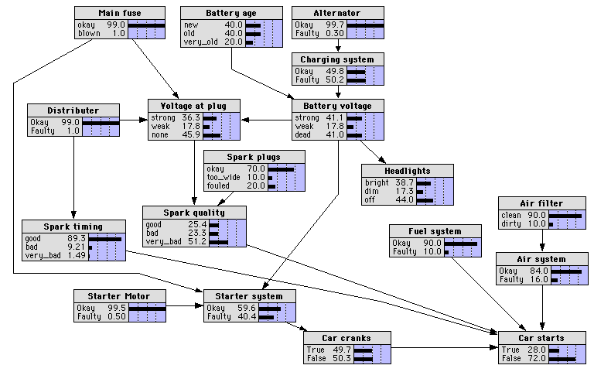
Cmput 466 / 551



Learning Belief Net Structures

Readings: HTF ~Ch17

+ Bayesian Networks without the Tears (Charniak)

R Greiner

University of Alberta

Some material taken from C Guesterin (CMU), K Murphy (UBC)

Outline

- Motivation
- What is a Belief Net?
- Learning a Belief Net
  - Goal?
  - Learning Parameters – Complete Data
  - Learning Parameters – Incomplete Data
  - Learning Structure
    - Learning best TREE Structure
- Dynamical Belief Nets … HMMs

Learning Belief Nets

Jump

Data	Structure: Known	Structure: Unknown
Complete	Easy	NP-hard
Missing	Hard … EM	Very hard!!

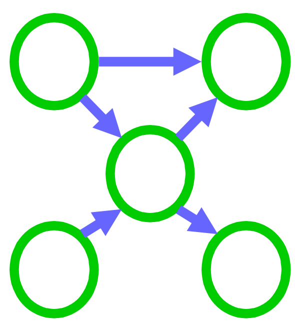
Data

$\mathbf{x}^{(1)}$

…

$\mathbf{x}^{(m)}$

$\Rightarrow$



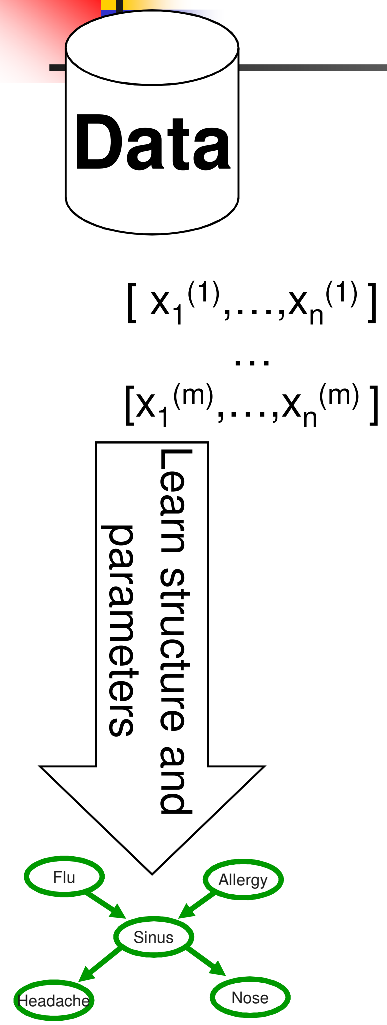
structure

$+$ CPTs :

