DSC250: Advanced Data Mining

Machine Learning Basics

Zhiting Hu

Lecture 4, Jan 16, 2025

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

Outline

- Probability
  - Bayes' rule
  - Exponential family
  - Probabilistic graphical models
  - Entropy, KL divergence, cross entropy

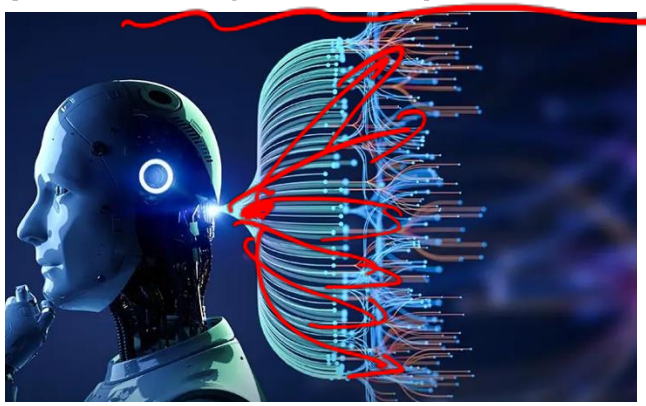
- Functional derivatives

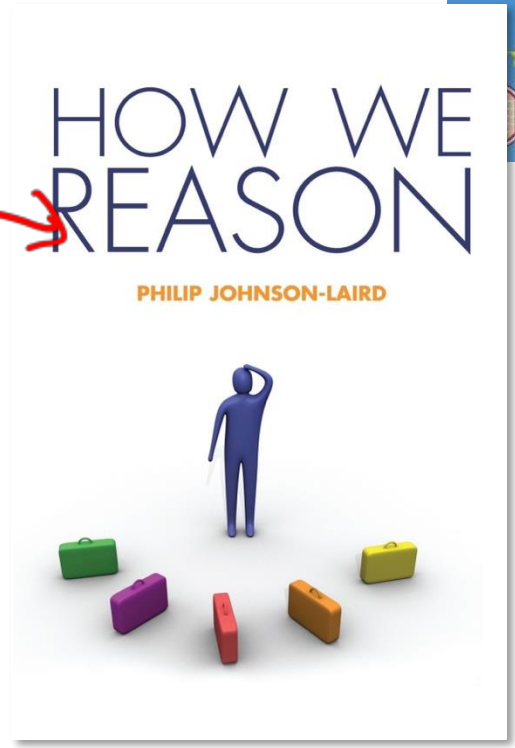
- Practice: MLE vs Maximum entropy

Probability

Why Probability?

- The world is a very uncertain place
  - "What will the weather be like today?"
  - "Will I like this movie?"
- We often can't prove something is true, but we can still ask how likely different outcomes are or ask for the most likely explanations
  - Indeed, this is how humans reason
  - We reason by "thinking about possibilities"



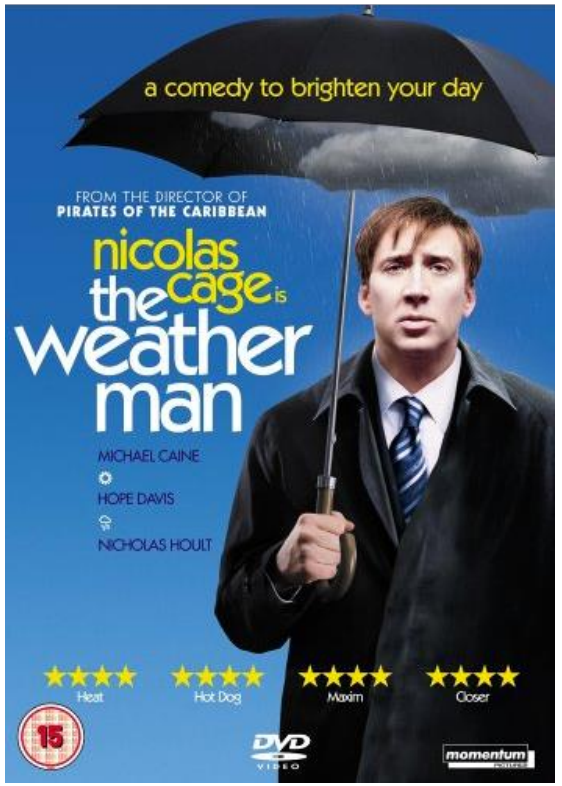


Mental model theory

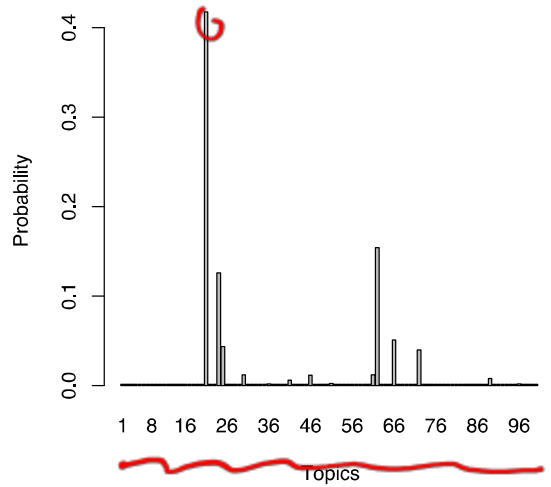
[CS60020, Bhattacharya; CSC2515, Wang]

Why Probability?

