EECS 496: Sequential Decision Making

Soumya Ray

sray@case.edu

Office: Olin 516

Office hours: T 4-5:30 or by appointment

Announcement

- Assignment posted
  - Due Saturday 9/21 11:59pm
- Grader: Zicheng Gao (zxg109)

Recap

- What is probability theory?
- What is a random variable?
- What is an atomic event? Event? Sample space?
- What are the axioms of probability?
- What is the “joint pdf”?
- What is $p(X|Y)$?
- What is Bayes’ Rule? What is its significance?
- When are two r.v.’s independent? What is the significance of independence?
- What is the expectation of a rv?
- What is the variance of a rv?
- What is conditional independence?
- In probabilistic inference, we want the pdf over a _________ given ________.
- One method for probabilistic inference is ____________.

Today

- Probabilistic inference (Ch 13)
- Bayesian Networks (Ch 14, Russell and Norvig)

Inference by Enumeration

- We are given a pdf over a collection of r.v.'s $\mathbf{X}$
  - Of these, we observe evidence $\mathbf{E}=\mathbf{e}$ ($\mathbf{E} \subseteq \mathbf{X}$)
  - We are interested in the query variable V
  - Let $\mathbf{Y}$ be $\mathbf{X} \setminus \{\mathbf{E}, V\}$ (everything in $\mathbf{X}$ not in $\mathbf{E}$ and not V)
    - Note $\mathbf{X} = \mathbf{Y} \cup \mathbf{E} \cup V$
    - Sometimes called "nuisance" variables
- We want $p(V=v|\mathbf{E}=\mathbf{e})$

Inference by Enumeration

$$p(V=v \mid \mathbf{E}=\mathbf{e}) = \frac{p(V=v, \mathbf{E}=\mathbf{e})}{p(\mathbf{E}=\mathbf{e})}$$

Marginalization

$$p(V=v, \mathbf{E}=\mathbf{e}) = \sum_{\mathbf{y}} p(V=v, \mathbf{E}=\mathbf{e}, \mathbf{Y}=\mathbf{y}), \quad \mathbf{Y} = \mathbf{X} \setminus \{\mathbf{E}, V\}$$

Atomic Event

$$p(\mathbf{E}=\mathbf{e}) = \sum_{v} \sum_{\mathbf{y}} p(V=v, \mathbf{E}=\mathbf{e}, \mathbf{Y}=\mathbf{y})$$

Normalization Factor

$$p(V=v \mid \mathbf{E}=\mathbf{e}) = \frac{\sum_{\mathbf{y}} p(V=v, \mathbf{E}=\mathbf{e}, \mathbf{Y}=\mathbf{y})}{\sum_{v} \sum_{\mathbf{y}} p(V=v, \mathbf{E}=\mathbf{e}, \mathbf{Y}=\mathbf{y})}$$

Example

CloudyTomorrow	RainTomorrow	WetGrass	Probability
No	No	No	0.4
No	No	Yes	0.01
No	Yes	No	0
No	Yes	Yes	0.01
Yes	No	No	0.15
Yes	No	Yes	0.02
Yes	Yes	No	0.01
Yes	Yes	Yes	0.4

$$p(WetGrass = Yes \mid CloudyTomorrow = Yes)?$$

Solution

$$p(WetGrass = Yes \mid CloudyTomorrow = Yes) \propto$$
$$p(W = Yes, C = Yes, R = Yes) + p(W = Yes, C = Yes, R = No) \propto$$
$$0.4 + 0.02 = 0.42$$

$$p(WetGrass = No \mid CloudyTomorrow = Yes) \propto$$
$$p(W = No, C = Yes, R = Yes) + p(W = No, C = Yes, R = No) \propto$$
$$0.01 + 0.15 = 0.16$$
$$c = \frac{1}{(0.42 + 0.16)} = 1.724$$
$$p(WetGrass = Yes \mid CloudyTomorrow = Yes) = 0.42c = 0.724$$
$$p(WetGrass = No \mid CloudyTomorrow = Yes) = 0.16c = 0.276$$

Bayesian Networks (Motivation)

- A key need in all of AI is to reason with uncertain information
  - i.e., given you know the values of some quantities (“evidence”), find the distribution over values of some other quantities (“query variables”)
  - I observe a sequence of sensor measurements from my location. Each sensor is noisy. Where am I located?
- This is hard!
  - General probabilistic inference is NP-hard