$\{ P(X_i | \mathbf{Pa}_{X_i}) \}$

parameters

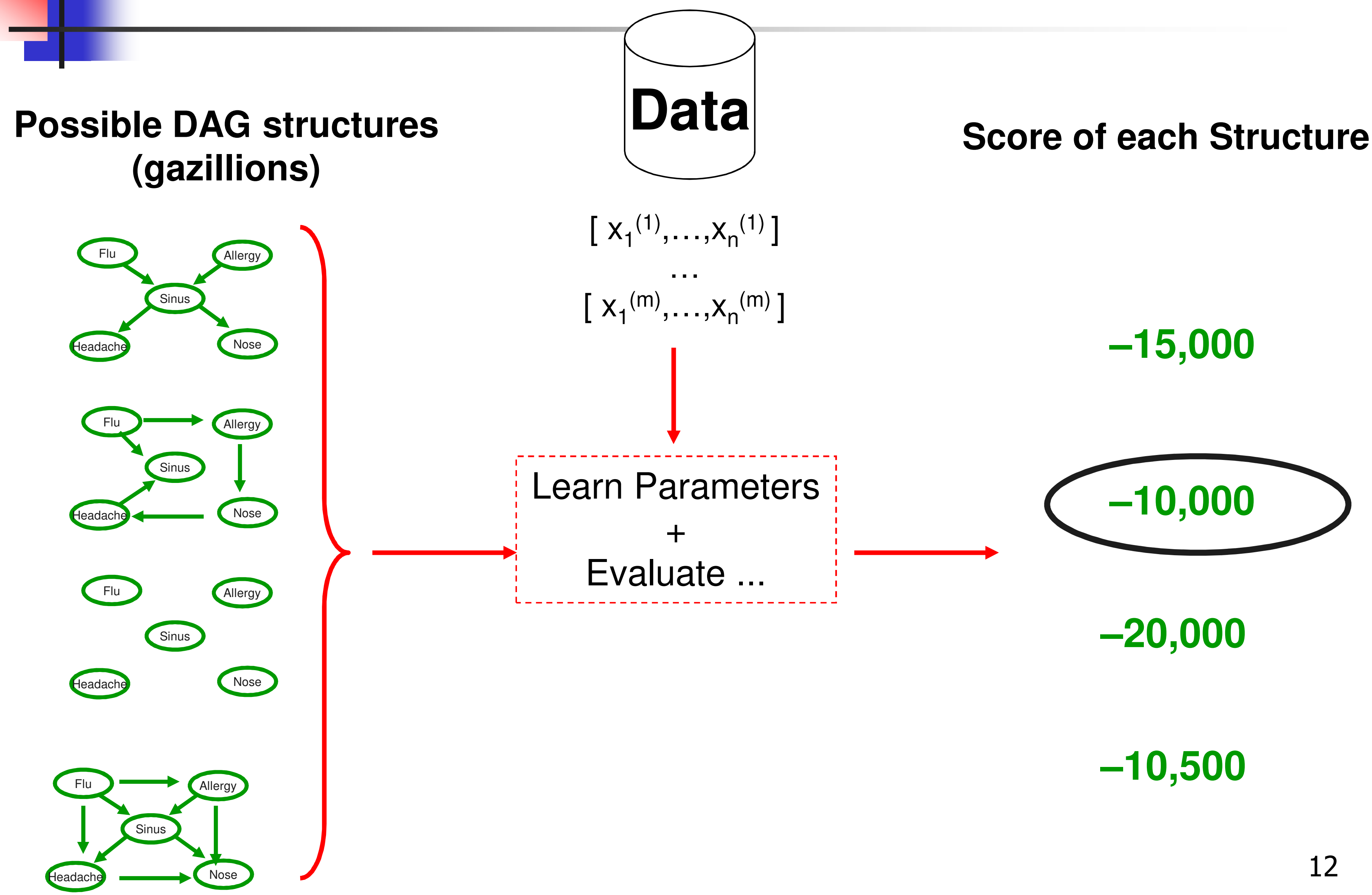
Learning the structure of a BN



Take Cmput659…

- **Constraint-based approach**
  - BN encodes conditional independencies
  - Test conditional independencies in data
  - Find an I-map (?P-map?)
    - Only include link if dependency (given other links)
- **Score-based approach**
  - Finding structure + parameters = *density estimation*
  - Evaluate model as we evaluated parameters
    - Maximum likelihood
    - Bayesian
    - etc.

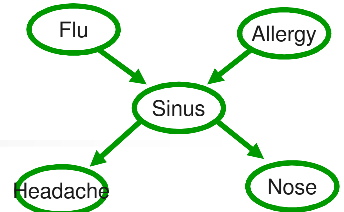
Score-based Approach



Just use MLE parameters

- $\max_{\mathcal{G},\, \theta_{\mathcal{G}}} L(\langle \mathcal{G}, \theta_{\mathcal{G}} \rangle : \mathcal{D}) =$
  $\max_{\mathcal{G}} \boxed{\max_{\theta_{\mathcal{G}}} L(\langle \mathcal{G}, \theta_{\mathcal{G}} \rangle : \mathcal{D})} =$
  $\max_{\mathcal{G}} L(\langle \mathcal{G}, \boxed{\theta^*(\mathcal{G})} \rangle : \mathcal{D})$
  - $\theta^*(\mathcal{G}) = \max_{\theta_{\mathcal{G}}} L(\langle \mathcal{G}, \theta_{\mathcal{G}} \rangle : \mathcal{D})$
- So…
  seek the structure $\mathcal{G}$ that achieves highest likelihood,
  given its MLE parameters $\theta^*(\mathcal{G})$
- $\mathrm{Score}(\mathcal{G}, \mathcal{D}) = \log L(\langle \mathcal{G}, \theta^*(\mathcal{G}) \rangle : \mathcal{D})$