- The world is a very uncertain place
  - "What will the weather be like today?"
  - "Will I like this movie?"
- We often can't prove something is true, but we can still ask how likely different outcomes are or ask for the most likely explanations
  - Indeed, this is how humans reason
- Predictions need to have associated confidence
  - Confidence -> probability
- Not all machine learning models are probabilistic
  - … but most of them have probabilistic interpretations



Example: topic modeling



"Genetics"	"Evolution"	"Disease"	"Computers"
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

For documents in a large collection of text, model $p(\text{Word}|\text{Topic})$, $p(\text{Topic})$

Figure from (Blei, 2011), shows topics and top words learned automatically from reading 17,000 Science articles

Example: image segmentation

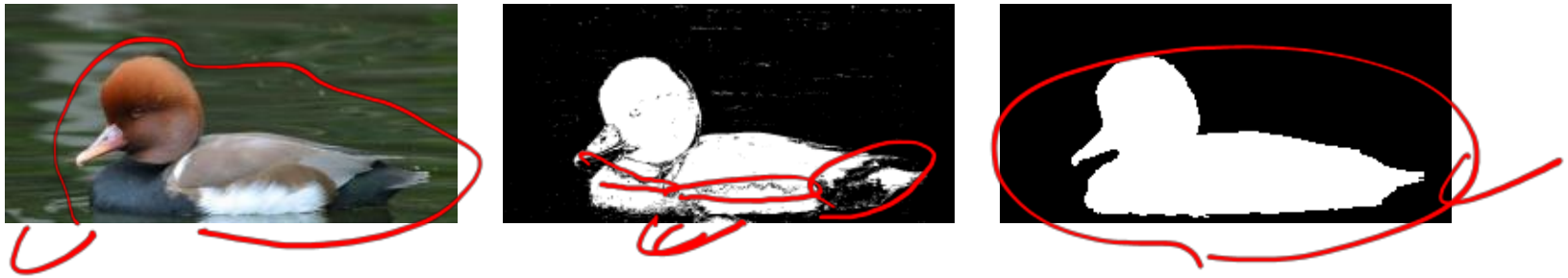
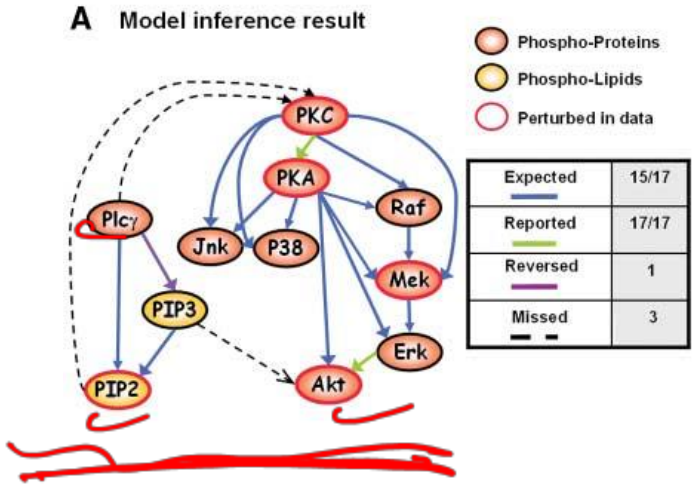


Figure (Nowozin and Lampert, 2012) shows image segmentation problem, original image on left, where goal is to separate foreground from background

Middle figure shows a segmentation where each pixel is individually classified as belonging to foreground or background

Right figure shows a segmentation where the segmentation is inferred from a probability model over all pixels jointly (encoding probability that neighboring pixels tend to belong to the same group)

Example: modeling protein networks



In cellular modeling, can we automatically determine how the presence or absence of some proteins affects other proteins?

Figure from (Sachs et al., 2005) shows automatically inferred protein probability network, which captured most of the known interactions using data-driven methods (far less manual effort than previous methods)

Notations

- A random variable x represents outcomes or states of the world.
  - We write $p(x_0)$ to mean Probability($x = x_0$)
- Sample space: the space of all possible outcomes (may be discrete, continuous, or mixed)
- $p(x)$ is the probability mass (density) function
  - Assigns a number to each point in sample space
  - Non-negative, sums (integrates) to 1
  - Intuitively: how often does x occur, how much do we believe in x.

$\sum p(x) = 1$ $\int p(x) = 1$

Notations

- Joint distribution $p(\boldsymbol{x}, \boldsymbol{y})$
- Conditional distribution $p(\boldsymbol{y}|\boldsymbol{x})$
  - $p(\boldsymbol{y}|\boldsymbol{x}) = \frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{x})}$
- Expectation:

$$\mathbb{E}[f(\boldsymbol{x})] = \sum_{\boldsymbol{x}} f(\boldsymbol{x})\, p(\boldsymbol{x})$$

or

$$\mathbb{E}[f(\boldsymbol{x})] = \int_{\boldsymbol{x}} f(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x}$$

