Readings:
K&F: 3.3, 3.4, 16.1, 16.2, 16.3, 16.4

Learning P-maps
Param. Learning

Graphical Models – 10708
Carlos Guestrin
Carnegie Mellon University
September 24th, 2008

Perfect maps (P-maps)

- I-maps are not unique and often not simple enough
- Define "simplest" *G* that is I-map for *P*
  - A BN structure *G* is a **perfect map** for a distribution *P* if I(*P*) = I(*G*)
- Our goal:
  - Find a perfect map!
  - Must address equivalent BNs

Inexistence of P-maps 1

- XOR (this is a hint for the homework)

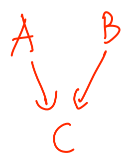
$A = B \text{ xor } C$

$A \perp B$
$B \perp C$
$C \perp A$

$\neg A \perp B \mid C$
$\neg A \perp C \mid B$
$\neg B \perp C \mid A$

P-MAP?
extra credit

I map :
Minimal
$A \perp B$



not P-map
can't read
$B \perp C$
$A \perp C$

Obtaining a P-map

- Given the independence assertions that are true for *P*
- Assume that there exists a perfect map G^*
  - Want to find G^*
- Many structures may encode same independencies as G^*, when are we done?
  - Find all equivalent structures simultaneously!

I-Equivalence

- Two graphs G_1 and G_2 are **I-equivalent** if $I(G_1) = I(G_2)$
- **Equivalence class** of BN structures
  - Mutually-exclusive and exhaustive partition of graphs

X→Y→Z
X←Y←Z
X←Y→Z

…

…

X→Y

Y→X

⋮

- How do we characterize these equivalence classes?

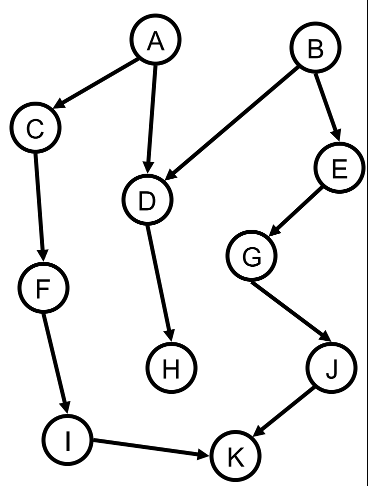
Skeleton of a BN

- **Skeleton** of a BN structure G is an **undirected graph** over the same variables that has an edge X–Y for every X→Y or Y→X in G

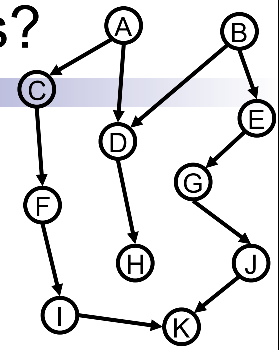
Skeleton

- (Little) **Lemma:** Two I-equivalent BN structures must have the same skeleton

Counter example



What about V-structures?



- **V-structures are key property of BN structure**

  $X \rightarrow Z \leftarrow Y$

- **Theorem:** If G_1 and G_2 have the same skeleton and V-structures, then G_1 and G_2 are I-equivalent

  not if and only if

Same V-structures not necessary

- **Theorem:** If G_1 and G_2 have the same skeleton and V-structures, then G_1 and G_2 are I-equivalent
- Though sufficient, same V-structures not necessary

$I(G_1)=\emptyset$ $\quad X \xrightarrow{G_1} Z$ $\quad$ I-equivalent $\quad$ $X \xleftarrow{G_2} Z$ $\quad I(G_2)=\emptyset$

$W \perp XZ \mid Y$ $\quad Y$ $\quad$ not same V-structures $\quad Y$ $\quad W \perp XZ \mid Y$

