Fundamentals of Probability (Part Deux)

Alexander G. Ororbia II

Introduction to Machine Learning

CSCI-635

9/6/2023

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

Independence

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

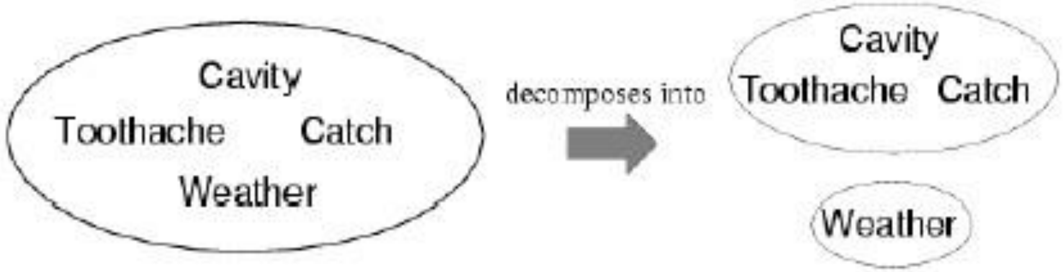
Two events are **independent**, **statistically independent**, or **stochastically independent** if occurrence of one does not affect probability of occurrence of the other

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



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

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

Absolute independence is powerful, but rare.

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) $\mathbf{P}(catch \mid toothache, cavity) = \mathbf{P}(catch \mid cavity)$

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

(2) $\mathbf{P}(catch \mid toothache, \neg cavity) = \mathbf{P}(catch \mid \neg cavity)$

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

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

Equivalent statements:

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

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

Conditional independence

We can now write out the full joint distribution as:

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

$= \mathbf{P}(Toothache, Catch \mid Cavity)\, \mathbf{P}(Cavity)$ *// product rule*

$= \mathbf{P}(Toothache \mid Cavity)\, \mathbf{P}(Catch \mid Cavity)\, \mathbf{P}(\mathrm{Cavity})$ *// 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

Bayes, in English Please?

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

$$P(B|A)$$

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

  - By having already known:

$$P(A|B)$$

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):

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

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

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

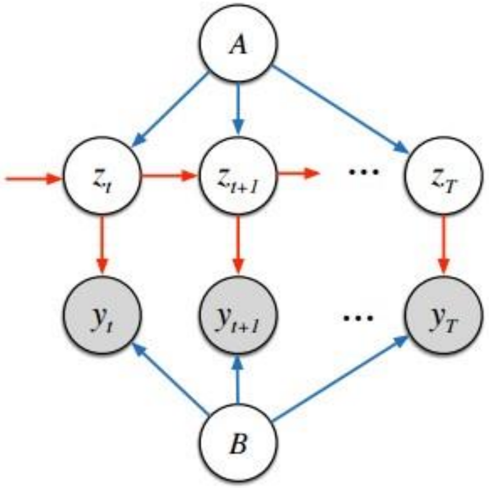
Note: posterior probability of meningitis still very small!

Summary of Probability

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

- Notions apply equally well to vector/matrix variables

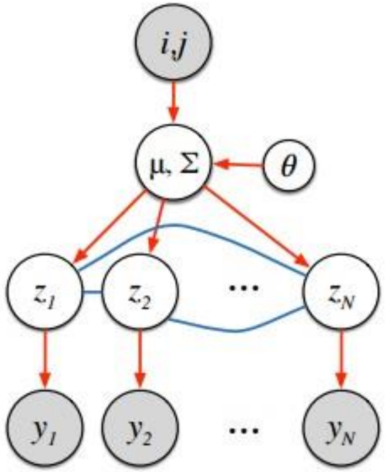
Probability allows us to build models of stochastic, data-generating processes….



Gaussian Linear State Space Model
Kalman Filter

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

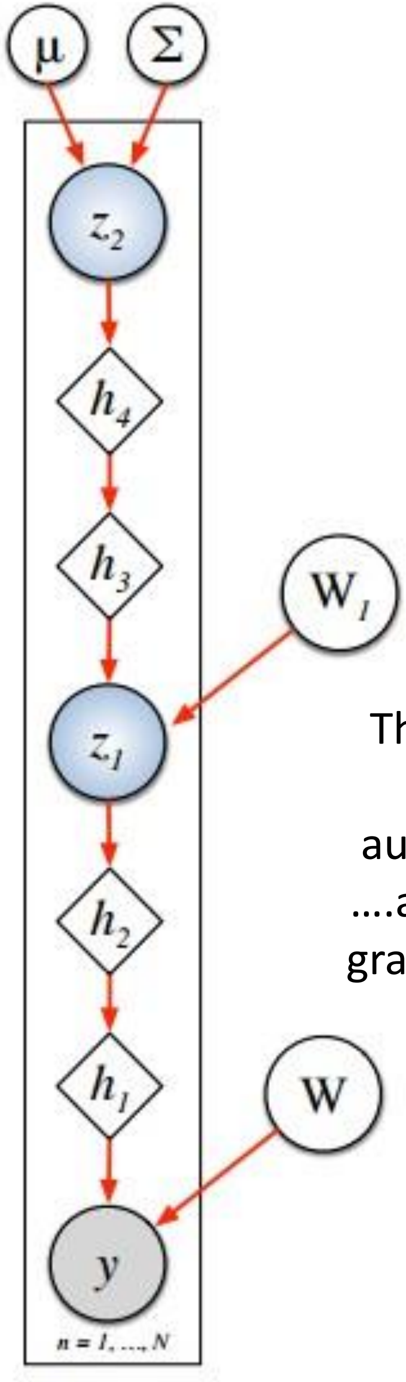
$$y_t \sim \mathcal{N}(y_t|Bz_t, \sigma_y^2 I)$$