Rules of Probability

- Sum rule

$$p(x) = \sum_{y} p(x, y) \qquad \text{(Marginalize out } y\text{)}$$

$$p(x_1) = \sum_{x_2} \sum_{x_3} \dots \sum_{x_N} p(x_1, x_2, \dots, x_N)$$

- Product/chain rule

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

$$p(x_1, \dots, x_N) = p(x_1)\, p(x_2 \mid x_1) \dots p(x_N \mid x_1, \dots, x_{N-1})$$

Bayes' Rule

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

$$\frac{P(x,y)}{P(x)}$$

- This gives us a way of "reversing" conditional probabilities
- We call $p(\boldsymbol{y})$ the "prior", and $p(\boldsymbol{y}|\boldsymbol{x})$ the "posterior"
- Ex: Bayes' Rule in machine learning:
  - $\mathcal{D}$: data (evidence)
  - $\boldsymbol{\theta}$: unknown quantities, such as model parameters, predictions

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})}$$

Likelihood: How likely is the observed data under the particular unknown quantities $\boldsymbol{\theta}$

Prior belief on the unknown quantities **before** you see data $\mathcal{D}$

Posterior belief on the unknown quantities you see data $\mathcal{D}$

Independence

- Two random variables are said to be **independent** iff their joint distribution factors

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{x})p(\boldsymbol{y})$$

- Two random variables are **conditionally independent** given a third if they are independent after conditioning on the third

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

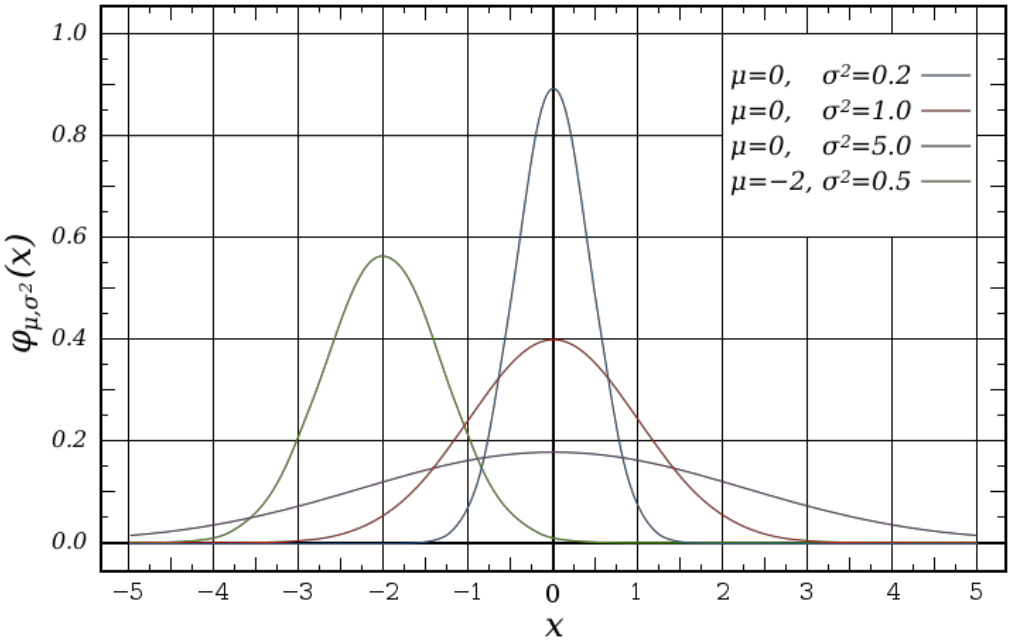
Some common distributions - Gaussian distribution

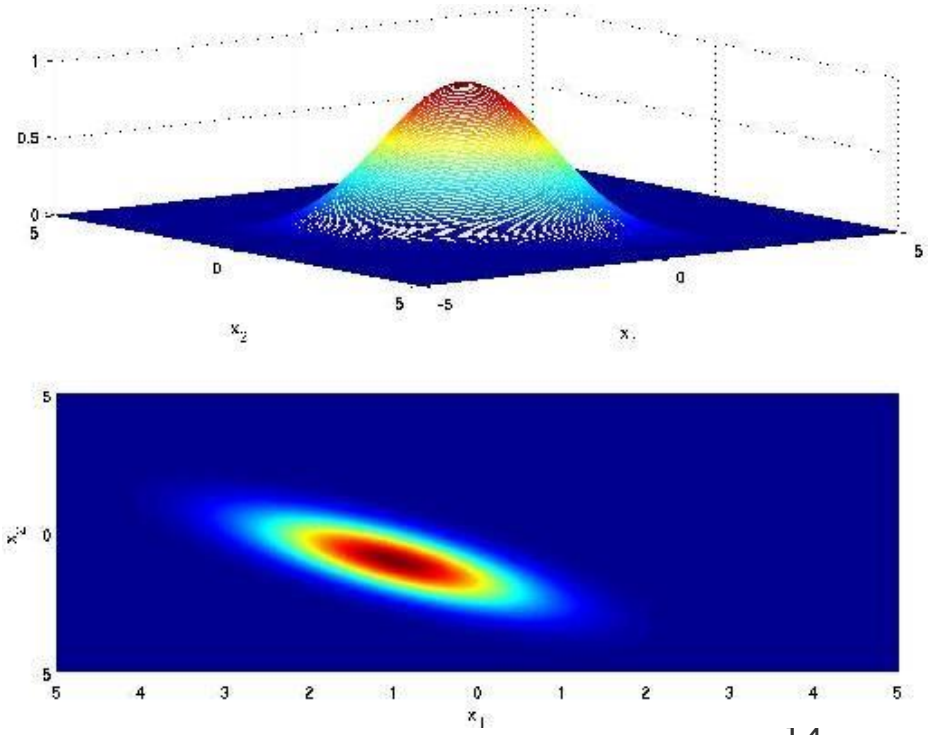
- Gaussian distribution

$$P(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

(Multivariate)

$$P(x \mid \mu, \Sigma) = |2\pi\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}$$