W $\quad W$

Immoralities & I-Equivalence

immoral

moral

immorality ⇐

X Z not married

- Key concept not V-structures, but "immoralities" (unmarried parents ☺)
  - $X \rightarrow Z \leftarrow Y$, with no arrow between X and Y
  - Important pattern: X and Y independent given their parents, but not given Z
  - (If edge exists between X and Y, we have *covered* the V-structure)
- **Theorem:** G_1 and G_2 have the same skeleton and immoralities if and only if G_1 and G_2 are I-equivalent

Obtaining a P-map

- Given the independence assertions that are true for *P*
  - Obtain skeleton
  - Obtain immoralities
- From skeleton and immoralities, obtain every (and any) BN structure from the equivalence class

Measuring Independence

- many ways → oracle → estimate from data ← not always easy, but doable
- A very simple approach (related to MLE)

$(X \perp Y \mid Z) \Leftrightarrow I(X,Y \mid Z) = 0$

$$I(X,Y \mid Z) = \sum_{xyz} P(x,y,z) \log \frac{P(x,y \mid z)}{P(x \mid z)\,P(y \mid z)}$$

data: don't have $P(x,y,z)$, estimate MLE $\hat{P}(x,y,z) \overset{MLE}{=} \frac{Count(x,y,z)}{m}$ ← # data points

in practice: $I(X,Y \mid Z) > 0$

independent "enough" when $I(X,Y \mid Z) < \varepsilon$

Identifying the skeleton 1

- When is there an edge between X and Y?

is it when $\neg X \perp Y$? NO: $X \to Z \to Y$ $\neg X \perp Y$ but no edge $X - Y$

$\neg X \perp Y \mid$ everything else? $X \to Z \leftarrow Y$

- When is there no edge between X and Y?

X is not a parent of Y & vice-versa

Local Markov assumption

$\exists\, U \subseteq \mathcal{X} - \{X,Y\}$, such that $X \perp Y \mid U$

additional assumption # parents $\leq d$

$\exists\, U \subseteq \mathcal{X} - \{X,Y\}, |U| \leq d$, $X \perp Y \mid U$

Identifying the skeleton 2

- Assume d is max number of parents (d could be n)
- For each X_i and X_j $\sum_{k=0}^{d} \binom{n-2}{k}$
  - $E_{ij} \leftarrow$ true
  - For each $\mathbf{U} \subseteq \mathbf{X} - \{X_i, X_j\}$, $|\mathbf{U}| \leq d$ U could be $\emptyset$
    - Is $(X_i \perp X_j \mid \mathbf{U})$?
      - $E_{ij} \leftarrow$ false ✓ break
  - If E_{ij} is true
    - Add edge X – Y to skeleton

Identifying immoralities

$X \to U \to Y$, $\neg X \perp Y$ but still immoral

- Consider X – Z – Y in skeleton, when should it be an immorality? when $X \perp Y$ (and $\neg X \perp Y \mid Z$ but no need to test in this simple case)

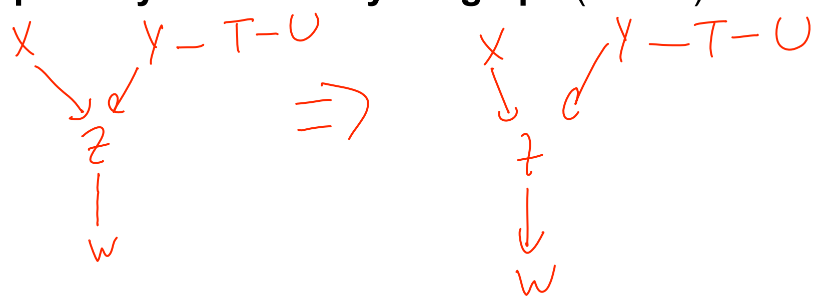
  $X \to Z \to Y$ then there must a way to make $X \perp Y \mid Z$ since Z is a parent of Y
- Must be $X \rightarrow Z \leftarrow Y$ (immorality):
  - When X and Y are **never independent** given **U**, if $Z \in \mathbf{U}$