Information-theoretic interpretation of maximum likelihood



- Given structure $\mathcal{G}$, parameters $\theta_{\mathcal{G}}$, log likelihood of data $\mathcal{D}$:

$$\log P(\mathcal{D} \mid \theta_{\mathcal{G}}, \mathcal{G}) = \sum_{j=1}^{m} \sum_{i=1}^{n} \log P\left(X_i = x_i^{(j)} \mid \mathbf{Pa}_{X_i} = \mathbf{x}^{(j)} \right)$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{m} \log P\left(X_i = x_i^{(j)} \mid \mathbf{Pa}_{X_i} = \mathbf{x}^{(j)} \right)$$

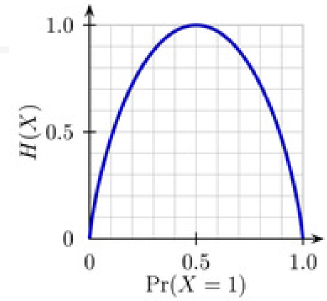
$$= \sum_{i=1}^{n} \sum_{x_i, \mathbf{u}} \overbrace{\#(X_i = x_i, \mathbf{Pa}_{X_i} = u)}^{N_{ijk}} \log \overbrace{P\left(X_i = x_i \mid \mathbf{Pa}_{X_i} = \mathbf{u}\right)}^{\theta_{ijk}}$$

$$= m \sum_{i=1}^{n} \sum_{x_i, \mathbf{u}} \boxed{\frac{\#(X_i = x_i, \ \mathbf{Pa}_{X_i} = u)}{m}} \log P\left(X_i = x_i \mid \mathbf{Pa}_{X_i} = \mathbf{u}\right)$$

$$\boxed{\hat{P}(X_i = x_i, \mathbf{Pa}_{X_i} = u)}$$

$$= m \boxed{\sum_{i=1}^{n} \sum_{x_i, \mathbf{u}} \hat{P}(X_i = x_i, \ \mathbf{Pa}_{X_i} = \mathbf{u}) \log P\left(X_i = x_i \mid \mathbf{Pa}_{X_i} = \mathbf{u}\right)}$$

Entropy & Conditional Entropy



- Entropy of Distribution
  - $H(X) = -\sum_i P(x_i) \log P(x_i)$
  - "How `surprising' variable is"
  - Entropy = 0 when know everything… eg P(+x)=1.0
- Conditional Entropy $H(X \mid \mathbf{U})$
  - $H(X|\mathbf{U}) = -\sum_{\mathbf{u}} P(\mathbf{u}) \sum_i P(x_i|\mathbf{u}) \log P(x_i|\mathbf{u})$
  - How much uncertainty is left in X, after observing **U**

$$H(X_i \mid \mathbf{Pa}_{X_i}) = -\sum_{x_i, \mathrm{u}} \hat{P}(X_i = x_i, \mathrm{Pa}_{X_i} = \mathrm{u}) \log P\left(X_i = x_i^{(j)} \mid \mathbf{Pa}_{X_i} = \mathrm{u}\right)$$

Information-theoretic interpretation of maximum likelihood … 2

- Given structure $\mathcal{G}$, parameters $\theta_{\mathcal{G}}$, log likelihood of data $\mathcal{D}$ is…

$$
\begin{aligned}
\log \hat{P}(\mathcal{D} \mid \theta, \mathcal{G}) &= m \sum_i \sum_{x_i, \mathbf{u}} \hat{P}(x_i, \mathbf{Pa}_{x_i,\mathcal{G}} = \mathbf{u}) \log \hat{P}(x_i \mid \mathbf{Pa}_{x_i,\mathcal{G}} = \mathbf{u}) \\
&= m \sum_i -\hat{H}(X_i | \mathbf{Pa}_{x_i,\mathcal{G}}\ ) \\
&= -m \sum_i \boxed{\hat{H}(X_i | \mathbf{Pa}_{x_i,\mathcal{G}}\ )}
\end{aligned}
$$

So $\log P(\mathcal{D} \mid \theta, \mathcal{G})$ is LARGEST

when each $H(X_i \mid \mathbf{Pa}_{X_i,\mathcal{G}})$ is SMALL…

…ie, when parents of X_i are very INFORMATIVE about X_i !