[CSC2515, Wang]

Some common distributions - Multinomial distribution

- Multinomial distribution
  - Discrete random variable $\boldsymbol{x}$ that takes one of M values $\{1, \dots, M\}$
  - $p(\boldsymbol{x} = i) = \pi_i, \quad \sum_i \pi_i = 1$
  - Out of n independent trials, let k_i be the number of times $\boldsymbol{x} = i$ was observed
  - The probability of observing a vector of occurrences $\boldsymbol{k} = [k_1, \dots, k_M]$ is given by the *multinomial distribution* parametrized by $\boldsymbol{\pi}$

$$\mathrm{p}(\mathbf{k}|\boldsymbol{\pi}, \mathrm{n}) = \mathrm{p}(\mathrm{k}_1, \dots, \mathrm{k}_\mathrm{m}|\pi_1, \dots, \pi_\mathrm{m}, \mathrm{n}) = \frac{\mathrm{n}!}{\mathrm{k}_1!\mathrm{k}_2! \dots \mathrm{k}_\mathrm{m}!} \prod_{\mathrm{i}=1} \pi_\mathrm{i}^{\mathrm{k}_\mathrm{i}}$$

  - E.g., describing a text document by the frequency of occurrence of every distinct word
  - For $n = 1$, a.k.a. categorical distribution
    - $p(\boldsymbol{x} = i \mid \boldsymbol{\pi}) = \pi_i$
    - In $\boldsymbol{k} = [k_1, \dots, k_M]$: $k_i = 1$, and $k_j = 0$ for all $j \neq i$ → a.k.a., one-hot representation of i

Exponential family

- A distribution

$$p_\theta(\boldsymbol{x}) = h(\boldsymbol{x}) \exp\{\boldsymbol{\theta} \cdot T(\boldsymbol{x})\} / Z(\boldsymbol{\theta})$$

  is an exponential family distribution

  - $\boldsymbol{\theta} \in R^d$: natural (canonical) parameter
  - $T(\boldsymbol{x}) \in R^d$: sufficient statistics, features of data $\boldsymbol{x}$
  - $Z(\boldsymbol{\theta}) = \sum_{x,y} h(\boldsymbol{x}) \exp\{\boldsymbol{\theta} \cdot T(\boldsymbol{x})\}$: normalization factor

- Examples: Bernoulli, multinomial, Gaussian, Poisson, gamma,...

Example: Multivariate Gaussian Distribution

- For a continuous vector random variable $\boldsymbol{x} \in R^k$

$$p(x|\mu,\Sigma) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}$$

$$= \frac{1}{(2\pi)^{k/2}} \exp\left\{-\tfrac{1}{2}\mathrm{tr}\left(\Sigma^{-1}xx^T\right) + \mu^T\Sigma^{-1}x - \tfrac{1}{2}\mu^T\Sigma^{-1}\mu - \log|\Sigma|\right\}$$

Moment parameter

- Exponential family representation

$$\boldsymbol{\theta} = \left[\Sigma^{-1}\mu; -\frac{1}{2}\mathrm{vec}\left(\Sigma^{-1}\right)\right] = \left[\boldsymbol{\theta}_1, \mathrm{vec}\left(\boldsymbol{\theta}_2\right)\right], \ \boldsymbol{\theta}_1 = \Sigma^{-1}\mu \text{ and } \boldsymbol{\theta}_2^- = -\frac{1}{2}\Sigma^{-1}$$

$$T(\boldsymbol{x}) = \left[\boldsymbol{x}; \mathrm{vec}\left(\boldsymbol{x}\boldsymbol{x}^T\right)\right]$$

$$A(\boldsymbol{\theta}) = \frac{1}{2}\mu^T\Sigma^{-1}\mu + \log|\Sigma| = -\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\theta}_2\boldsymbol{\theta}_1\boldsymbol{\theta}_1^T\right) - \frac{1}{2}\log\left(-2\boldsymbol{\theta}_2\right)$$

$$h(x) = (2\pi)^{-k/2}$$

Probabilistic Graphical Models

Example

- Consider three binary-valued random variables

$$X_1, X_2, X_3 \qquad \mathrm{Val}(X_i) = \{0, 1\}$$

- Let outcome space Ω be the cross-product of their states:

$$\Omega = \mathrm{Val}(X_1) \times \mathrm{Val}(X_2) \times \mathrm{Val}(X_3)$$

- $X_i(\omega)$ is the value for X_i in the assignment $\omega \in \Omega$
- Specify $p(\omega)$ for each outcome $\omega \in \Omega$ by a big table:

x_1	x_2	x_3	$p(x_1, x_2, x_3)$
0	0	0	.11
0	0	1	.02
⋮	⋮		
1	1	1	.05

- How many parameters do we need to specify?

