Readings: K&F 2.1, 2.2, 2.3, 3.1

Introduction to Probabilistic Graphical Models

Lecture 1 – Mar 28, 2011
CSE 515, Statistical Methods, Spring 2011

Instructor: Su-In Lee
University of Washington, Seattle

Logistics

- Teaching Staff
  - Instructor: Su-In Lee (suinlee@uw.edu, PAC 536)
    - Office hours: Fri 9-10am or by appointment (PAC 536)
  - TA: Andrew Guillory (guillory@cs.washington.edu)
    - Office hours: Wed 1:30-2:20 pm or by appointment (PAC 216)
- Course website
  - cs.washington.edu/515
  - Discussion group: course website
- Textbook
  - (required) Daphne Koller and Nir Friedman, Probabilistic Graphical Models: Principles and Techniques, MIT Press
  - Various research papers (copies available in class)

Course requirement

- 4 homework assignments (60% of final grade)
  - Theory / implementation exercises
  - First one goes out next Monday!
  - 2 weeks to complete each
  - HW problems are long and hard
    - Please, please, please start early!
  - Late/collaboration policies are described on the website
- Final exam (35%)
  - Date will be announced later.
- Participation (5%)

Probabilistic graphical models (PGMs)

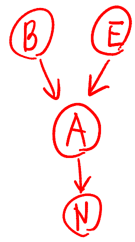
- One of the most exciting developments in machine learning (knowledge representation, AI, EE, Stats, …) in the last two decades…
- Tool for representing complex systems and performing sophisticated reasoning tasks
- Why have a model?
  - Compact and modular representation of complex systems
  - Ability to efficiently execute complex reasoning tasks
  - Make predictions
  - Generalize from particular problem

Probabilistic graphical models (PGMs)

- Many classical probabilistic problems in statistics, information theory, pattern recognition, and statistical mechanics are special cases of the formalism
  - Graphical models provides a common framework
  - Advantage: specialized techniques developed in one field can be transferred between research communities
- PGMs are a marriage between graph theory and probability theory
  - Representation: graph
  - Reasoning: probability theory
  - Any simple example?

A simple example

- We want to know/model whether our neighbor will inform us of the alarm being set off
- The alarm can set off (A) if
  - There is a burglary (B)
  - There is an earthquake (E)



- Whether our neighbor calls (N) depends on whether the alarm is set off (A)
- "Variables" in this system
  - Whether alarm being set off (**A**); burglary (**B**); earthquake (**E**); our neighbor calls (**N**)

A si

Probabilistic Inference

Task I Say that the alarm is set off (A=True), then how likely is it to get a call from our neighbor (N=True)?

Task II Given that my neighbor calls (N=True), how likely is it that a burglary occurred (B=True)?

- Var
  - Earthquake (E), Burglary (B), Alarm (A), NeighborCalls (N)

E	B	A	N	Prob.
F	F	F	F	0.01
F	F	F	T	0.04
F	F	T	F	0.05
F	F	T	T	0.01
F	T	F	F	0.02
F	T	F	T	0.07
F	T	T	F	0.2
F	T	T	T	0.1
T	F	F	F	0.01
T	F	F	T	0.07
T	F	T	F	0.13
T	F	T	T	0.04
T	T	F	F	0.06
T	T	F	T	0.05
T	T	T	F	0.1
T	T	T	T	0.05

$\leftarrow$ 1%

$\leftarrow$ 5%

$2^n - 1$

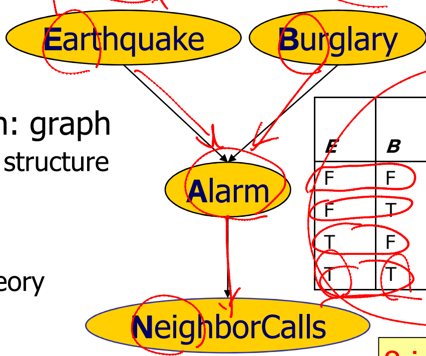
2^4-1 independent parameters

$\sum_i P_i = 1 \quad 1 - \sum_{i \neq 1} P_i$

A simple example

E	
F	T
0.9	0.1

B	
F	T
0.7	0.3



- Representation: graph
  - Intuitive data structure
- Reasoning
  - Probability theory