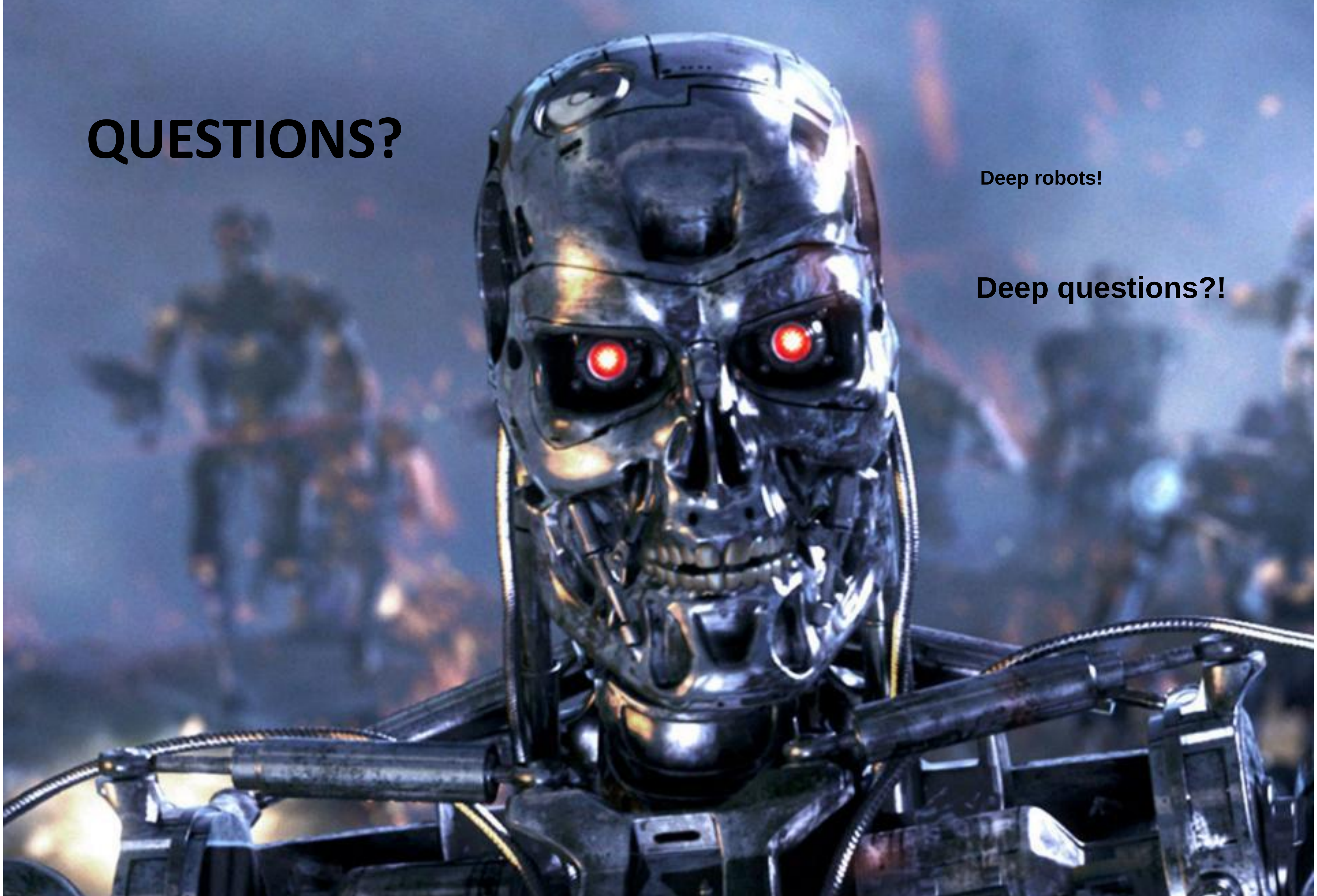
Latent Gaussian Cox Point Process

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

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



The venerable variational autoencoder….
….a probabilistic graphical model.



QUESTIONS?

Deep robots!

Deep questions?!