Fundamentals of Probability (Part 2)

Alexander G. Ororbia II and Viet Nguyen

Introduction to Machine Learning CSCI-335

2/4/2026

Computing Marginal Probability with the Sum Rule

$$\forall x \in \mathrm{x}, P(\mathrm{x} = x) = \sum_y P(\mathrm{x} = x, \mathrm{y} = y). \tag{3.3}$$

Summation → *Discrete* random variables!

$$p(x) = \int p(x, y) dy. \tag{3.4}$$

Integration → *Continuous* random variables!

Also known as marginalization!

Conditional Probability

$$P(\mathrm{y} = y \mid \mathrm{x} = x) = \frac{P(\mathrm{y} = y, \mathrm{x} = x)}{P(\mathrm{x} = x)}$$

In probability theory, **conditional probability** is a measure of the probability of an event given that (by assumption, presumption, assertion or evidence) another event has occurred

Chain Rule of Probability

$$P(\mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(n)}) = P(\mathrm{x}^{(1)}) \Pi_{i=2}^{n} P(\mathrm{x}^{(i)} \mid \mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(i-1)})$$

In probability theory, the **chain rule** (also called the **general product rule**) permits the calculation of any member of the joint distribution of a set of random variables using only conditional probabilities

Just remember: **P(A, B)** = P(A | B) x P(B) = **P(B | A) x P(A)**

P(B | A) x P(A) Meaning: The chance of both A and B happening is the chance of A happening `**P(A)**` multiplied by the chance of B happening assuming A is already true `**P(B | A)**`.

Example: P(Flu, Cough) = P(Cough | Flu) x P(Flu)

60% probs that people with flu will have a cough

There are 2% people having flu in the population

Chain Rule of Probability

Also remember: **$P(A, B \mid C)$** = $P(A \mid B, C) \times P(B \mid C)$ = **$P(B \mid A, C) \times P(A \mid C)$**
$P(B \mid A, C) \times P(A \mid C)$ Meaning: if both events A and B occur, given that some background condition C is already true => event B occurs given A and C, and event A occurs given C

Example: Job interview
A: solving coding challenges
B: receiving job offer
C (already given): you are in the final round interview
$P(A, B \mid C) = P(B \mid A, C) \times P(A \mid C)$
$P(B \mid A, C) = 90\%$: if coding challenges are solved => most likely receiving the offer
$P(A \mid C) = 40\%$: there are about 40% of people in the final interview can solve the coding challenges

Fundamentals of Probability (Part 3)

Alexander G. Ororbia II and Viet Nguyen

Introduction to Machine Learning CSCI-635

2/6/2026

Independence

$$\forall x \in \mathrm{x}, y \in \mathrm{y},\ p(\mathrm{x} = x, \mathrm{y} = y) = p(\mathrm{x} = x)p(\mathrm{y} = y)$$

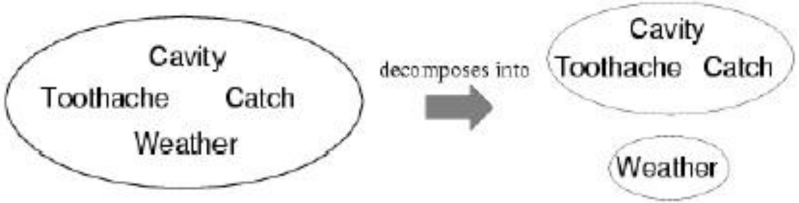
Two events are **independent**, **statistically independent**, or **stochastically independent** if occurrence of one does not affect probability of occurrence of the other

Similarly, two random variables are independent ***if*** realization of one does not affect probability distribution of other

(Absolute) Independence

A and *B* are independent iff (Note: following are equivalent)

$\mathbf{P}(A|B) = \mathbf{P}(A)$ or $\mathbf{P}(B|A) = \mathbf{P}(B)$ or $\mathbf{P}(A, B) = \mathbf{P}(A)\,\mathbf{P}(B)$



$\mathbf{P}(Toothache, Catch, Cavity, Weather)$
$= \mathbf{P}(Toothache, Catch, Cavity)\,\mathbf{P}(Weather)$

Because we know they will not depend on weather, we can remove the weather condition

32 (2^3 * 4 (*Weather*)) entries reduced to 12 (2^3 + 4 (*Weather*))

Absolute independence is powerful, but rare.

For your review!