E	B	A: F	A: T
F	F	0.99	0.01
F	T	0.1	0.9
T	F	0.3	0.7
T	T	0.01	0.99

+4

A	N: F	N: T
F	0.9	0.1
T	0.2	0.8

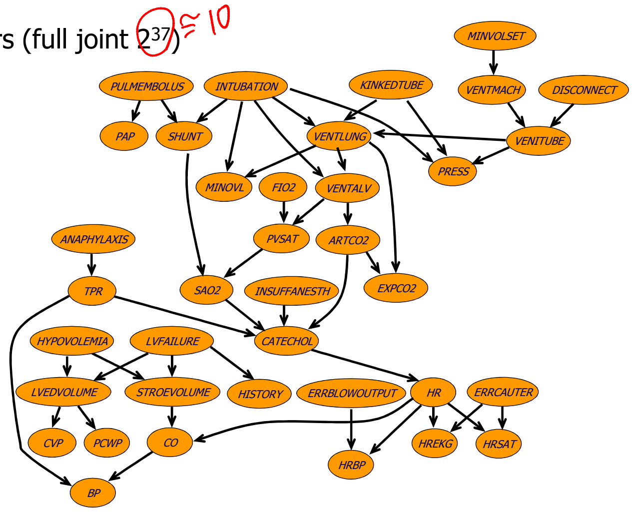
+2

8 independent parameters

15

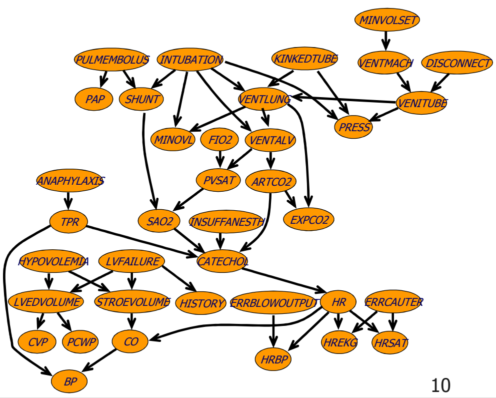
Example Bayesian network

- The "Alarm" network for monitoring intensive care patients
  - 37 variables
  - 509 parameters (full joint 2^{37}) $\cong 10^{11}$



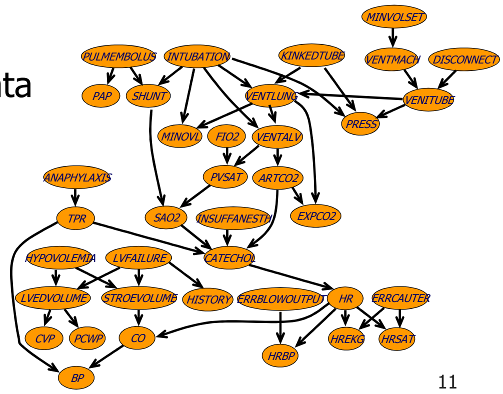
Representation: graphs

- Intuitive data structure for modeling highly-interacting sets of variables
  - Compact representation
  - Explicit model of modularity
- Data structure that allows for design of efficient general-purpose algorithms



Reasoning: probability theories

- Well understood framework for modeling uncertainty
  - Partial knowledge of the state of the world
  - Noisy observations
  - Phenomenon not covered by our model
  - Inherent stochasticity
- Clear semantics
- Can be learned from data



Probabilistic reasoning

- This course covers:
  - Probabilistic graphical model (PGM) representation
    - Bayesian networks (directed graph)
    - Markov networks (undirected graph)
  - Answering queries in PGMs ("inference")
    - What is the probability of X given some observations?
    - What is the most likely explanation for what is happening?
  - Learning PGMs from data ("learning")
    - What are the right/good parameters/structure of the model?
  - Application & special topics
    - Modeling temporal processes with PGMs
      - Hidden Markov Models (HMMs) as a special case
    - Modeling decision-making processes
      - Markov Decision Processes (MDPs) as a special case

