Bayesian Networks in Educational Assessment Tutorial

Session II: Bayesian Networks

Russell Almond, Bob Mislevy, David Williamson and Duanli Yan

Unpublished work © 2002-2014 ETS
# Agenda

<table>
<thead>
<tr>
<th>SESSION</th>
<th>TOPIC</th>
<th>PRESENTERS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Session 1</td>
<td>Evidence Centered Design</td>
<td>David Williamson</td>
</tr>
<tr>
<td>Session 2</td>
<td>Bayesian Networks</td>
<td>Russell Almond</td>
</tr>
<tr>
<td>Session 3</td>
<td>Bayes Net Tools &amp; Applications</td>
<td>Duanli Yan</td>
</tr>
<tr>
<td>Session 4</td>
<td>ACED: ECD in Action</td>
<td>Russell Almond &amp; Duanli Yan</td>
</tr>
<tr>
<td>Session 5</td>
<td>Refining Bayes Nets with Data</td>
<td>Bob Mislevy</td>
</tr>
</tbody>
</table>
The Urn Model

- Urn contains $b$ black balls and $w$ white balls, all same size & weight
- Draw a ball out without looking
- Probability of drawing black ball is $b/(b+w)$
Cup and Cap notation

- In probability theory, events are sets (sets of balls in the urn).
- Let $A$ and $B$ be two events
- Either $A$ or $B$ occurs
  - Corresponds to union of sets
  - $A \cup B$
- Both $A$ and $B$ occur
  - Corresponds to intersection of sets
  - $A \cap B$
  - Sometimes $Pr(A, B)$
- Not $A$ – the balls in the urn that are not in event $A$
  - $\neg A$
  - $Pr(\neg A) = 1 - Pr(A)$
Conditional Probability

- **Definition**

\[
\Pr(E|H) = \frac{\Pr(E \cap H)}{P(H)}
\]

- **Law of Total Probability**

\[
\Pr(E) = \Pr(E|H) \Pr(H) + \Pr(E|\neg H) \Pr(\neg H)
\]
Bayes Theorem

\[ \Pr(H|E) = \frac{\Pr(E|H) \Pr(H)}{\Pr(E)} = \frac{\Pr(E|H) \Pr(H)}{\Pr(E|H) \Pr(H) + \Pr(E|\neg H) \Pr(\neg H)} \]

- Prior \( \Pr(H) \)
- Likelihood \( \Pr(E|H) \)
- Posterior \( \Pr(H|E) \)
Aside on Objective vs Subjective

• All procedures discussed are agnostic with respect to interpretation
• Subjective probabilities are judgments that an event is like a game of chance which has certain frequency properties
• Probabilities are objective when the basis for the judgment is documented
• Probabilities are subjective wrt sources of information used
• Some people have problems with strong assumptions about “prior” but no hesitation about stronger assumptions about “likelihood”
Independence

\[ \Pr(B) = \Pr(B|A) = \Pr(B|\neg A) \]

\[ \Pr(A) = \Pr(A|B) = \Pr(A|\neg B) \]

\[ \Pr(A \cap B) = \Pr(B|A) \Pr(B) = \Pr(A) \Pr(B) \]

- A provides no information about B
Accident Proneness (Feller, 1968)

- Driving Skill: 5/6 Normal, 1/6 Accident Prone
- Probability of an accident in a given year
  - 1/100 for Normal drivers
  - 1/10 for Accident prone drivers
- Accidents happen independently in each year
- What is the probability a randomly chosen driver will have an accident in Year 1?
- Given a driver had an accident in Year 1, what is probability of accident in Year 2?
Accident Proneness (cont)

- What is the probability a randomly chosen driver will have an accident in Year 1? Year 2?

\[
P(A_i) = P(A_i|N)P(N) + P(A_i|\bar{N})P(\bar{N})
\]

\[
= \frac{0.05}{6} + \frac{1}{6} = \frac{0.15}{6} = 0.025.
\]
Accident Proneness (cont)

• Given a driver had an accident in Year 1, what is probability of accident in Year 2?

\[
P(A_1 \cap A_2) = P(A_1 \cap A_2|N)P(N) + P(A_1 \cap A_2|\overline{N})P(\overline{N})
\]
\[
= P(A_1|N)P(A_2|N)P(N) + P(A_1|\overline{N})P(A_2|\overline{N})P(\overline{N})
\]
\[
= .01 \times .01 \times \frac{5}{6} + .1 \times .1 \times \frac{1}{6}
\]
\[
= \frac{.0005}{6} + \frac{.01}{6} = \frac{.0105}{6} = .00175 .
\]

Note that

\[
P(A_2|A_1) = \frac{P(A_1 \cap A_2)}{P(A_2)} = \frac{.00175}{.025} = .07 .
\]
Conditional Independence

- Years are \textit{conditionally independent} given driving skill

\[
p(\text{Y}_1, \text{Y}_2 | S) = p(\text{Y}_1 | S)p(\text{Y}_2 | S)
\]

- Years are \textit{marginally dependent}
- Separation in graph tells the story
Review of Graph Theory