Example

- Consider three binary-valued random variables

$$X_1, X_2, X_3 \qquad \mathrm{Val}(X_i) = \{0, 1\}$$

- Let outcome space Ω be the cross-product of their states:

$$\Omega = \mathrm{Val}(X_1) \times \mathrm{Val}(X_2) \times \mathrm{Val}(X_3)$$

- $X_i(\omega)$ is the value for X_i in the assignment $\omega \in \Omega$
- Specify $p(\omega)$ for each outcome $\omega \in \Omega$ by a big table:

x_1	x_2	x_3	$p(x_1, x_2, x_3)$
0	0	0	.11
0	0	1	.02
	⋮		
1	1	1	.05

- How many parameters do we need to specify?

$$2^3 - 1$$

Marginalization

- Suppose X and Y are random variables with distribution $p(X, Y)$
  X: Intelligence, $\quad \text{Val}(X) = \{\text{“Very High”}, \text{“High”}\}$
  Y: Grade, $\quad \text{Val}(Y) = \{\text{“a”}, \text{“b”}\}$
- Joint distribution specified by:

		X = vh	X = h
Y	a	0.7	0.15
	b	0.1	0.05

- $p(Y = a) = ?= 0.85$
- More generally, suppose we have a joint distribution $p(X_1, \ldots, X_n)$. Then,

$$p(X_i = x_i) = \sum_{x_1} \sum_{x_2} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \cdots \sum_{x_n} p(x_1, \ldots, x_n)$$

Conditioning

- Suppose X and Y are random variables with distribution $p(X, Y)$

  X: Intelligence, $\quad \text{Val}(X) = \{\text{“Very High”}, \text{“High”}\}$

  Y: Grade, $\quad \text{Val}(Y) = \{\text{“a”}, \text{“b”}\}$

		X	
		vh	h
Y	a	0.7	0.15
	b	0.1	0.05

- Can compute the conditional probability

$$p(Y = a \mid X = vh) = \frac{p(Y = a, X = vh)}{p(X = vh)}$$

$$= \frac{p(Y = a, X = vh)}{p(Y = a, X = vh) + p(Y = b, X = vh)}$$

$$= \frac{0.7}{0.7 + 0.1} = 0.875.$$

Example: Medical diagnosis

- Variable for each **symptom** (e.g. "fever", "cough", "fast breathing", "shaking", "nausea", "vomiting")
- Variable for each **disease** (e.g. "pneumonia", "flu", "common cold", "bronchitis", "tuberculosis")
- Diagnosis is performed by **inference** in the model:

$$p(\text{pneumonia} = 1 \mid \text{cough} = 1, \text{fever} = 1, \text{vomiting} = 0)$$

- One famous model, Quick Medical Reference (QMR-DT), has 600 diseases and 4000 findings

Representing the distribution

- Naively, could represent multivariate distributions with table of probabilities for each outcome
- How many outcomes are there in QMR-DT? 2^{4600}
- **Estimation** of joint distribution would require a huge amount of data
- **Inference** of conditional probabilities, e.g.

$$p(\text{pneumonia} = 1 \mid \text{cough} = 1, \text{fever} = 1, \text{vomiting} = 0)$$

  would require summing over exponentially many variables' values
- Moreover, defeats the purpose of probabilistic modeling, which is to make predictions with *previously unseen observations*

Structure through independence

- If $X_1, \ldots, X_n$ are independent, then

$$p(x_1, \ldots, x_n) = p(x_1)p(x_2)\cdots p(x_n)$$

- 2^n entries can be described by just n numbers (if $|\text{Val}(X_i)| = 2$)!
- However, this is not a very *useful* model – observing a variable X_i cannot influence our predictions of X_j
- If $X_1, \ldots, X_n$ are *conditionally independent* given Y, denoted as $X_i \perp \mathbf{X}_{-i} \mid Y$, then

$$\begin{aligned} p(y, x_1, \ldots, x_n) &= p(y)p(x_1 \mid y)\prod_{i=2}^{n} p(x_i \mid x_1, \ldots, x_{i-1}, y) \\ &= p(y)p(x_1 \mid y)\prod_{i=2}^{n} p(x_i \mid y). \end{aligned}$$

Bayesian networks (directed PGMs)