Course outline

Week	Topic	Reading
1	Introduction, Bayesian network representation	2.1-3, 3.1
	Bayesian network representation cont.	3.1-3
2	Local probability models	5
	Undirected graphical models	4
3	Exact inference	9.1-4
	Exact inference cont.	10.1-2
4	Approximate inference	12.1-3
	Approximate inference cont.	12.1-3
5	Parameter estimation	17
	Parameter estimation cont.	17
6	Partially observed data (EM algorithm)	19.1-3
	Structure learning BNs	18
7	Structure learning BNs cont.	18
	Partially observed data	19.4-5
8	Learning undirected graphical models	20.1-3
	Learning undirected graphical models cont.	20.1-3
9	Hidden Markov Models	TBD
	HMMs cont. and Kalman filter	TBD
10	Markov decision processes	TBD

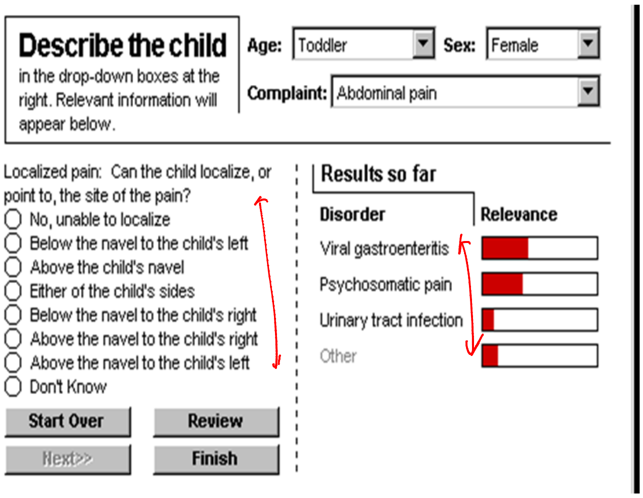
Application: recommendation systems

- Given user preferences, suggest recommendations
- Example: Amazon.com

- Input: movie preferences of many users
- Solution: model correlations between movie features
  - Users that like comedy, often like drama
  - Users that like action, often do not like cartoons
  - Users that like Robert Deniro films often like Al Pacino films
  - Given user preferences, can predict probability that new movies match preferences

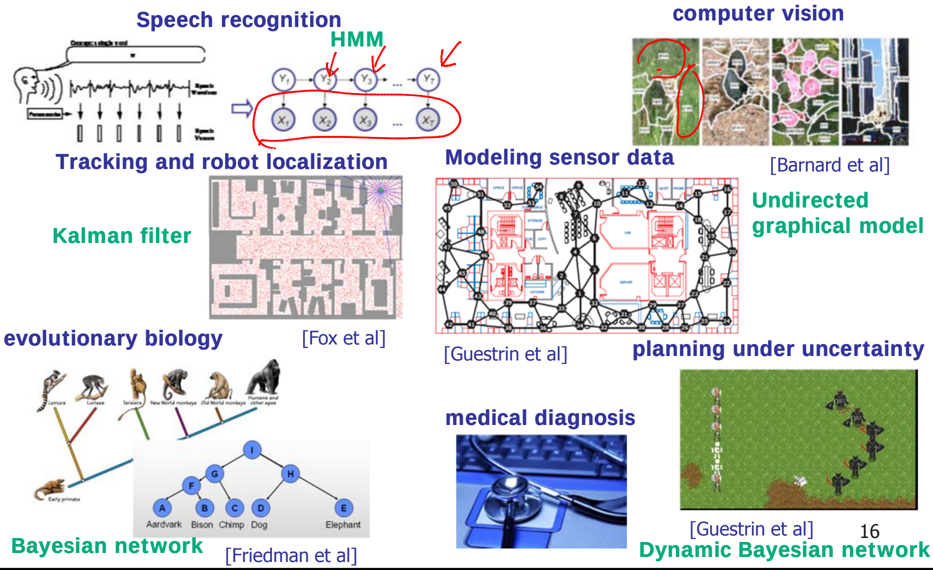
Diagnostic systems

- Diagnostic indexing for home health site at microsoft
- Enter symptoms → recommend multimedia content



Many research areas in CS

- Full of tasks that require reasoning under uncertainty



Enjoy!