• Graph consists of set of
  – Nodes: Vertexes, Variables
  – Edges: Arcs, Links, Arrows

• Four kinds of graph
  – Simple: Edges are unordered pairs of nodes
  – Directed (Digraph): Edges are ordered pairs of nodes
  – Hypergraph: Hyperedges contain 1 or more nodes
  – Directed Hypergraph: Directed hyperedges contain set of to and from nodes
Simple Graph

- **Neighbors** are nodes connected by an edge
- **Complete** set of nodes are all neighbors
  - **Clique** is a maximal complete set
Directed Graph

- Edges go from *parent* to *child*
  - *Ancestors* (*descendants*) are produced through recursive application of the parent (child) relationship
  - Write $pa(X|G)$ for parents of Node $X$ in Graph $G$
Paths and Trees

- A sequence of nodes so that each is a neighbor of the previous one is called a path
  - For directed graphs, path must follow the direction of the arrows (undirected paths are called chains)
- Two nodes are connected if there is a path (chain) between them
- A path which returns to the starting node is called a cycle
- A graph with no cycles is called acyclic
  - An acyclic undirected graph is called a tree
Acyclic Digraphs

- Acyclic directed graphs play a special role in Bayesian networks
- Sometimes called DAG (Directed Acyclic Graph) although this is technically incorrect
Triangulated Graphs

- A shortcut through a cycle (of length >3) is called a chord
- A graph which has no cycles of length >3 without a chord is called triangulated
- A graph which is not triangulated can be “filled in” to make it triangulated
Graphical Models

• Nodes in graph are random variables in joint probability
• Markov Property: Separation in graph implies conditional independence
• Gibbs Property: Probability distribution factors according to graph
• *Bayesian Network*
  – Directed graph
  – Nodes usually represent discrete variables
Simple Graphs

- Separation in graph implies conditional independence given separating set
- A, B independent of D, E given C
- C independent of F given D and E
- Can break triangulated graph up into cliques and intersections
Competing Explanations

- Skill 1 and Skill 2 are (a priori) independent in population
- Task X requires both skills (conjunctive model)
- Answer the following questions:
  - What is posterior after learning $X=\text{False}$, and $\theta_1=\text{High}$?
  - What is posterior after learning $X=\text{False}$, and $\theta_2=\text{High}$?
  - What is true of joint posterior of $\theta_1$ and $\theta_2$ after learning $X=\text{False}$?
D-Separation

- For $\leftrightarrow$, $\rightarrow\rightarrow$, and $\leftrightarrow\leftrightarrow$ edges conditioning on middle variables renders outer variables independent.
- For $\rightarrow\leftrightarrow$ (collider) edges, if middle variable (or descendent is known) then variables are dependent.
- A path is *active* if collider with middle node observed, or non-collider with middle node unobserved.
D-Separation Exercise

• Are $A$ and $C$ independent if
  1. We have observed no other variables?
     • What could we condition on to make $A$ and $C$ independent?
  2. We have observed $F$ and $H$?
     • What else could we condition on to make $A$ and $C$ independent?
  3. We have observed $G$?
     • What else could we condition on to make $A$ and $C$ independent?
Building Up Complex Networks

- Recursive representation of probability distributions:

\[ p(x_1, \ldots, x_n) = p(x_n \mid x_{n-1}, \ldots, x_1)p(x_{n-1} \mid x_{n-2}, \ldots, x_1) \cdots p(x_2 \mid x_1)p(x_1) \]

\[ = \prod_{j=1}^{n} p(x_j \mid x_{j-1}, \ldots, x_1) = \prod_{j=1}^{n} p(x_j \mid Pa(x_j)), \]

- All orderings are equally correct, but some are more beneficial because they capitalize on causal, dependence, time-order, or theoretical relationships that we posit:

Terms simplify when there is conditional independence – in ed measurement, due to unobservable student variables.
Building Up Complex Networks, cont.

- For example, in IRT, item responses are conditionally independent given $\theta$:

\[
p(x_1, \ldots, x_n, \theta) = p(x_n | x_{n-1}, \ldots, x_1, \theta)p(x_{n-1} | x_{n-2}, \ldots, x_1, \theta) \cdots p(x_2 | x_1, \theta)p(x_1 | \theta)p(\theta)
\]

\[
= p(x_n | \theta) \quad p(x_{n-1} | \theta) \quad \cdots p(x_2 | \theta) \quad p(x_1 | \theta)p(\theta)
\]

\[
= \prod_{j=1}^{n} p(x_j | \theta)p(\theta).
\]
Bayes net

- One factor for each node in graph in recursive representation
- This factor is conditioned on parents in graph
- “Prior” nodes have no parents
- \( p(A)p(B)p(C|A,B)p(D|C)p(E|C)p(F|D,E) = p(A,B,C,D,E,F) \)
- Digraph must be acyclic
Directed Hypergraph

- Augment picture with “boxes” representing probability distributions
  - Can use icons to distinguish parameterizations of distributions (Conjunctive, Compensatory, &c)
- Boxes are in fact directed hyperedges
- 2-section of this graph is formed by connected all nodes which participate in each hyperedge
Simple (Moralized) Graph