Easier Form

- Note
  $$H_P(\, X \mid \mathbf{U}\,) = \sum_{x,\mathbf{u}} P(x, \mathbf{u}) \log P(x \mid \mathbf{u})$$
  is not symmetric in X, $\mathbf{U}$
- Better to use symmetric
  $$\mathbf{I}_{\boldsymbol{P}}(X, \mathbf{U}) \;=\; \sum_{x,\mathbf{u}} P(x, \mathbf{u}) \log \frac{P(x,\mathbf{u})}{P(x)\, P(\mathbf{u})}$$

(Conditional) Mutual Information

- **Mutual information:** $\mathbf{I}_{\boldsymbol{P}}(\mathrm{X},\mathrm{U}) = \sum_{\mathrm{x,y}} \mathrm{P}(\mathrm{x},\mathrm{u}) \log \frac{\mathrm{P(x,u)}}{\mathrm{P(x)\,P(u)}}$
- Mutual information and independence:
  - X and U independent if and only if I(X,U)=0
  - $X \perp U \quad \Leftrightarrow \quad P(x,u) = P(x)\,P(u) \Leftrightarrow \log[\,P(x,u)/P(x)P(u)\,] = 0$
- **Conditional mutual information:**

$$\mathrm{I_P}(\mathrm{X},\mathrm{Y}\,|\mathrm{Z}) = \mathrm{E_Z}[\,\mathrm{I}(\mathrm{X},\mathrm{Y}\,|\mathrm{Z}=\mathrm{z})] = \sum_{z}\sum_{\mathrm{x,y}} \mathrm{P}(\mathrm{x},\mathrm{y}|z)\log\frac{\mathrm{P}(\mathrm{x},\mathrm{y}\,|\mathrm{z})}{\mathrm{P}(\mathrm{x}\,|\mathrm{z})\,\mathrm{P}(\mathrm{y}|\mathrm{z})}$$

  - $X \perp Y \mid Z$ *iff* $P(X,Y|Z) = P(X|Z)\; P(Y|Z)$ *iff* $I(X,Y|Z) = 0$
- Using the data *D*
  - Empirical distribution: $\widehat{\boldsymbol{P}}(\boldsymbol{x},\boldsymbol{y}) = \frac{\mathbf{Count}_{\mathrm{D}}(\boldsymbol{X}=\boldsymbol{x},\boldsymbol{Y}=\boldsymbol{y})}{|\boldsymbol{D}|}$
  - Mutual information: $\mathrm{I}_{\widehat{\mathrm{P}}}(\mathrm{X},\mathrm{Y}) = \sum_{\mathrm{x,y}} \widehat{\boldsymbol{P}}(\mathrm{x},\mathrm{y})\log\frac{\widehat{\boldsymbol{P}}(\mathrm{x,y})}{\widehat{\boldsymbol{P}}(\mathrm{x})\widehat{\boldsymbol{P}}(\mathrm{y})}$

Mutual Information

- **Mutual information:** $\mathbf{I}_P(X, U) = \sum_{x,u} P(x,u) \log \frac{P(x,u)}{P(x)\,P(u)}$

- Mutual information and independence:
  - X and **U** independent if and only if $I(X, \mathbf{U}) = 0$
  - $X \perp \mathbf{U} \quad \Leftrightarrow \quad \forall x, \mathbf{u}\; P(x, \mathbf{u}) = P(x)\, P(\mathbf{u}) \quad \Leftrightarrow \quad \forall x, \mathbf{u} \quad \log[\, \frac{P(x,\mathbf{u})}{P(x)P(\mathbf{u})} \,] = 0$

- $\mathbf{I}_P(X, \mathbf{U}) = \sum_{x,\mathbf{u}} P(x, \mathbf{u}) \log \frac{P(x,\mathbf{u})}{P(x)\,P(\mathbf{u})} = \sum_{x,\mathbf{u}} P(x, \mathbf{u}) \log \frac{P(x \mid \mathbf{u})}{P(x)}$

$$= \sum_{x,\mathbf{u}} P(x, \mathbf{u}) \log P(x \mid \mathbf{u}) - \sum_{x,\mathbf{u}} P(x, \mathbf{u}) \log P(x)$$

$$= \sum_{x,\mathbf{u}} P(x, \mathbf{u}) \log P(x \mid \mathbf{u}) - \sum_{x} P(x) \log P(x) \;\cancelto{1}{\sum_{\mathbf{u}} P(\mathbf{u}|x)}$$

$$= \quad -H(X \mid \mathbf{U}) \quad + \quad H(X)$$

Score for Belief Network

- $\mathcal{I}(X, \mathbf{U}) = H(X) - H(X \mid \mathbf{U})$

  $\Rightarrow \; H(X \mid \mathbf{Pa}_{X,\mathcal{G}}) = H(X) - \mathcal{I}(X, \mathrm{Pa}_{X,\mathcal{G}})$

Doesn't involve the structure, $\mathcal{G}$!