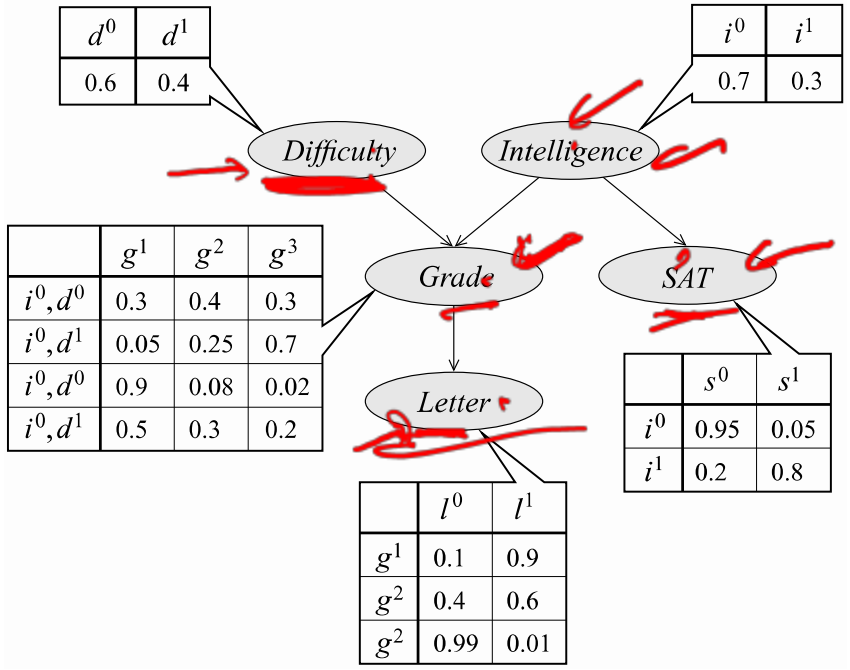
- A **Bayesian network** is specified by a directed *acyclic* graph $G = (V, E)$ with:
  1. One node $i \in V$ for each random variable X_i
  2. One conditional probability distribution (CPD) per node, $p(x_i \mid \mathbf{x}_{\mathrm{Pa}(i)})$, specifying the variable's probability conditioned on its parents' values
- Corresponds 1-1 with a particular factorization of the joint distribution:

$$p(x_1, \ldots x_n) = \prod_{i \in V} p(x_i \mid \mathbf{x}_{\mathrm{Pa}(i)})$$

- Powerful framework for designing *algorithms* to perform probability computations

Example

- Consider the following Bayesian network:



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^0, d^0	0.9	0.08	0.02
i^0, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^2	0.99	0.01

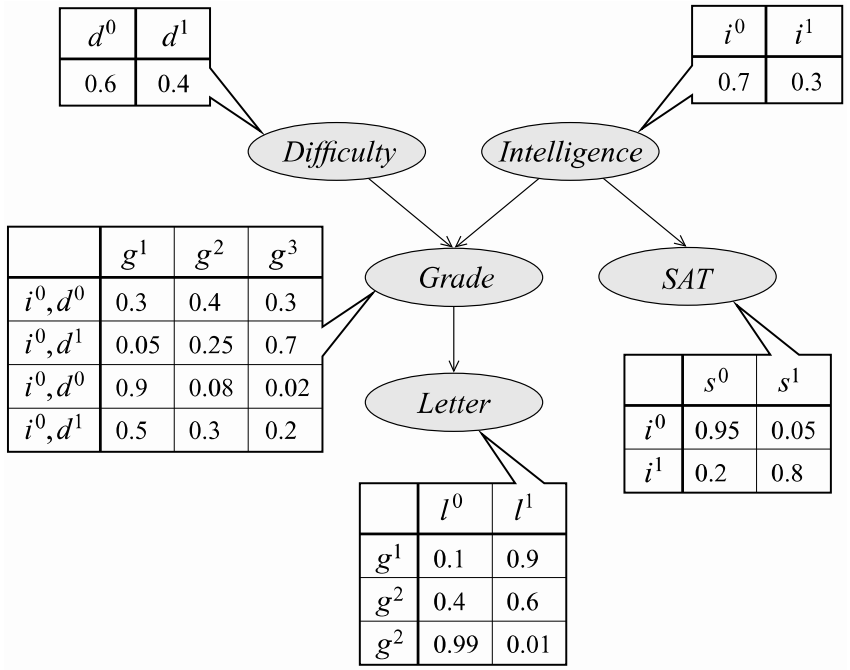
- What is its joint distribution?

$$p(x_1, \ldots x_n) = \prod_{i \in V} p(x_i \mid \mathbf{x}_{\mathrm{Pa}(i)})$$

$$p(d, i, g, s, l) = p(d)p(i)p(g \mid d, i)p(s \mid i)p(l \mid g)$$

Example

- Consider the following Bayesian network:



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^0, d^0	0.9	0.08	0.02
i^0, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^2	0.99	0.01

- What is its joint distribution?

