	a	0.0575
2	b	0.0128
3	c	0.0263
4	d	0.0285
5	e	0.0913
6	f	0.0173
7	g	0.0133
8	h	0.0313
9	i	0.0599
10	j	0.0006
11	k	0.0084
12	l	0.0335
13	m	0.0235
14	n	0.0596
15	o	0.0689
16	p	0.0192
17	q	0.0008
18	r	0.0508
19	s	0.0567
20	t	0.0706
21	u	0.0334
22	v	0.0069
23	w	0.0119
24	x	0.0073
25	y	0.0164
26	z	0.0007
27	–	0.1928

- Random Variable X
- Outcome $A_X = \{a_1, a_2, \cdots, a_I\}$
- Probability

$$T \subset A_X \qquad p(T) = \sum_{a_i \in T} P(x = a_i)$$

Probability



- Joint Probability $X \quad Y$

  $p(x, y)$

- Marginal Probability

$$p(x = a_i) = \sum_{y \in A_Y} p(x = a_i, y)$$

- Conditional Probability

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

Probability Theory

- product rule

$$p(x, y \mid H) = p(x \mid y, H)p(y \mid H) = p(y \mid x, H)p(x \mid H)$$

- sum rule

$$p(x \mid H) = \sum_y p(x, y \mid H) = \sum_y p(x \mid y, H)p(y \mid H)$$

- Bayes' theorem

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

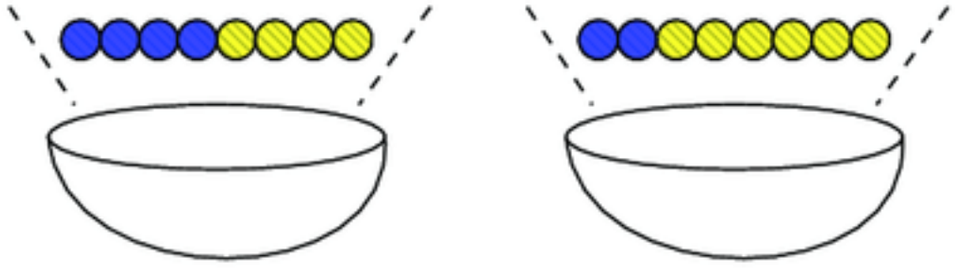
- independence

$$p(x, y) = p(x)p(y)$$

- conditional independence

$$p(x, y \mid z) = p(x \mid z)p(y \mid z)$$

Probability Theory



- What is the probability distribution of the number of times a blue ball is draw?

- forward probability $p(x|\theta)$ likelihood

- After 10 draws and 3 blues have been drawn, what is the probability that the urn Alice is using is the left urn from Bob's view?

- inverse probability $p(\theta|x)$ posterior

Probability Theory

Alice, a scientist, conducted 12 trials and obtains 3 successes and 9 failures. Then she left the lab. Bob, another scientist, kept trying until he obtained 3 successes. Assume success probability is θ .

- What is the probability for Alice that out of 12 trials 3 were successes?

$$L(\theta \mid X = 3) = \binom{12}{3} \theta^3 (1-\theta)^9$$

- What is the probability for Bob needing to conduct 12 experiments?

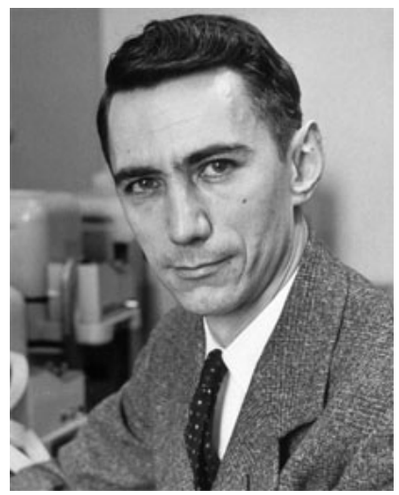
$$L(\theta \mid Y = 12) = \binom{11}{2} \theta^3 (1-\theta)^9$$

The likelihood principle: given a generative model for data d given parameters θ, $p(d \mid \theta)$, and having observed a particular outcome d_1, all inferences and predictions should depend only on the function $p(d_1 \mid \theta)$.