Rule-based Expert Systems

- To get around the hardness of the problem, early AI researchers devised heuristic solutions
- “Expert Systems” had lots of weighted rules:
  - “If sensor1>5 and sensor2>4, x-location=5: 5.6”
  - “If sensor2<7 and sensor3>6, x-location=7: 3.4”
- What if multiple rules were true?
  - Then you had to combine the weights using arcane combining rules
  - This procedure was very ad-hoc and sometimes led to conflicting results

Key Point 1

- Real-world systems can be described by large numbers of variables, but typically *only a few interact with each other*

- So we can take advantage of *statistical independence* during inference
  - This makes probabilistic inference practical on a large scale

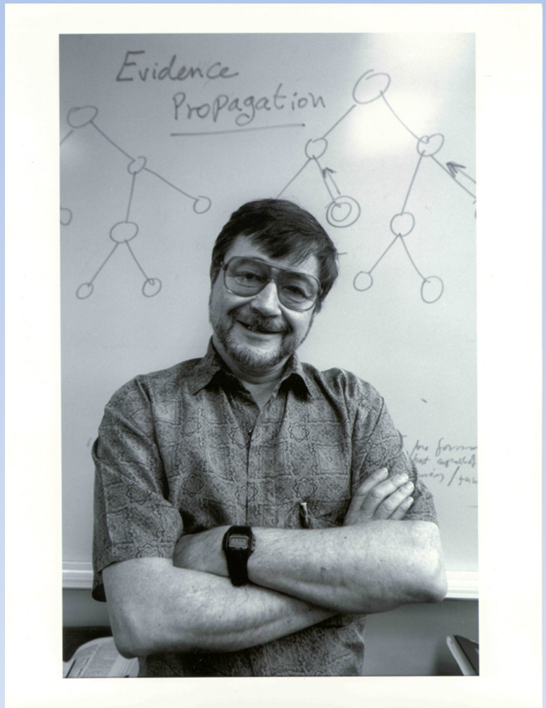
Statistical Independence

- Two r.v.'s X and Y are statistically independent if

$$p_{X,Y}(X=x, Y=y) = p_X(X=x)\, p_Y(Y=y)$$

Key Point 2

- Once the probability distributions are *factored* using independence, they can be *represented as graphs*
- These ideas lead to *Bayesian Networks*
  - Developed by Judea Pearl (Turing award winner 2012) among others



Bayesian Networks

- A way of representing the probability distribution over a collection of random variables

- The probability distribution is represented as a ***graph***
  - This is a kind of “graphical model”

- Inference operations can be made efficient by taking advantage of the graph structure

The Chain Rule

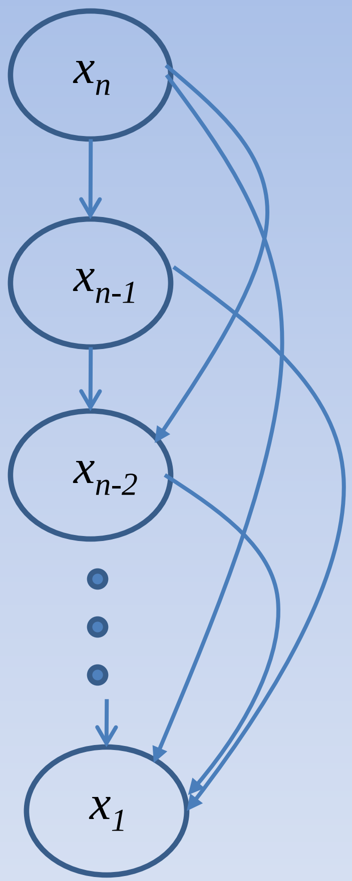
- Consider n random variables $X_1,\ldots,X_n$