  $\nexists\, U \subseteq X - \{X, Y\}, Z \in U, X \perp Y \mid U$ (if I have at most d parents $|U| \leq d$)
- Must **not** be $X \rightarrow Z \leftarrow Y$ (not immorality):
  - When there exists **U** with $Z \in \mathbf{U}$, such that X and Y are **independent** given **U**

  $\exists\, U \subseteq X - \{X, Y\}, Z \in U, X \perp Y \mid U$

From immoralities and skeleton to BN structures

- Representing BN equivalence class as a **partially-directed acyclic graph** (PDAG)



- Immoralities force direction on some other BN edges
- Full (polynomial-time) procedure described in reading

What you need to know

- Minimal I-map
  - every P has one, but usually many
- Perfect map
  - better choice for BN structure
  - not every P has one
  - can find one (if it exists) by considering I-equivalence
  - Two structures are I-equivalent if they have same skeleton and immoralities

Announcements

- Recitation tomorrow ✓
  - Don't miss it!
- ~~No class on Monday ☹~~

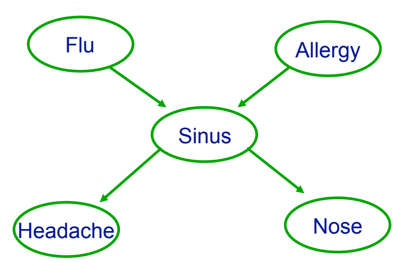
Everything so far Chapter 3

↑ your HW

Now → Chapter 16

Review

- Bayesian Networks
  - Compact representation for probability distributions
  - Exponential reduction in number of parameters
  - Exploits independencies



- Next – Learn BNs
  - parameters ✓ ←
  - structure ✓

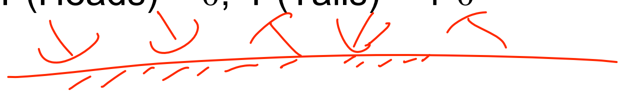
Thumbtack – Binomial Distribution

- P(Heads) = θ, P(Tails) = 1-θ

$\theta = \frac{3}{5}$

why??

MLE



- Flips are i.i.d.:
  - Independent events
  - Identically distributed according to Binomial distribution
- Sequence *D* of α_H Heads and α_T Tails

$$P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

Maximum Likelihood Estimation

- **Data:** Observed set *D* of α_H Heads and α_T Tails
- **Hypothesis:** Binomial distribution
- Learning θ is an optimization problem
  - What's the objective function? MLE

$\max_\theta \; P(D|\theta)$

- MLE: Choose θ that maximizes the probability of observed data:

$$\hat{\theta}_{MLE} = \arg\max_\theta \; P(\mathcal{D} \mid \theta)$$

$$= \arg\max_\theta \; \ln P(\mathcal{D} \mid \theta)$$

Your first learning algorithm

$\ln a^b = b \ln a$

$$\widehat{\theta} = \arg\max_{\theta} \ \ln P(\mathcal{D} \mid \theta)$$
$$= \arg\max_{\theta} \ \ln \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$
$$= \arg\max_{\theta} \ \alpha_H \ln \theta + \alpha_T \ln(1-\theta)$$

$\frac{\partial}{\partial \theta} \ln \theta = \frac{1}{\theta}$

$\frac{\partial}{\partial \theta} \ln(1-\theta) = \frac{-1}{1-\theta}$

- Set derivative to zero: $\boxed{\frac{d}{d\theta} \ln P(\mathcal{D} \mid \theta) = 0}$