- Probabilistic graphical models are having significant impact in science, engineering and beyond
- This class should give you the basic foundation for applying PGMs and developing new methods
- The fun begins …

Today

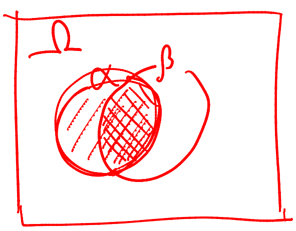
- **Basics of probability**
  - Conditional probabilities
  - Statistical independence
  - Random variable
- Simple Bayesian networks
  - Two nodes make a BN
  - Naïve Bayes
- Should be a review for everyone – Setting up notation for the class

Sample spaces, events and probabilities

- Probability
  - A degree of confidence that an "event" of an uncertain nature will occur.
- Begin with a set Ω -- the sample space
  - Space of possible outcomes
  - e.g. if we consider dice, we might have a set $\Omega=\{1,2,3,4,5,6\}$
  - $\alpha\in\Omega$ is a sample point / atomic event.
- A probability space is a sample space with an assignment $P(\alpha)$ for every ~~$\alpha\in\Omega$~~ s.t.
  - $0\leq P(\alpha)\leq 1$
  - $\sum_{\alpha} P(\alpha)=1$
  - e.g. $P(1)= P(2)= P(3)= P(4)= P(5) = P(6)=1/6$
- An event A is any subset of Ω
  - $P(A) = \sum_{\{\alpha \in A\}} P(\alpha)$
- E.g., $P(\text{die roll}<4) = P(1)+ P(2)+P(3) = 0.5$

Conditional probabilities

- Consider two events α and β,
  - e.g. α = getting admitted to the UW CSE, $p(\alpha)$
    β = getting a job offer from Microsoft. $p(\beta)$
- After learning that α is true, how do we feel about β?
  - $P(\beta|\alpha)$ ~~?~~ $p(\beta)$?



$$p(\beta|\alpha)=\frac{p(\alpha\cap\beta)}{p(\alpha)}$$

Two of the most important rules of the quarter: 1. The chain rule

- From the definition of the conditional distribution, we immediately see that
  - $P(\alpha \cap \beta) = P(\alpha)P(\beta|\alpha)$

$p(\beta|\alpha) = \frac{p(\alpha\cap\beta)}{p(\alpha)}$

- More generally: $\alpha_1, \ldots \alpha_k \quad (k > 2)$
  - $P(\alpha_1 \cap \ldots \cap \alpha_k) = P(\alpha_1) P(\alpha_2|\alpha_1) \cdots P(\alpha_k|\alpha_1 \cap \ldots \cap \alpha_{k-1})$

Two of the most important rules of the quarter: 2. Bayes rule

- Another immediate consequence of the definition of conditional probability is:

$$P(\alpha \mid \beta) = \frac{P(\beta \mid \alpha)P(\alpha)}{P(\beta)}$$

$p(\alpha|\beta) = \frac{p(\alpha\cap\beta)}{p(\beta)} = \frac{p(\alpha)\,p(\beta|\alpha)}{p(\beta)}$

"invert"

- A more general version of Bayes' rule, where all the probabilities are conditioned on some "background" event γ:

$$P(\alpha \mid \beta \cap \gamma) = \frac{P(\beta \mid \alpha \cap \gamma)P(\alpha \mid \gamma)}{P(\beta \mid \gamma)}$$

Most important concept of the quarter: a) Independence

- α and β are independent, if $P(\beta|\alpha)=P(\beta)$
  - Denoted $P \models (\alpha \perp \beta)$

- **Proposition:** α and β are independent if and only if $P(\alpha\cap\beta)=P(\alpha)P(\beta)$

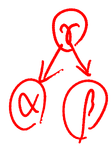
$$P(\alpha\cap\beta) = P(\alpha)\cdot P(\beta|\alpha) = P(\alpha)P(\beta)$$

Most important concept of the quarter: b) Conditional independence

- Independence is rarely true, but conditionally…

- α and β conditionally independent given γ if $P(\beta|\alpha\cap\gamma)=P(\beta|\gamma)$
  - $P \models (\alpha \perp \beta \mid \gamma)$