$$\Pr(x_1,\ldots,x_n) = \Pr(x_1,\ldots,x_{n-1} \mid x_n)\Pr(x_n)$$

$$= \Pr(x_1,\ldots,x_{n-2} \mid x_{n-1},x_n)\Pr(x_{n-1} \mid x_n)\Pr(x_n)$$

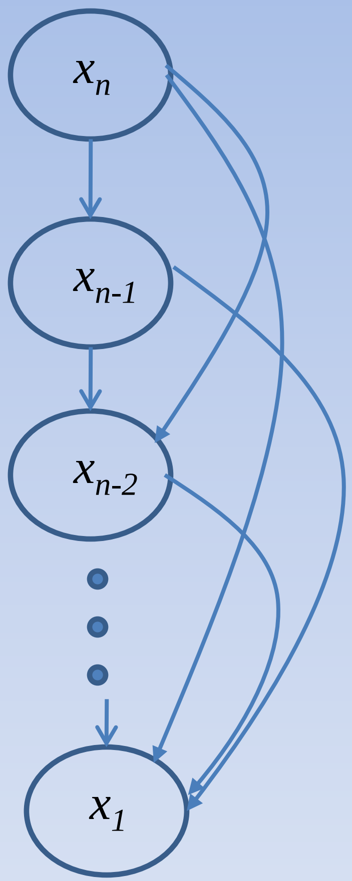
$$= \Pr(x_n)\prod_{i=1}^{n-1}\Pr(x_i \mid \{x_j\}_{j=i+1}^{n})$$

The Chain Rule as a Graph



$$\Pr(x_1,\ldots,x_n) = \Pr(x_n)\prod_{i=1}^{n-1}\Pr(x_i \mid \{x_j\}_{j=i+1}^{n})$$

The Chain Rule as a Graph

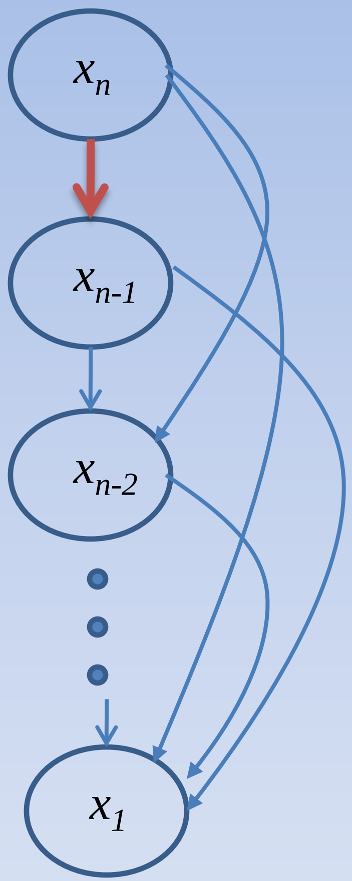


- Each node represents a random variable
- A directed edge represents a conditioning dependence
  - So x_{n-1} is conditioned on x_n, x_{n-2} on x_n and x_{n-1}, etc
- The "**parents**" of a node x_i, $Pa(x_i)$, are the other nodes x_j with edges to x_i

Properties of the Network

- It is a DAG
  - "Directed Acyclic Graph"–If you follow the directed edges, you can't start from x_i and get back to it
  - But there are lots of *undirected* cycles
- It is not unique
  - Reorder the variables
  - Therefore, any probability distribution can be represented using many graphical structures

The Chain Rule as a Graph



- What would happen if I *deleted* an edge from this graph?

$$\Pr(x_1,\ldots,x_n)=\Pr(x_n)\prod_{i=1}^{n-1}\Pr(x_i\mid\{x_j\}_{j=i+1}^{n})$$

$$\Pr(x_n)\Pr(x_{n-1})\prod_{i=1}^{n-2}\Pr(x_i\mid\{x_j\}_{j=i+1}^{n})$$

$= \Pr(x_1,\ldots,x_n)$ iff x_{n-1} is independent of x_n

Key idea

- The "chain rule graph" is always an exact representation for any joint distribution
- Suppose for some x_i, we *know* that it is independent of an ancestor x_{i+k} given the other parents:

$$\Pr(x_i \mid x_{i+1},\ldots,x_{i+k},\ldots x_n) =$$
$$\Pr(x_i \mid x_{i+1},\ldots,x_{i+k-1},x_{i+k+1},\ldots,x_n)$$

- In the chain rule graph, we can delete the edge $x_{i+k} \rightarrow x_i$ and it will *still* represent the joint distribution

The Bayesian network assumption

- Consider an **arbitrary DAG** over n random variables

- This DAG *still* represents the joint probability distribution iff for all x_i, x_i is independent of all its *non-descendants* given its *parents*
  - Called the "Bayesian Network Assumption"

