ECE 175B: Probabilistic Reasoning and Graphical Models
Lecture 12: Model Requirements and Simplifications

Florian Meyer & Ken Kreutz-Delgado

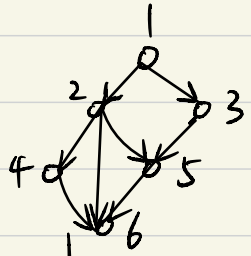
Electrical and Computer Engineering Department
University of California San Diego

UC San Diego
JACOBS SCHOOL OF ENGINEERING

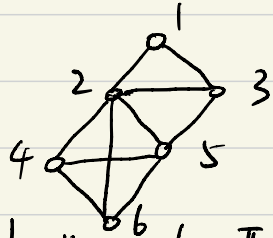
Ancestral Graph Method for Testing CI statements in BNs

- We simplify notation with $\perp\!\!\!\perp_G$: $X \perp\!\!\!\perp_G Y \mid Z \triangleq \langle X \mid Z \mid Y \rangle_G$
  Then $i \perp\!\!\!\perp_G j \mid k \implies X_i \perp\!\!\!\perp X_j \mid X_k$.
- Testing for graph separation in MNs is easier than in BNs.
- However given a candidate CI statement $X \perp\!\!\!\perp_G Y \mid Z$ for a BN G, we can't test if it is true in the moral graph $M(G)$ (Recall that $\mathbb{I}(M(G)) \subset \mathbb{I}(G)$) because CI statements involving collider nodes are lost during moralization.

Example:



$\implies$ $M(G)$:



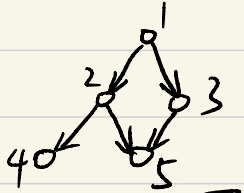
"$4 \perp\!\!\!\perp_G 5 \mid 2$" $\in \mathbb{I}(G)$ "$4 \perp\!\!\!\perp_G 5 \mid 2$" $\notin \mathbb{I}(M(G))$

- Note that it is the "downstream" collider node "6" that causes the problem because it results in the moralization link "4-5".

Ancestral Graph Method – Cont'd

- For $\alpha =$ "$X \perp\!\!\!\perp_G Y \mid Z$", define $U = X \cup Y \cup Z$.
  For our example: $U = \{2, 4, 5\}$.
- Define $An_G(U) = U \cup \underbrace{Ancestor(U)}_{\text{nodes upstream of } U} =$ ancestral graph
- We focus on the subgraph $An_G(U)$, meaning that we throw away all nodes of G not in $An_G(U)$ as those nodes (including the downstream nodes) are irrelevant for ascertaining if $\alpha \in \Pi(G)$.
- I.e., only the nodes in $An_G(U)$ matter for ascertaining if $\alpha \in \Pi(G)$.
- For our example,

$An_G(U)$:

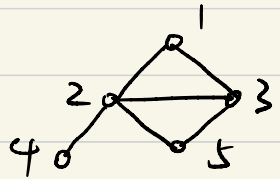


- It is the case that $\Pi(An_G(G)) \subset \Pi(G)$, i.e., $An_G(G)$ is faithful to G.

Ancestral Graph Method - Cont'd

- Now define $MA_U(G) = M(A_U(G))$ the moral graph of $A_U(G)$ then $\Pi(MA_U(G)) \subset \Pi(A_U(G)) \subset \Pi(G)$.
  i.e., $\forall \alpha \in \Pi(MA_U(G)) \Rightarrow \alpha \in \Pi(G)$.
- For our example,

$MA_U(G)$ =



$\langle 4 \mid 2 \mid 5 \rangle$ in this MN.

$\Downarrow$

$\alpha =$ "$4 \perp\!\!\!\perp_{MA_U(G)} 5 \mid 2$" $\in \Pi(MA_U(G))$

$\Downarrow$

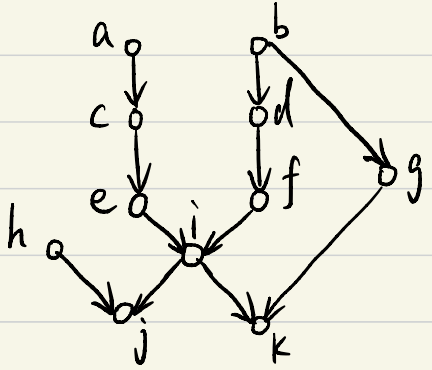
$\alpha \in \Pi(G)$, i.e $X_4 \perp\!\!\!\perp X_5 \mid X_2$.