Conditional Independence

$$\forall x \in \mathrm{x}, y \in \mathrm{y}, z \in \mathrm{z},\ p(\mathrm{x}=x, \mathrm{y}=y \mid \mathrm{z}=z) = p(\mathrm{x}=x \mid \mathrm{z}=z)p(\mathrm{y}=y \mid \mathrm{z}=z)$$

Two events *x* and *y* are **conditionally independent** given a third event *z* if occurrence of *x* *and* occurrence of *y* are independent events in their conditional probability distribution given *z*

In other words, *x* and *y* are conditionally independent given *z* if and only if (***iff***), given **knowledge** that *z* occurs, **knowledge** of whether *x* occurs provides **no information** on likelihood of *y* occurring, and **knowledge** of whether *y* occurs provides **no information** on likelihood of *x* occurring

Conditional Independence

If I have a cavity, probability the probe catches doesn't depend on whether I have a toothache:

(1) **P**(*catch* | *toothache, cavity*) = **P**(*catch* | *cavity*)

The same independence holds if I haven't got a cavity:

(2) **P**(*catch* | *toothache*, ¬*cavity*) = **P**(*catch* | ¬*cavity*)

Catch is conditionally independent of *Toothache* given *Cavity*:

(3) P(*Toothache, Catch* | *Cavity*) = **P**(*Toothache* | *Cavity*) **P**(*Catch* | *Cavity*)

Equivalent statements:

P(*Toothache* | *Catch, Cavity*) = **P**(*Toothache* | *Cavity*)

P(*Catch* | *Toothache, Cavity*) = **P**(*Catch* | *Cavity*)

For your review!

Conditional independence

We can now write out the full joint distribution as:

P(*Toothache, Catch, Cavity*)
= **P**(*Toothache, Catch* | *Cavity*) **P**(*Cavity*) // *product rule*
= **P**(*Toothache* | *Cavity*) **P**(*Catch* | *Cavity*) **P**(Cavity) // *cond. ind.*

In many cases

Use of conditional independence reduces the size of the representation of the joint distribution from exponential in *n* to linear in *n*

Conditional independence

Our most basic and robust form of knowledge about uncertain environments

For your review!

Bayes, in English Please?

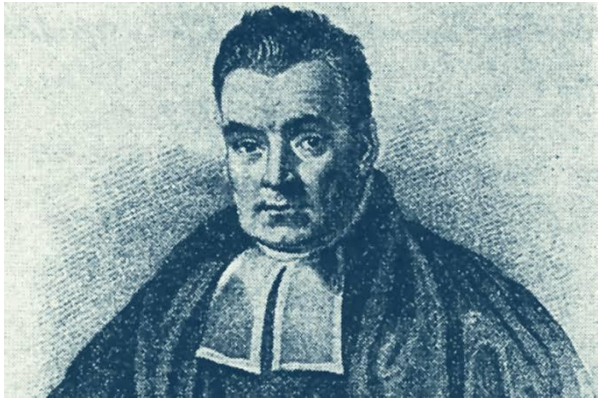
- What does Bayes' Formula helps to find?
  - Helps us to find:

$$P(B \mid A)$$

  - By having already known:

$$P(A \mid B)$$

$$P(x \mid y) = \frac{P(x)P(y \mid x)}{P(y)}$$



Thomas Bayes, 1701-1761

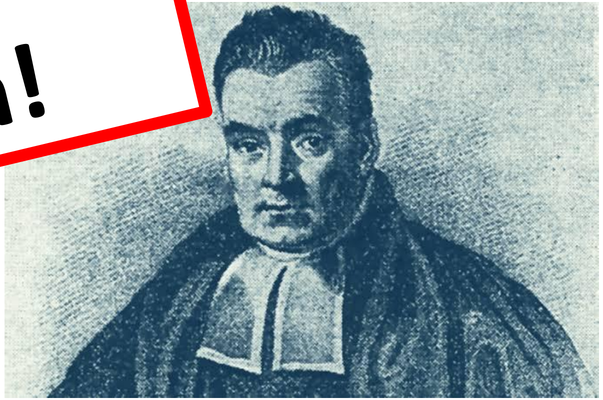
Bayes, in English Please?