BNA and the Chain rule

- So for an arbitrary DAG,

$$\Pr(x_1, \ldots, x_n)$$

$$= \Pr(x_n) \prod_{i=1}^{n-1} \Pr(x_i \mid \{x_j\}_{j=i+1}^{n})$$

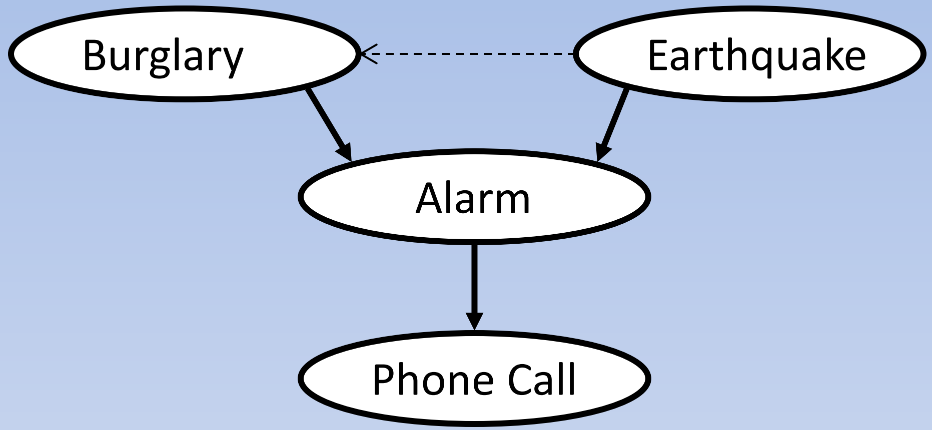
$$\boxed{= \prod_{i=1}^{n} \Pr(x_i \mid Pa(x_i))}$$

By BNA

Example (J. Pearl et al.)

- My house has a sensitive burglar alarm which occasionally also goes off if there is an earthquake. If the alarm goes off, my neighbor might call and tell me about it.
- How to describe this with a BN?
  - Nodes?
  - Edges?

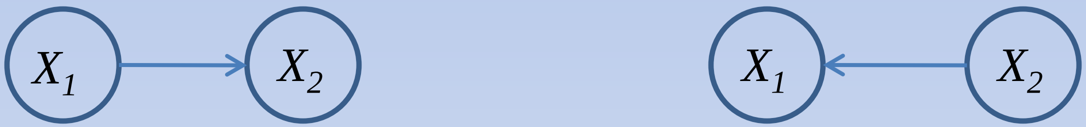
Alarm network



Pr(B,E,A,P)=?

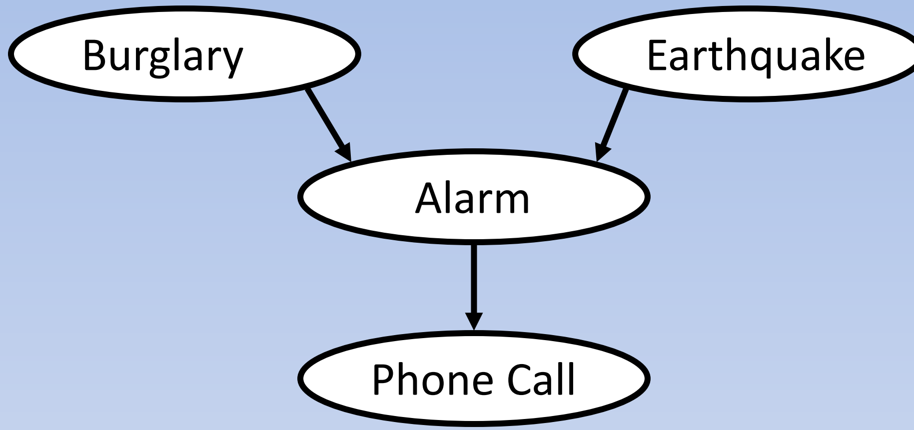
The Meaning of an Edge

- Sometimes, it is useful to think of an edge $x_i \longrightarrow x_j$ as being a “causal” relationship
- Consider the two networks:



- These represent the exact same probability distribution, $Pr(X_1,X_2)$
  - **Independence** is **symmetric**, **causality** is not (usually)
- A network constructed to be causal will be a BN, but not all BNs are causal

Explaining Away



- When Alarm is unknown, Burglary and Earthquake are independent
- But if the Alarm goes off, then they become dependent because they “compete” to explain the Alarm