- Procedure: step 1: Given G and $\alpha =$ "$X \perp\!\!\!\perp_G Y \mid Z$", form $U = X \cup Y \cup Z$, and $A_U(G) = U \cup \text{ancestor}(U)$
  step 2: Form $MA_U(G) = M(A_U(G))$
  step 3: Test if $X \perp\!\!\!\perp_{MA_U(G)} Y \mid Z$ is true.

Ancestral Graph Method – Cont'd

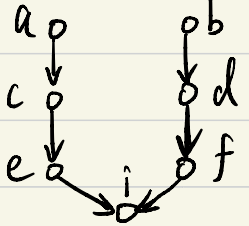
Example: (see. BRML Fig. 4.4). Is $a \perp\!\!\!\perp_G b \mid \{d, i\} \Rightarrow X_a \perp\!\!\!\perp X_b \mid \{X_d, X_i\}$?

G:

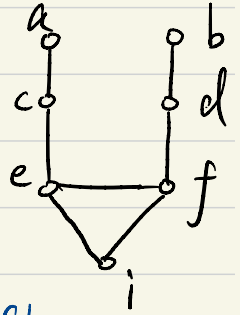


step 1 $\Rightarrow U = \{a, b, d, i\}$

$An_U(G)$:

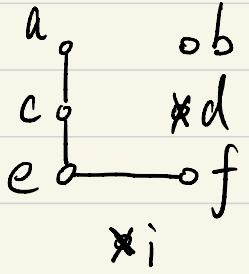


step 2 $\Rightarrow MA_U(G)$



step 3 ⇓

$$
\begin{array}{lll}
a & \perp\!\!\!\perp_{MA_U(G)} & b \mid d, i \\
a & \perp\!\!\!\perp_G & b \mid d, i \\
X_a & \perp\!\!\!\perp & X_b \mid \{X_d, X_i\}
\end{array}
\quad \Leftarrow
$$

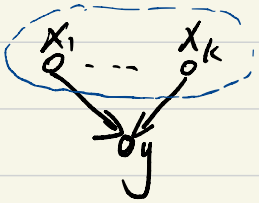


$\langle a \mid d, i \mid b \rangle$

Model Simplification

- Graphical Models can quickly become intractable complex.
- Model approximations can often be made that provide significant reductions in complexity which provides reasonably accurate performance.
- We discuss three:
  - Logistic Regression Model
  - "Naive Bayes" assumption
  - Gaussian Linear Model / Gaussian BNs

Logistic Regression Model



- Parents-to-node is a "fan-in" situation.

$$P(y, \underbrace{x_1, \cdots, x_k}_{pa(y)}) = P(y \mid \underbrace{x_1, \cdots, x_k}_{pa(y)}) P(x_1) \cdots P(x_k)$$

If all binary, 2^k values must be specified or learned (e.g. $2^{10} = 1024$)

- Assume a Logistic Regression probability model. This is an "abstraction step" in the modeling process.

A). Regression step: Learn or specify a linear regression model.

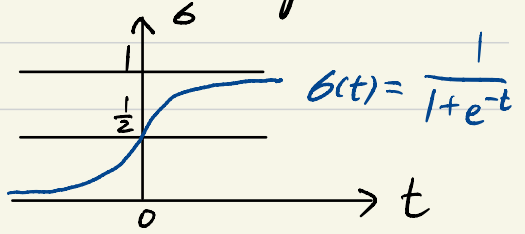
$$t(pa(y)) = b_0 + b_1 x_1 + \cdots + b_k x_k$$

$k+1$ parameters are needed, (e.g. $k = 10 \Rightarrow 11$ values)

B). Logistic step: Apply the nonlinear "sigmoid" Logistic function.