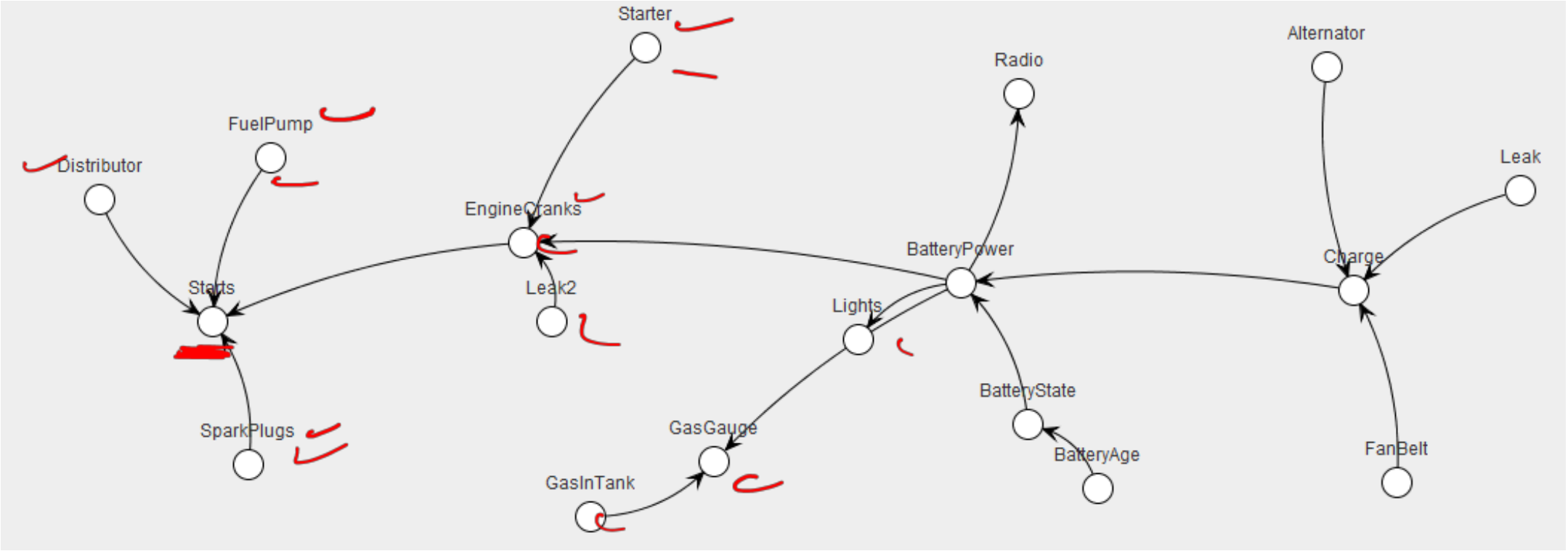
$$p(x_1, \ldots x_n) = \prod_{i \in V} p(x_i \mid \mathbf{x}_{\mathrm{Pa}(i)})$$

$$p(d, i, g, s, l) = p(d)p(i)p(g \mid i, d)p(s \mid i)p(l \mid g)$$

More Examples

$$p(x_1, \ldots x_n) = \prod_{i \in V} p(x_i \mid \mathbf{x}_{\mathrm{Pa}(i)})$$

Will my car start this morning?



Heckerman *et al.*, Decision-Theoretic Troubleshooting, 1995

More Examples

$$p(x_1, \ldots x_n) = \prod_{i \in V} p(x_i \mid \mathbf{x}_{\mathrm{Pa}(i)})$$

What is the differential diagnosis?