- Log data likelihood

$$\log P(D|\theta, G) = m \sum_i \mathcal{I}(X_i, \mathbf{Pa}_{X_i,\mathcal{G}}) - m \boxed{\sum_i H(X_i)}$$

- So use score: $\sum_i \mathcal{I}(X_i, \mathbf{Pa}_{X_i,\mathcal{G}})$

Best Tree Structure

$\log P(D \mid \theta, G)$ is monotonic with $\sum_i I(x_i, Pa_{x_i,G})$

- Identify tree with set $\mathcal{F} = \{ Pa(X) \}$
  - each $Pa(X)$ is either $\{\}$, or another variable
- Optimal tree, given data, is
  $\text{argmax}_{\mathcal{F}} \sum_i I(X_i, Pa(X_i))$

- So … want parents $\mathcal{F}$ s.t.
  - tree structure
  - maximizes $\sum_i I(X_i, Pa(X_i))$

Chow-Liu Tree Learning Alg

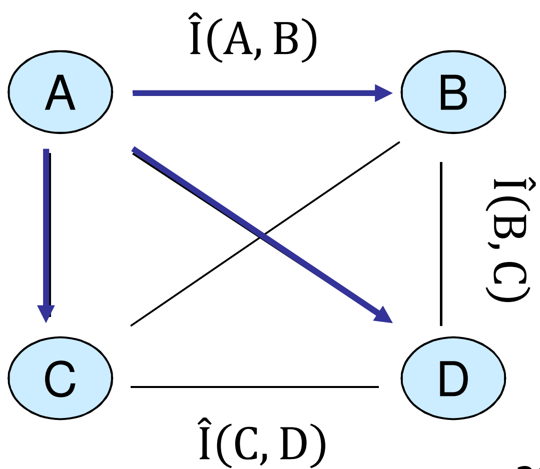
- For each pair of variables X_i, X_j
  - Compute empirical distribution:

$$\widehat{P}(x_i, x_j) = \frac{\text{count}(x_i, x_j)}{m}$$

  - Compute mutual information:

$$\hat{I}(X_i, X_j) = \sum_{x_i, x_j} \widehat{P}(x_i, x_j) \log \frac{\widehat{P}(x_i, x_j)}{\widehat{P}(x_i)\widehat{P}(x_j)}$$

