EECS 391: Introduction to AI

Soumya Ray

Website: http://vorlon.case.edu/~sray/eecs391_sp12/index.html

Email: sray@case.edu

Office: Olin 516

Office hours: Tue 2:30-4:00 or by appointment

Announcements

- PA3 due
- HW5 online, due 27th
- Short meeting to discuss midterm grades, Quiz 3 and 4, Friday 3-4pm here

Today

- Probabilistic Inference (Ch 13)
- Bayesian Networks (Ch 14 sections 1, 2, 4.1, 4.2, 5.1)

Review

- Atomic event?
- Event?
- PDF?
- Conditional Probability?
- Bayes Rule?
- Independence?

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

Atomic Event

Event

Joint Probability Density Function

$$\Pr(C = No \mid R = No, W = No) = ?$$

Tasks

- There are two general tasks we will be interested in when working with random variables and their associated distributions
  - Inference
  - Estimation

Inference (Reasoning)

- We are given a collection of random variables defining a sample space and a joint density function
- Some of these variables are observed and their values fixed (evidence)
- We want to find the probability distribution of *some* of the remaining variables (query variables)

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

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

Query variable

evidence

Inference by Enumeration

- We have a collection of r.v.'s $\mathbf{X}$
  - Of these, we observe $\mathbf{E}=\mathbf{e}$ ($\mathbf{E} \subseteq \mathbf{X}$)
  - 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 $\mathbf{X} = \mathbf{Y} \cup \mathbf{E} \cup V$

- We want $p(V=v|\mathbf{E}=\mathbf{e})$

Inference by Enumeration

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

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

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

Marginalization

Atomic Event

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

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

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

Example

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

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 2011) among others

Bayesian Networks

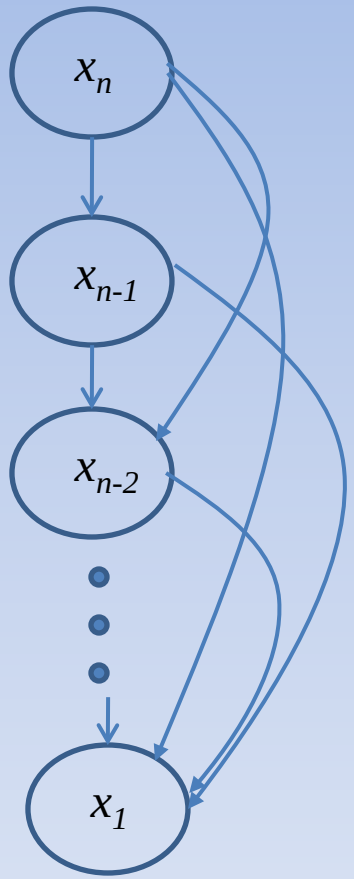
- A way of representing the probability distribution over a collection of random variables
- The probability distribution is represented as a ***graph***
  - This is a kind of “graphical model”
- Query operations can be made efficient by taking advantage of the graph structure

The Chain Rule

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

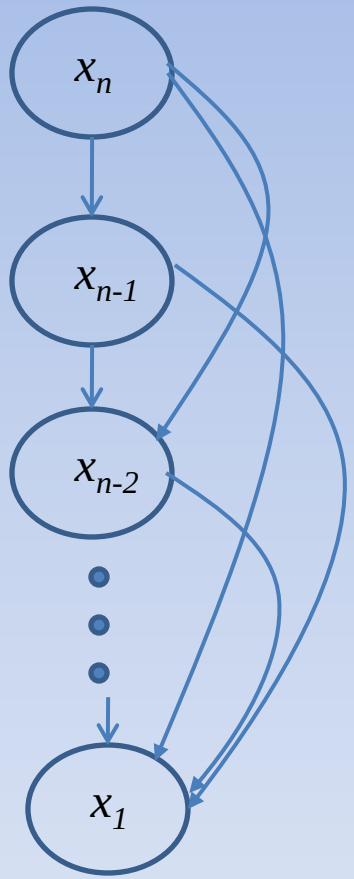
$$\Pr(x_1,\ldots,x_n) = \Pr(x_1,\ldots,x_{n-1} \mid x_n)\Pr(x_n)$$
$$= \Pr(x_1,\ldots,x_{n-2} \mid x_{n-1},x_n)\Pr(x_{n-1} \mid x_n)\Pr(x_n)$$
$$= \Pr(x_n)\prod_{i=1}^{n-1}\Pr(x_i \mid \{x_j\}_{j=i+1}^{n})$$

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

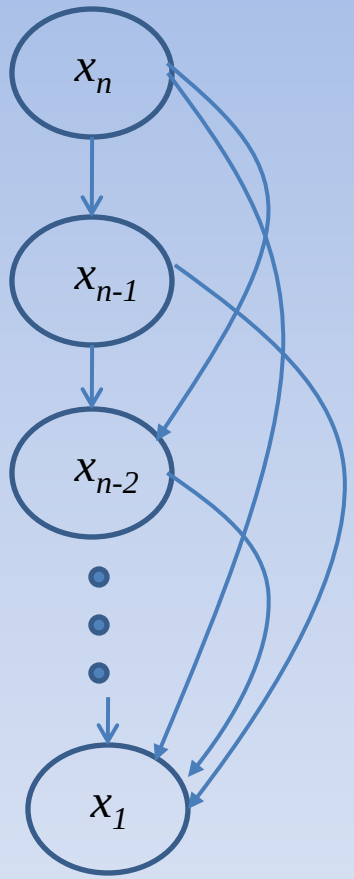


- Each node represents a random variable
- A directed edge represents a conditioning dependence
  - So x_{n-1} is conditioned on x_n, x_{n-2} on x_n and x_{n-1}, etc
- The "**parents**" of a node x_i, $Pa(x_i)$, are the other nodes x_j with edges to x_i
- Each node stores the distribution $\Pr(x_i|Pa(x_i))$

Properties of the Graphical Model

- 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

The Bayesian network assumption

- The "chain rule graph" was an exact representation for any joint distribution
- Suppose for some x_i, we *knew* that it was 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 graph, we can delete this edge

The Bayesian network assumption

- Consider an arbitrary DAG over n random variables
- This DAG 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"
  - Or sometimes (confusingly) the "Markov condition"
- How to get non-descendants?

BNA and the Chain rule

- So for an arbitrary DAG,

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

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

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

By BNA

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