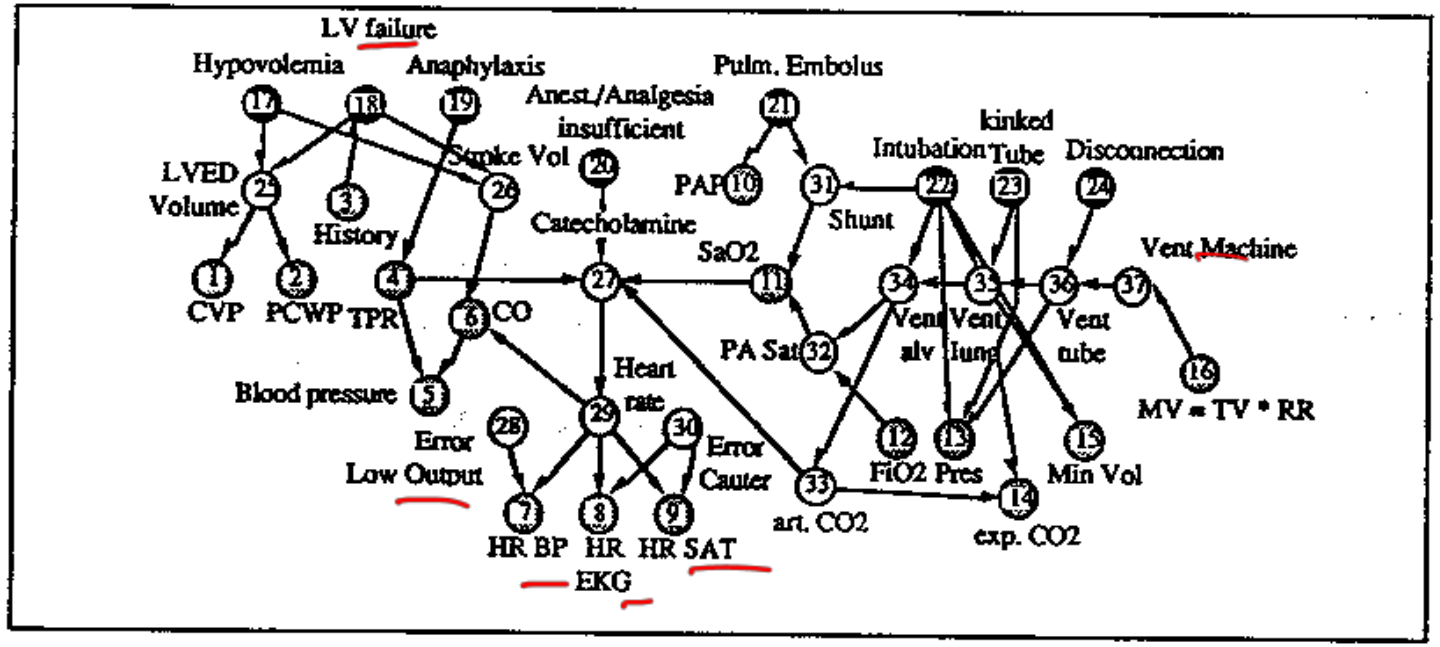


Fig. 1 The ALARM network representing causal relationships is shown with diagnostic (●), intermediate (○) and measurement (◎) nodes. CO: cardiac output, CVP: central venous pressure, LVED volume: left ventricular end-diastolic volume, LV failure: left ventricular failure, MV: minute ventilation, PA Sat: pulmonary artery oxygen saturation, PAP: pulmonary artery pressure, PCWP: pulmonary capillary wedge pressure, Pres: breathing pressure, RR: respiratory rate, TPR: total peripheral resistance, TV: tidal volume

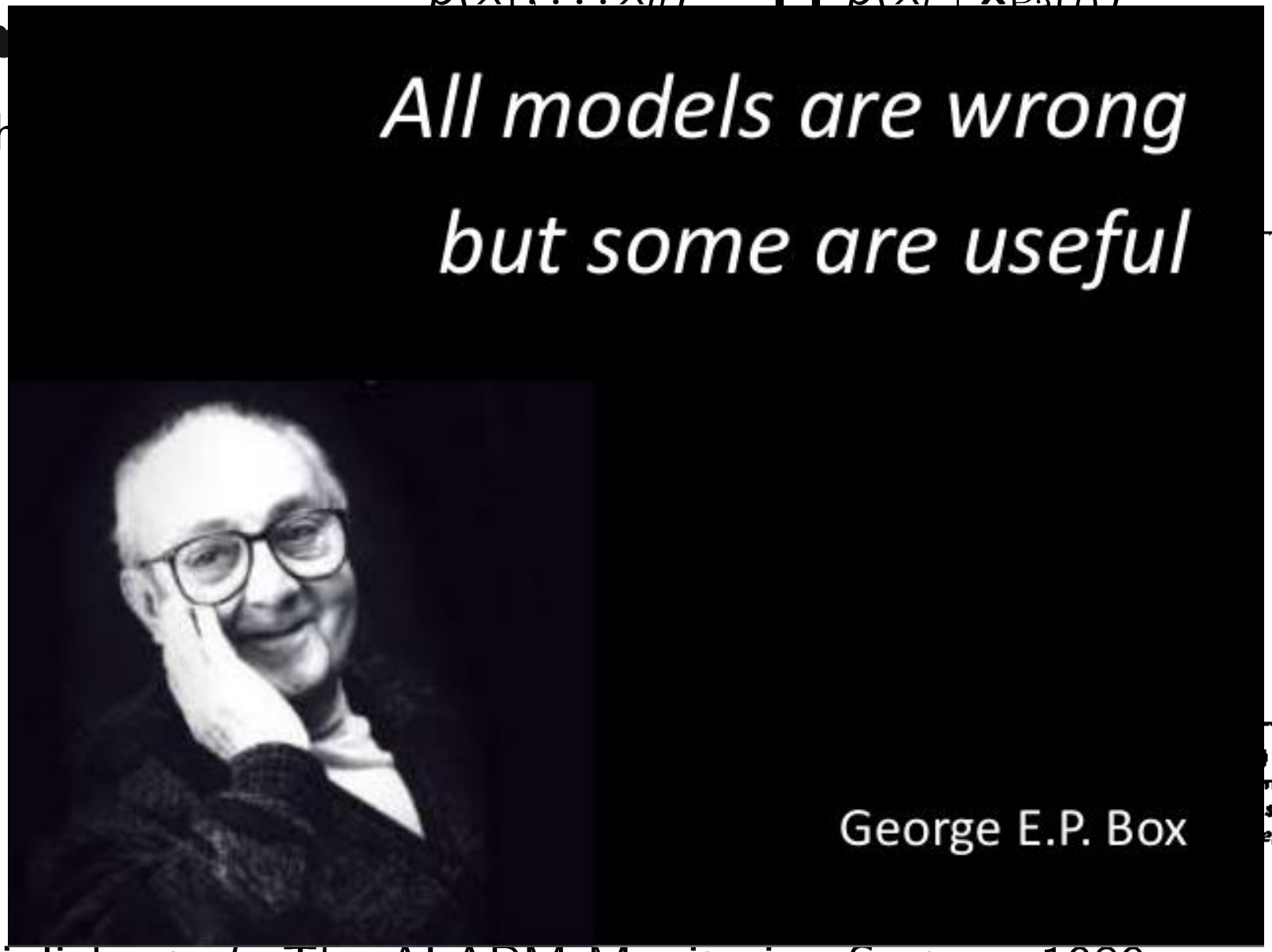
Beinlich *et al.*, The ALARM Monitoring System, 1989

[Sontag]

More Exan

$$p(x_1, \ldots, x_n) = \prod_i p(x_i \mid \mathbf{x}_{\mathrm{Pa}(i)})$$

Wh



Beinlich *et al.*, The ALARM Monitoring System, 1989

[Sontag]

Questions?