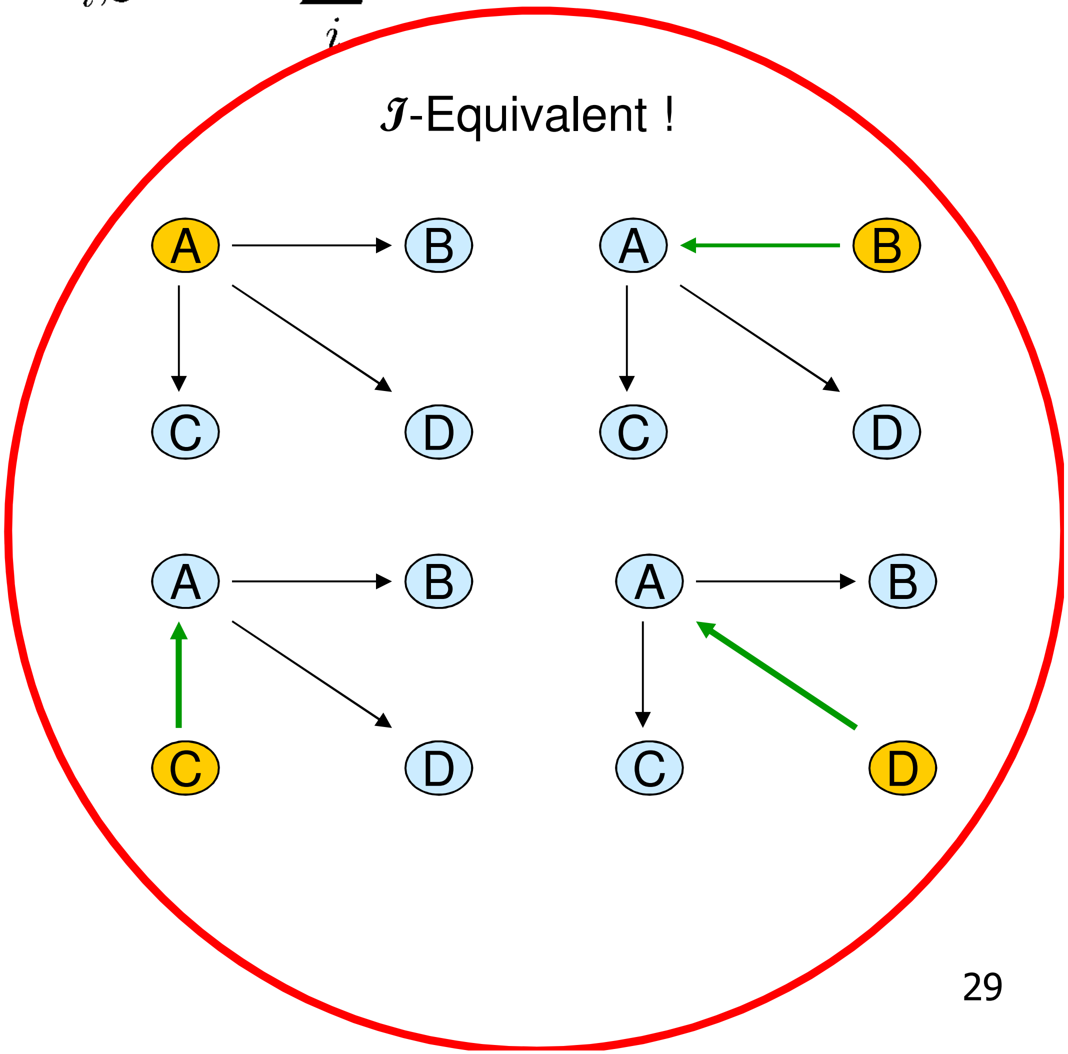
- Define a graph
  - Nodes $X_1, \ldots, X_n$
  - Edge (i,j) gets weight $\hat{I}(X_i, X_j)$
- Find Maximal Spanning Tree
- Pick a node for root, dangle…



Chow-Liu Tree Learning Alg … 2

$$\log \hat{P}(\mathcal{D} \mid \theta, \mathcal{G}) = m \sum_i \hat{I}(x_i, \mathbf{Pa}_{x_i,\mathcal{G}}) - m \sum_i \hat{H}(X_i)$$

- Optimal tree BN
  - …
  - Compute maximum weight spanning tree
  - Directions in BN:
    - pick any node as root, …doesn't matter which!
    - breadth-first-search defines directions
- Score Equivalence:
  If $\mathcal{G}$ and $\mathcal{G}'$ are $\mathcal{I}$-equiv, then scores are same



Chow-Liu (CL) Results

- If distribution P is tree-structured,
  CL finds CORRECT one

- If distribution P is NOT tree-structured,
  CL finds tree structured Q that
  has min'l KL-divergence: $\text{argmin}_Q\ KL(P; Q)$

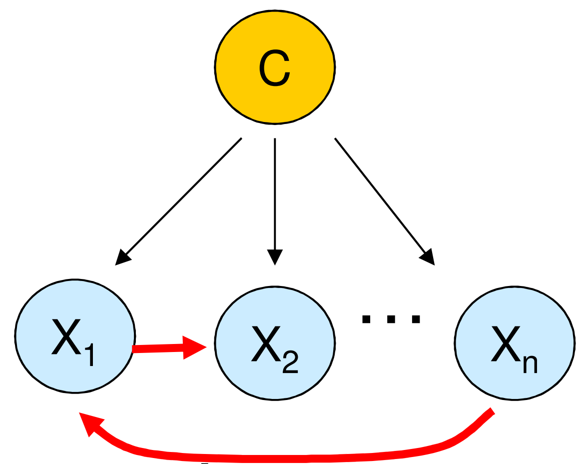
- Even though $2^{\theta(n \log n)}$ trees,
  CL finds BEST one in poly time $O(n^2 [m + \log n])$

number of variables

number of instances

Using Chow-Liu to Improve NB

- Naïve Bayes model
  - $X_i \perp X_j \mid C$
  - Ignores correlation between features
  - What if $X_1 = X_2$? **Double count…**



- Avoid by conditioning features on one another
- Tree Augmented Naïve bayes (TAN)
  [Friedman et al. '97]
  - Same as Chow-Liu, but score edges with:

$$I(X_i, X_j \mid C) = \sum_{x_i, x_j, c} P(x_i, x_j, c) \log \frac{P(x_i, x_j \mid c)}{P(x_i \mid c)\, P(x_j \mid c)}$$

All but ONE feature have 2 parents: C, X_i

Can we extend Chow-Liu ?