α: UW CSE
β: MS
γ: Smart, diligent, letter, goal



Proposition: $P \models (\alpha \perp \beta \mid \gamma)$ if and only if $P(\alpha\cap\beta \mid\gamma)=P(\alpha \mid\gamma)P(\beta \mid\gamma)$

Random variables

- Probability distributions are defined for events
- Events are complicated – so, let's think about attributes
  - Age, Grade, HairColor

Grade A

Job Offer

- A random variable (such as *Grade*), is defined by a function that associates each outcome in Ω (each person) with a value.
  - Grade = A – shorthand for event $\{w \in \Omega : f_{Grade}(w) = A\}$
  - Grade = B – shorthand for event $\{w \in \Omega : f_{Grade}(w) = B\}$

    :

- Properties of a random variable X:
  - Val(X) = a set of possible values of random variable X
  - For discrete (categorical): $\sum_{i=1,\ldots,|Val(X)|} P(X=x_i) = 1$
  - $P(x) \geq 0$

X: Grade

Val(X) = {A, B, C, .. F}

$P(X=A)$

Basic concepts for random variables

- Atomic event: assignment $x_1,\ldots,x_n$ to $X_1,\ldots,X_n$

A Grade

- Conditional probability: $\boxed{P(Y|X)=P(X,Y)/P(X)}$
  - For all values $x \in Val(X)$, $y \in Val(Y)$

$P(Y|X) = \frac{P(X,Y)}{P(X)}$

- Bayes rule: $P(X|Y) = \frac{P(X)P(Y|X)}{P(Y)}$

$P(\alpha \cap \beta)$

- Chain rule:
  - $P(X_1,\ldots,X_n) =$

$P(X,Y)$

$P(Grade \mid Intelligence)$

$P(Grade{=}A \mid Intell{=}H)$

$P(Grade{=}A \mid Intell{=}L)$

$P(G=a, I=h) = \cancel{0.3}\ 0.5$

$P(X_1)P(X_2|X_1)$ .. $P(X_n|X_1 \cdots X_{n-1})$

$P(\alpha \cap \beta)$

Joint distribution, marginalization

- Two random variables – Grade & Intelligence

$G \in \{a, b\}$, $I \in \{vh, h\}$ $\qquad P(G, I)$

G \ I	vh	h
a	0.7	0.1
b	0.15	0.05

$P(G=a, I=h)$ $\qquad P(G=a, I=vh)$

- Marginalization – Compute marginal over single variable

$P(G, I) \to P(G)$ or $P(I)$

$P(G=b) = P(G=b, I=vh) + P(G=b, I=h)$

$P(G=a) = 1 - P(G=b)$

Ω $\quad I=h \quad I=vh$

Marginalization – the general case

- Compute marginal distribution $P(X_i)$ from joint distribution $P(X_1,\ldots,X_i,\ldots,X_n)$:

$G \qquad I$

$$P(X_1, X_2, \ldots, X_i) = \sum_{x_{i+1},\ldots,x_n} P(X_1, X_2, \ldots, X_i, x_{i+1}, \ldots, x_n)$$

$I \qquad G$

$$P(X_i) = \sum_{x_1,\ldots,x_{i-1}} P(x_1, \ldots, x_{i-1}, X_i)$$

$\sum_{x_{i+1}} \sum_{x_{i+2}} \ldots \sum_{x_n}$

2^{n-1}

Today

- Basics of probability
  - Conditional probabilities
  - Statistical independence
  - Random variable
- **Two nodes make a BN**
- Naïve Bayes

- Should be a review for everyone – Setting up notation for the class

Representing joint distributions

- Random variables: $X_1,\ldots,X_n$
- P is a joint distribution over $X_1,\ldots,X_n$

$p(x_1, \cdots x_n)$

$x_i \in \{T, F\}$

If $X_1,..,X_n$ binary, need 2^n parameters to describe P

Can we represent P more compactly?
- Key: Exploit independence properties

Independent random variables

- If $X_1,\ldots,X_n$ are independent then:
  - $P(X_1,\ldots,X_n) = P(X_1)\ldots P(X_n)$
  - $O(n)$ parameters
  - All 2^n probabilities are implicitly defined
  - Cannot represent many types of distributions