- What does Bayes' Formula helps to find?
  - Helps us to find:

$$P(B \mid A$$

Bayes Theorem shall return!

$$y) = \frac{P(\mathrm{x})P(\mathrm{y} \mid \mathrm{x})}{P(\mathrm{y})}$$



Thomas Bayes, 1701-1761

  - By having already known:

$$P(A \mid B)$$

Bayes' Rule

Product rule: $P(a \wedge b) = P(a \mid b) P(b) = P(b \mid a) P(a)$

$\Rightarrow$ Bayes' rule: $P(a \mid b) = P(b \mid a) P(a) / P(b)$

or in distribution form

$$\mathbf{P}(Y|X) = \mathbf{P}(X|Y) \mathbf{P}(Y) / \mathbf{P}(X) = \alpha \mathbf{P}(X|Y) \mathbf{P}(Y)$$

Causal Probability (useful for diagnostics):

We are having a stiff neck, we want to know if we actually have the disease:

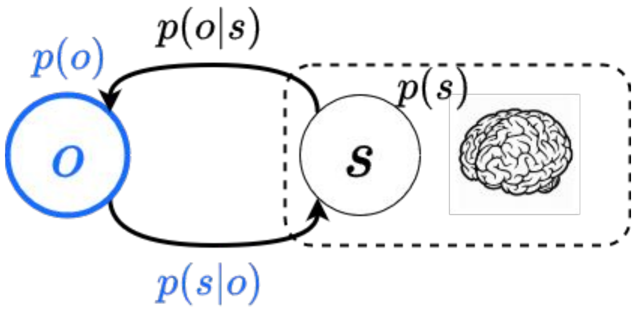
P(Cause | Effect) = P(Effect|Cause) P(Cause) / P(Effect)

E.g., let *M* be meningitis, *S* be stiff neck:

$$P(m|s) = P(s|m) P(m) / P(s) = 0.8 \times 0.0001 / 0.1 = 0.0008$$

Note: posterior probability of meningitis still very small!

Bayes' Rule



$$p(s|o) = \frac{p(o|s)p(s)}{p(o)}$$

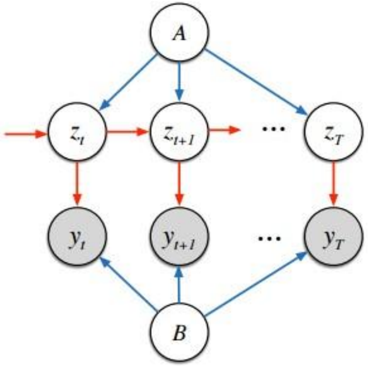
- p(s): **prior** over **state / knowledge / cause** (meningitis)
- p(o): **model evidence / observation / effect** (stiff neck)
- p(o|s): likelihood function (how likely do we have stiff neck given meningitis)
- p(s|o): **posterior** (P of meningitis given the stiff neck observation)

Summary of Probability

- Probability is rigorous formalism for *uncertain* knowledge
- Joint probability distribution specifies probability of every atomic event (in sample/event space)
- *Queries* can be answered by summing over atomic events
- For nontrivial domains, we must find ways to reduce joint probability distributional search space
  - *Independence* & *conditional independence* = your tools for reducing joint probability distribution table size

- ***Note***: These ideas/axioms apply equally well to vector/matrix variables

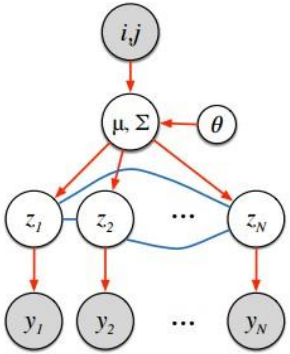
Probability allows us to build models of stochastic, data-generating processes….