Entropy

- Shannon Information

$$h(x) = \log_2 \frac{1}{P(x)}$$

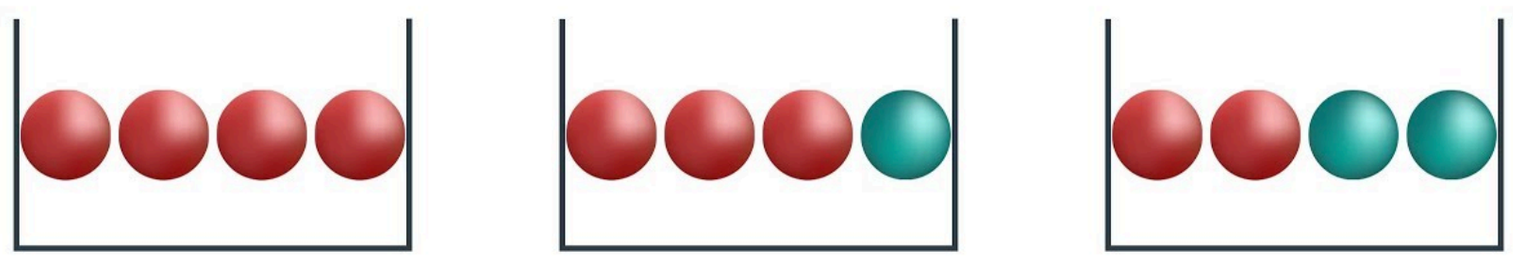


Claude Shannon

- Base of log is unimportant — will only change the units
- Information as amount of "surprise"
- Concerned with abstract possibilities, not their meaning

Entropy

- Entropy $$H(X) = \sum_{x \in A_X} P(x) \log \frac{1}{P(x)}$$



low medium high

  - $H(X) \geq 0$ with equality iff $p_i = 1$ for one i.
  - entropy is maximized if p is uniform

- Joint Entropy $$H(X, Y) = \sum_{x \in A_X, y \in A_Y} P(x, y) \log \frac{1}{P(x, y)}$$
  - Entropy is additive for independent random variables
  - $H(X, Y) = H(X) + H(Y) \iff p(x, y) = p(x)p(y)$

Entropy

- Decomposability

$$x = \{0,1,2\} \qquad p(x) = (\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$$

$$H(x) = \frac{1}{2}\log 2 + \frac{1}{4}\log 4 + \frac{1}{4}\log 4 = 1.5$$

$$H(x) = \frac{1}{2}\log 2 + \frac{1}{2}\log 2 + \frac{1}{2}(\frac{1}{2}\log 2 + \frac{1}{2}\log 2) = 1.5$$

$$H(p) = H(p_1, 1-p_1) + (1-p_1)H(\frac{p_2}{1-p_1}, \frac{p_2}{1-p_1}, \cdots, \frac{p_I}{1-p_1})$$

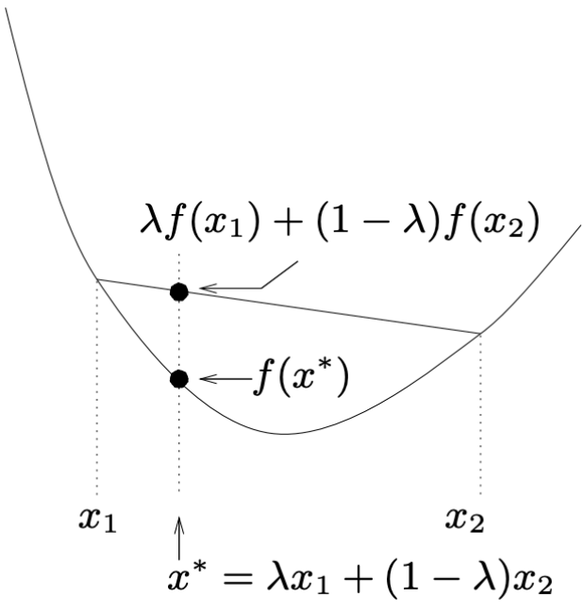
- Relative entropy/Kullback-Leibler Divergence

$$D_{KL}(P||Q) = \sum_{x \in A_X} P(x)\log\frac{P(x)}{Q(x)}$$

- Gibbs Inequality

$$D_{KL}(P||Q) \geq 0$$

Jensen's Inequality



- A function is convex

$$\mathbb{E}[f(x)] \geq f(\mathbb{E}[x])$$