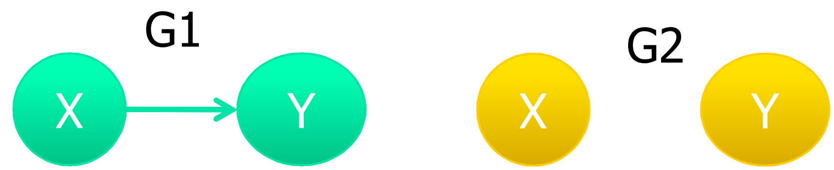
- (Approximately learning) models with tree-width up to *k*
  - [Narasimhan & Bilmes '04]
  - But, $O(n^{k+1})$…
    - and more subtleties

More Elaborate Models

- Space is sooo large
- Heuristic search
  - Pick a decomposable score
  - Pick a reasonable start, check neighbourhood
    - Add arc
    - Delete arc
    - Reverse arc

    All easy checks, compare 1 or 2 CPDs, (cache Δ's)
  - Move to better score
    - If no better: terminate

Likelihood Overfits To Data



- Consider
  - $\log P_{G1}(\mathbf{D}) = I(X;Y) - H(X) - H(Y)$
  - $\log P_{G2}(\mathbf{D}) = \qquad\qquad - H(X) - H(Y)$

$\Rightarrow$ difference is mutual information

- Note:

  $\log P_{G1}(\mathbf{D}) - \log P_{G2}(\mathbf{D}) = I(X;Y) \geq 0$

$\Rightarrow$ G1 always preferred than G2

Maximum likelihood score overfits!

$$\log \hat{P}(\mathcal{D} \mid \theta, \mathcal{G}) = m \sum_i \hat{I}(X_i, \mathbf{Pa}_{X_i,\mathcal{G}}) - m \sum_i \hat{H}(X_i)$$

- Adding a parent never decreases score!!!
  - ***Facts:*** $H(X \mid Pa_{X,\mathcal{G}}) = H(X) - I(X, Pa_{X,\mathcal{G}})$

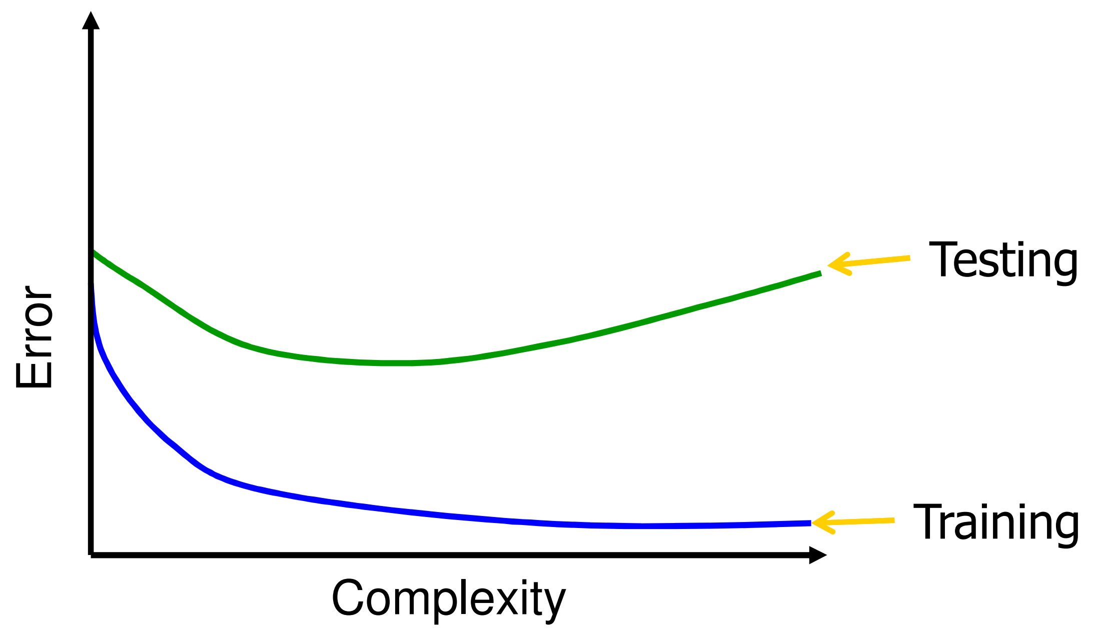
    $H(X \mid A) \geq H(X \mid A \cup Y)$
  - $I(X_i, \mathbf{Pa}_{X_i,G} \cup Y) = H(X_i) - H(X_i \mid \mathbf{Pa}_{X_i,G} \cup Y)$

    $\geq H(X_i) - H(X_i \mid \mathbf{Pa}_{X_i,G})$

    $= I(X_i, \mathbf{Pa}_{X_i,G})$
- So score increases as we add edges!
  - Best is COMPLETE Graph
  - Maximum likelihood score overfits !