Gaussian Linear State Space Model
Kalman Filter

$$z_t \sim \mathcal{N}(z_t|Az_{t-1}, \sigma_z^2 I)$$

$$y_t \sim \mathcal{N}(y_t|Bz_t, \sigma_y^2 I)$$

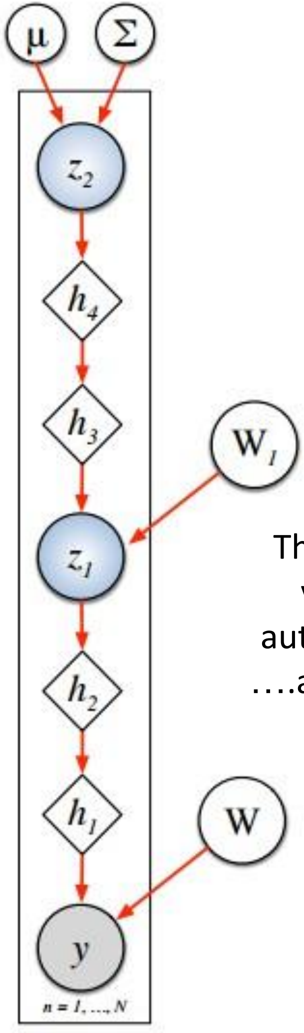


Latent Gaussian Cox Point Process

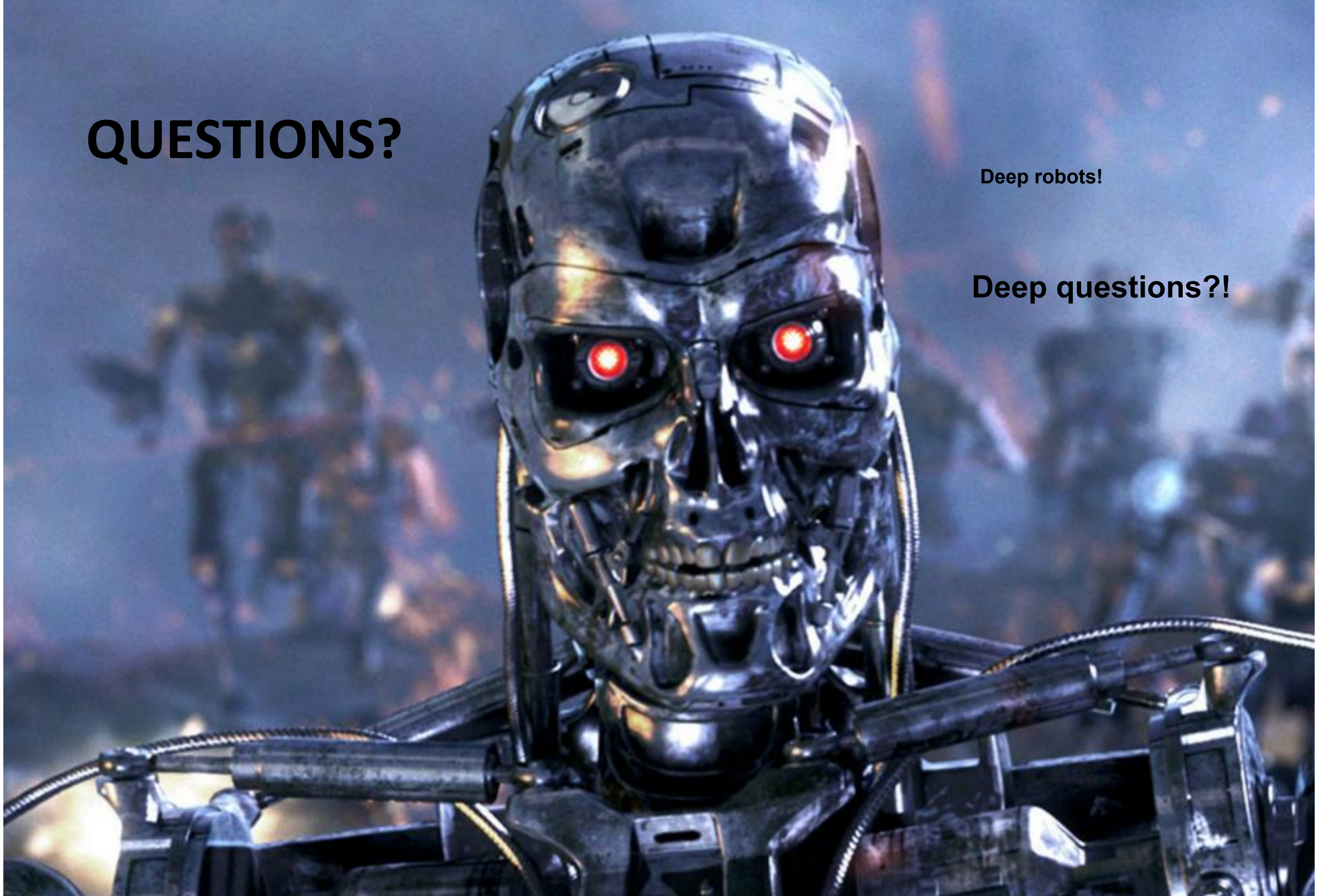
$$x \sim \mathcal{N}(x|\mu(i,j), \Sigma(i,j))$$

$$y_{ij} \sim \mathcal{P}(c\exp(x_{ij}))$$

Probabilistic graphical models (PGMs)



The venerable variational autoencoder….
….a neural PGM.



QUESTIONS?

Deep robots!

Deep questions?!