$X_i \in \{T,F\}$

X_1, X_2

$P(X_2) = P(X_2|X_1)$

$P(X_1,X_2) = P(X_1)P(X_2|X_1)$

$= P(X_1)\cdot P(X_2)$

- X and Y are conditionally independent given Z if
  - $P(X=x|Y=y, Z=z) = P(X=x|Z=z)$ for all values x, y, z
  - Equivalently, if we know Z, then knowing Y does not change predictions of X
  - Notation: $(X \perp Y \mid Z)$

$X \perp Y | Z$

Conditional parameterization

- S = SAT score, Val(S) = $\{s^0,s^1\}$
- I = Intelligence, Val(I) = $\{i^0,i^1\}$

$P(I,S) = P(I)P(S|I)$

P(I,S)

I	*S*	*P(I,S)*
i^0	s^0	0.665
i^0	s^1	0.035
i^1	s^0	0.06
i^1	s^1	0.24

Joint parameterization

3 parameters

P(I)

i^0	i^1
0.7	0.3

P(S|I)

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

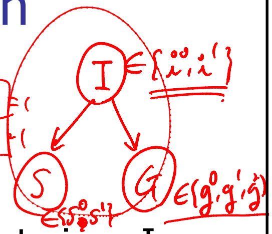
Conditional parameterization

3 parameters

Alternative parameterization: P(S) and P(I|S)

Conditional parameterization

- S = SAT score, Val(S) = $\{s^0,s^1\}$
- I = Intelligence, Val(I) = $\{i^0,i^1\}$
- G = Grade, Val(G) = $\{g^0,g^1,g^2\}$
- Assume that G and S are independent given I



⬇

- Joint parameterization
  - 2·2·3=12-1=11 independent parameters
- Conditional parameterization has
  - P(I,S,G) = P(I)P(S|I)P(G|I,S) = P(I)P(S|I)P(G|I)
  - P(I) – 1 independent parameter
  - P(S|I) – 2·1 independent parameters
  - P(G|I) - 2·2 independent parameters
  - ~~7 independent parameters~~

$P(S,I,G)$

$G \perp S \mid I$

$P(G|I,S) = P(G|I)$

$P(I,S,G) = P(I)P(S|I)P(G|I,S) = P(I)P(S|I)P(G|I)$

+1 +2 +4

Naïve Bayes model

- Class variable C, Val(C) = $\{c_1,\ldots,c_k\}$
- Evidence variables $X_1,\ldots,X_n$
- Naïve Bayes assumption: evidence variables are conditionally independent given C

⬇

$$P(C, X_1,\ldots, X_n) = P(C)\prod_{i=1}^{n} P(X_i \mid C)$$

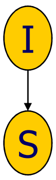
- Applications in medical diagnosis, text classification
- Used as a classifier:

$$\frac{P(C=c_1 \mid x_1,\ldots,x_n)}{P(C=c_2 \mid x_1,\ldots,x_n)} = \frac{P(C=c_1)}{P(C=c_2)}\prod_{i=1}^{n}\frac{P(x_i \mid C=c_1)}{P(x_i \mid C=c_2)}$$

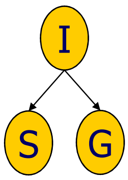
- Problem: Double counting correlated evidence

Bayesian network (informal)

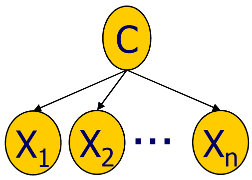
- Directed acyclic graph G
  - Nodes represent random variables
  - Edges represent direct influences between random variables
- Local probability models



Example 1



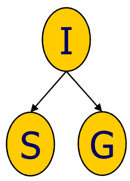
Example 2



Naïve Bayes

Bayesian network (informal)

- Represent a joint distribution
  - Specifies the probability for P(**X**=**x**)
  - Specifies the probability for P(**X**=**x**|**E**=**e**)
- Allows for reasoning patterns
  - Prediction (e.g., intelligent → high scores)
  - Explanation (e.g., low score → not intelligent)
  - Explaining away (different causes for same effect interact)