Likelihood Overfits To Data

- If additional arcs always favoured Then prefer "Complete Model", K_N
- With fixed data:
  - more complex models will overfit



Bayesian Score

- Prior distributions:
  - Over structures
  - Over parameters of a structure

  Goal: Prefer simpler structures... regularization …
- Posterior over structures given data:
  - $P(\mathcal{G}|\mathcal{D}) \propto P(\mathcal{D}|\mathcal{G}) \times P(\mathcal{G})$

    Posterior Likelihood Prior over Graphs

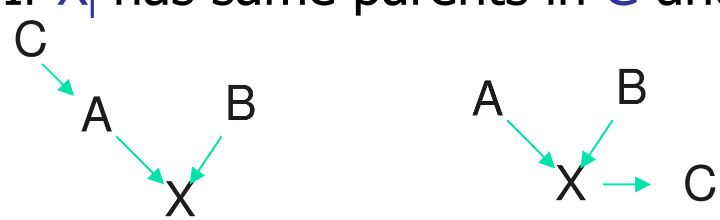
    Prior over Parameters
  - $P(\mathcal{D}|\mathcal{G}) = \int_\Theta P(\mathcal{D} \mid \mathcal{G}, \Theta)\, P(\Theta|\mathcal{G})\, d\Theta$

$$\log P(\mathcal{G} \mid D) \approx \log P(\mathcal{G}) + \log \int_{\theta_\mathcal{G}} P(D \mid \mathcal{G}, \theta_\mathcal{G}) P(\theta_\mathcal{G}|\mathcal{G}) d\theta_\mathcal{G}$$

Towards a decomposable Bayesian score

$$\log P(\mathcal{G} \mid D) \approx \log P(\mathcal{G}) + \log \int_{\theta_{\mathcal{G}}} P(D \mid \mathcal{G}, \theta_{\mathcal{G}}) P(\theta_{\mathcal{G}} \mid \mathcal{G}) d\theta_{\mathcal{G}}$$

- **Local and global parameter independence** $\theta_{Y|+x} \perp \theta_X$
- Prior satisfies **parameter modularity**:
  - If X_i has same parents in G and G′, then parameters have same prior



$\Theta(X; A,B)$ same in both structures

- Structure prior $P(\mathcal{G})$ satisfies **structure modularity**
  - Product of terms over families
  - Eg, $P(\mathcal{G}) / c^{|\mathcal{G}|}$ $|\mathcal{G}|$=#edges; c<1

- … then: Bayesian score decomposes along families!
  - $\log P(\mathcal{G}|\mathcal{D}) = \sum_X \text{ScoreFam}(X \mid Pa_X : \mathcal{D})$

Skip

Priors for General Graphs

- For finite datasets, prior is important!
- Prior over structure satisfying prior modularity
  - Eg, $P(\mathcal{G}) \propto c^{|\mathcal{G}|}$ $|\mathcal{G}|$=#edges; $c<1$

- What is good prior over *all* parameters?
  - *K2 prior* : fix $\alpha \in \Re^+$, set $\theta_{X_i|\boldsymbol{Pa}(X_i)} \sim \text{Dirichlet}(\alpha, \ldots, \alpha)$
  - Effective sample size, wrt X_i ?
    - If 0 parents: $k \times \alpha$
    - If 1 binary parent: $2 \; k \times \alpha$
    - If d k-ary parents: $k^d \; k \times \alpha$
  - So X_i *"effective sample size"* depends on #parental assignments
    - More parents $\Rightarrow$ strong prior… doesn't make sense!
  - K2 is "inconsistent"

Summary wrt Learning BN Structure

- Decomposable scores
  - Data likelihood
  - Information theoretic interpretation
  - Bayesian
- Priors
  - Structure and parameter assumptions
- Best tree (Chow-Liu)
- Best TAN
- Bayesian model averaging