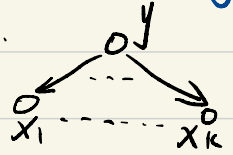
$$\sigma(t) = \frac{1}{1+e^{-t}} = \frac{e^t}{e^t+1} \quad \text{to } t = t(pa(y))$$

$$\Rightarrow P(y \mid pa(y)) = \sigma(t(pa(y)))$$

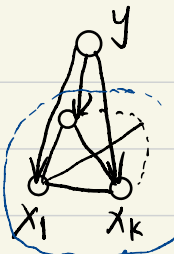


Naive Bayes Model

- The "fan-out" model.



as an significant simplification of



The naive Bayes Assumption:

$$P(x_1, \cdots x_K | y) = P(x_1|y) \cdots P(x_K|y)$$

If binary, need specify $2K$ values (e.g. $k=10 \to 20$ values)

$$P([x_1, \cdots, x_K]|y)$$

if binary, we need specify $2^{K+1}-1$ values (e.g. $k=10 \to 2047$ values)

- The naive Bayes assumption can work quite well if y is a "hidden factor" that explains a lot of the covariance of $x_1, \cdots, x_K$
  - The theory of finding such (often hidden) factors is called "Factor Analysis" (FA)
  - It is closely related to Principal Component Analysis (PCA), which attempts to explain the variance of $x_1, \cdots, x_K$.

The Gaussian Model.

- As we know, a fully connected graph is highly complex, and this giant clique is hard to unravel. In probability, $P(X) = P(x_1, \cdots, x_N)$ need $2^N - 1$ specific values even binary. (e.g. $N = 100, \rightarrow \approx 10^{30}$ values)
- Suppose $X_i \in \mathbb{R}^1 = \mathbb{R}$ are jointly Gaussian, $i = 1, \cdots, N$.
  i.e., that $X \in \mathbb{R}^N$ is a N-dimensional, Gaussian random vector, denoted as $X \sim \mathcal{N}(\mu, \Sigma)$, where the mean μ needs N parameters, and the covariance matrix Σ needs $\frac{N(N+1)}{2}$ parameters. Then a total of $\boxed{\frac{N(N+3)}{2}}$ parameters.
- The drop from $(2^N - 1)$ parameters to $\frac{N(N+3)}{2}$ parameters is a stunning reduction of complexity arising from the "abstracting step" of using an abstract probability model.
- For $N = 100$, we need 5150 parameters for the Gaussian model.
- We gain additional efficiency by modularizing (divide and conquer) the world using a Bayesian Network (BN).

Gaussian BN

- Let $X = \{x_i, i=1,\cdots,N\}$: $P(X) = \prod_{i=1}^{N} P(x_i | pa(x_i))$
- Gaussian Linear Model (see eq. 10.15 Murphy)

$$x_i = \mu_i + \sum_{x_s \in pa(x_i)} w_{is}(x_s - \mu_s) + e_i, \text{ where } e_i \sim N(0, \sigma_i^2), \text{ for } i = 1,\cdots,N \quad (*)$$

Note that this is equivalent to $P(x_i | pa(x_i)) \sim N(\mu_i + \sum_{x_s \in pa(x_i)} w_{is}(x_s - \mu_s), \sigma_i^2)$

- Stacking x_i, μ_i, e_i into vectors as X, μ, e, and w_{is} into matrix W, we have $(I - W)(X - \mu) = e$ (**) elements of W are either 0 or w_{is}

Assume ancestral ordering. W is lower-triangular (see eq 10.19 Murphy)

- Since one can empirically estimate $\mu = E(X)$ and $\Sigma = cov(X)$ from (**) one can obtain estimates of w_{is} and σ_i^2 as discussion in eq. 10.22 Murphy
- Let $|pa(x_i)| = p_i$, take the value $p = 5$ as the average. The number of parameters is N for means μ_i, N for variance σ_i^2, and $5N$ for weights w_{is}. (e.g., $N = 100 \rightarrow 700$ parameters). We have reduced the complexity $10^{30} \rightarrow 5150 \rightarrow 700$.