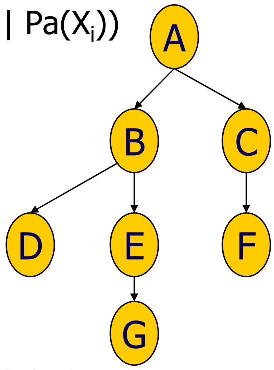


Example 2

Bayesian network structure

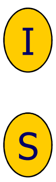
- Directed acyclic graph G
  - Nodes $X_1,\ldots,X_n$ represent random variables
- G encodes local Markov assumptions
  - X_i is independent of its non-descendants given its parents
  - Formally: $(X_i \perp NonDesc(X_i) \mid Pa(X_i))$

$E \perp \{A,C,D,F\} \mid B$ ⬅



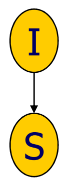
Independency mappings (I-maps)

- Let P be a distribution over **X**
- Let I(P) be the independencies $(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z})$ in P
- A Bayesian network structure is an I-map (independency mapping) of P if $I(G) \subseteq I(P)$



I	*S*	*P(I,S)*
i^0	s^0	0.25
i^0	s^1	0.25
i^1	s^0	0.25
i^1	s^1	0.25

I	*S*	*P(I,S)*
i^0	s^0	0.4
i^0	s^1	0.3
i^1	s^0	0.2
i^1	s^1	0.1



$I(G)=\{I \perp S\}$ $I(P)=\{I \perp S\}$ $I(P)=\varnothing$ $I(G)=\varnothing$

Factorization Theorem

- If G is an I-Map of P, then P factorizes over G.

$$P(X_1,\ldots,X_n) = \prod_{i=1}^{n} P(X_i \mid Pa(X_i))$$

Proof:

- wlog. (without loss of generality)
  $X_1,\ldots,X_n$ is an ordering consistent with G
- By chain rule: $P(X_1,\ldots,X_n) = \prod_{i=1}^{n} P(X_i \mid X_1,\ldots,X_{i-1})$
- From assumption: $Pa(X_i) \subseteq \{X_1,\ldots,X_{i-1}\}$

$$\{X_1,\ldots,X_{i-1}\} - Pa(X_i) \subseteq NonDesc(X_i)$$

- Since G is an I-Map → (X_i; NonDesc(X_i)| Pa(X_i))$\in$I(P)

⬇

$$P(X_i \mid X_1,\ldots,X_{i-1}) = P(X_i \mid Pa(X_i))$$

Factorization implies I-Map

- $P(X_1,\ldots,X_n) = \prod_{i=1}^{n} P(X_i \mid Pa(X_i))$ → G is an I-Map of P

Proof:

- Need to show (X_i; NonDesc(X_i)| Pa(X_i))$\in$I(P) or that
  P(X_i | NonDesc(X_i)) = P(X_i | Pa(X_i))
- wlog. $X_1,\ldots,X_n$ is an ordering consistent with G
-

$$\begin{aligned} P(X_i \mid NonDesc(X_i)) &= \frac{P(X_i, NonDesc(X_i))}{P(NonDesc(X_i))} \\ &= \frac{\prod_{k=1}^{i} P(X_k \mid Pa(X_k))}{\prod_{k=1}^{i-1} P(X_k \mid Pa(X_k))} \\ &= P(X_i \mid Pa(X_i)) \end{aligned}$$

Bayesian network definition

- A Bayesian network is a pair (G,P)
  - P factorizes over G
  - P is specified as set of CPDs associated with G's nodes (and its parents)
- Parameters
  - Joint distribution: 2^n
  - Bayesian network (bounded in-degree k): $n2^k$

Today and next class

- Next class
  - Details on semantics of BNs, relate them to independence assumptions encoded by the graph.
- Today's To-Do List
  - Visit the course website.
  - Reading K&F 2.1-3, 3.1.

Acknowledgement

- These lecture notes were generated based on the slides from Profs Eran Segal and Carlos Guestrin.