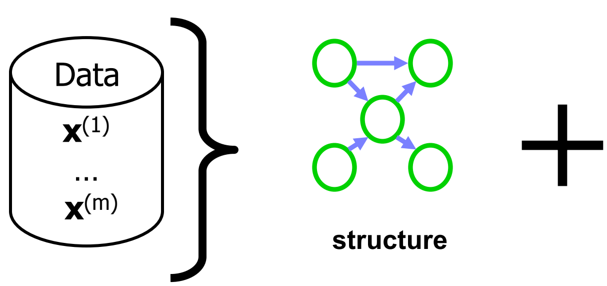
$$\alpha_H \frac{1}{\theta} - \alpha_T \frac{1}{1-\theta} = 0$$

$$\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

one binary node in this BN

Learning Bayes nets

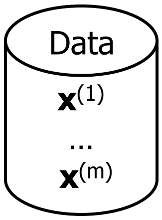
	Known structure	Unknown structure
Fully observable data	easy 1st	hard 2nd structure learning
Missing data	hard (EM) 3rd	very hard later in 4th semester



CPTs – $P(X_i \mid \mathbf{Pa}_{X_i})$

parameters

Learning the CPTs

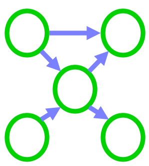


For each discrete variable X_i $\quad Pa_{X_i} = U$

$P(X_i \mid Pa_{X_i}) = P(X_i \mid U)$

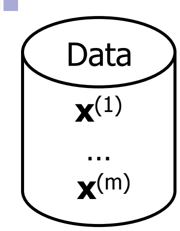
$$\hat{P}_{MLE}(X_i = x_i \mid U = u) = \frac{\text{Count}(X_i = x_i, U = u)}{\text{Count}(U = u)}$$

Why??



MLE: $$P(X_i = x_i \mid X_j = x_j) = \frac{\text{Count}(X_i = x_i, X_j = x_j)}{\text{Count}(X_j = x_j)}$$

Learning the CPTs

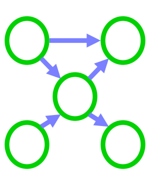


For each discrete variable X_i

MLE: $$P(X_i = x_i \mid X_j = x_j) = \frac{\text{Count}(X_i = x_i, X_j = x_j)}{\text{Count}(X_j = x_j)}$$

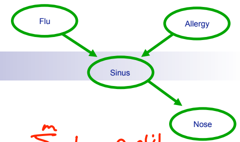
WHY??????????

if only one var
then take derivative, set to $\emptyset$
all is good



Maximum likelihood estimation (MLE) of BN parameters – example

$\log a \cdot b = \log a + \log b$



- Given structure, log likelihood of data:

$$\log P(\mathcal{D} \mid \theta_{\mathcal{G}}, \mathcal{G}) = \log \prod_{j=1}^{m} P(x^{(j)} \mid \theta_G, G) = \sum_{j=1}^{m} \log P(x^{(j)} \mid \theta_G, G)$$

for the example

$$\sum_{j=1}^{m} \log P(f^{(j)}, a^{(j)}, s^{(j)}, n^{(j)} \mid \theta_G, G) = \sum_{j=1}^{m} \log P(f^{(j)} \mid \theta_G, G) \cdot P(a^{(j)} \mid \theta_G, G)\, P(s^{(j)} \mid a^{(j)}, f^{(j)}, \theta_G, G)\, P(n^{(j)} \mid s^{(j)}, \theta_G, G)$$

$$= \sum_{j=1}^{m} \left[\log P(f^{(j)} \mid \theta_G, G) + \log P(a^{(j)} \mid \theta_G, G) + \log P(s^{(j)} \mid a^{(j)}, f^{(j)}, \theta_G, G) + \log P(n^{(j)} \mid s^{(j)}, \theta_G, G) \right]$$

$$= \underbrace{\sum_{j=1}^{m} \log P(f^{(j)} \mid \theta_F, G)} + \underbrace{\sum_{j=1}^{m} \log P(a^{(j)} \mid \theta_A, G)} + \underbrace{\sum_{j=1}^{m} \log P(s^{(j)} \mid a^{(j)}, f^{(j)}, \theta_{S|FA}, G)} + \underbrace{\sum_{j=1}^{m} \log P(n^{(j)} \mid s^{(j)}, \theta_{N|S}, G)}$$

one for each CPT

Broke up problem into independent subproblems :