- Note that in this view, nodes which are parents of the same child (A and B; D and E) are “married”
- Competing explanation induces dependency between (among) parents
- Joint probability factors according to potentials on cliques
- \( p(A,B,C,D,E,F) = \phi(A,B,C) \phi(C,D,E) \phi(D,E,F) \)
  - \( \phi(A,B,C) = p(A)p(B)p(C|A,B) \)
  - \( \phi(C,D,E) = p(D|C)p(E|C) \)
  - \( \phi(D,E,F) = p(F|D,E) \)
- Treewidth of graph is size of largest clique, cost is exponential in treewidth
Directing the Arrows

• Causal

![Causal Diagram]

• Diagnostic

![Diagnostic Diagram]

• Causal is usually more convenient, but strict causality is not necessary

• *Arc reversal* goes between representations
Gibbs--Markov Equivalence

• Moussouris (1974)
  – Factorization implies conditional independence
  – Conditional independence implies factorization
    (provided distribution is strictly positive)

• Can go back and forth between digraph and moral graph views of Bayes net

• Digraph for model construction, moral graph (junction tree) for computation
Decision Theory

• Influence Diagram
  – Round Nodes = Chance (random) variables
  – Square Boxes = Decision variables
  – Hexagons = Utilities (costs)

• Influence Diagram with only chance nodes is a Bayes net
Intervention Influence Diagram

Skill Time 1 -> Skill Time 2 -> Value of Skill
Intervention Influence Diagram
Intervention Influence Diagram

- Pretest Score
- Select Test
- Cost of Testing
- Skill Intervention
- Cost of Intervention
- Skill Time 1
- Skill Time 2
- Value of Skill
Bayes nets and Structural Equation Models

• Both provide a picture where nodes are variables, edges represent relationships among variables
• Bayes net graph must be acyclic, SEMs allow bidirectional edges
• Separation in Bayes nets implies conditional independence
  – SEM diagrams often picture marginal dependence
• Efficient computation algorithms always exist for Bayes nets, may or may not exist for SEM
Dance Competition

• Dance competition attempts to evaluate the ballet skills of a number of dancers

• Draw a Bayes net for each of the following situations:
  – Each dancer gives a single performance judged by a single rater
  – Each dancer gives a single performance judged by three raters
  – Each dancer gives three performances, each performance judged by a single rater
  – Each dancer gives three performances, each performance judged by (the same) three raters
Computation in Bayes Nets

The setup, with two random variables, X and Y:

- You know conditional probabilities, \( p(x_j | y_k) \). They tell you what to believe about X if you knew the value of Y.
- You learn \( X=x^* \); what should you believe about Y?
- You combine two things:
  - Conditional probability (the *likelihood*): \( p(x_j | y_k) \)
  - Previous probabilities about Y values: \( p(y_k) \)

Bayes Theorem with 2 Variables

\[
p(y_k | x^*) \propto p(x^* | y_k) p(y_k)
\]
Inference in a Markov Chain

Recursive representation:

\[ p(u,v,x,y,z) = p(z|y,x,v,u) \ p(y|x,v,u) \ p(x|v,u) \ p(v|u) \ p(u) \]
\[ = p(z|y) \ p(y|x) \ p(x|v) \ p(v|u) \ p(u). \]
Inference in a Chain

Suppose we learn the value of $X$:

Start here, by revising belief about $X$
Inference in a Chain

Propagate information down the chain using conditional probabilities:

From updated belief about $X$, use conditional probability to revise belief about $Y$
Inference in a Chain

Propagate information down the chain using conditional probabilities:

From updated belief about Y, use conditional probability to revise belief about Z.
Inference in a Chain

Propagate information up the chain using Bayes Theorem:

From updated belief about X, use Bayes Theorem to revise belief about V

\[
p(v|u) \quad p(x|v) \quad p(y|x) \quad p(z|y)
\]
Inference in a Chain

Propagate information up the chain using Bayes Theorem:

From updated belief about V, use Bayes Theorem to revise belief about U.
Inference in Trees

In a tree each variable has no more than one parent. Suppose we learn the value of X. We can update every variable in the tree using either the conditional probability relationship or Bayes theorem.
Inference in Multiply-Connected Nets
Inference in Multiply-Connected Nets

In a multiply-connected graph, in at least one instance there is more than one path from one variable to another variable. Repeated applications of Bayes theorem and conditional probability at the level of individual variables doesn’t work.
Inference in Multiply-Connected Nets

Key idea: Group variables into subsets ("cliques") such that the subsets form a tree.
Inference in Multiply-Connected Nets

Key idea: Group variables into subsets ("cliques") such that the subsets form a tree.
Inference in Multiply-Connected Nets

Key idea: Group variables into subsets (“cliques”) such that the subsets form a tree.
Inference in Multiply-Connected Nets

Key idea: Group variables into subsets ("cliques") such that the subsets form a tree.
Can the update cliques with a generalized version of updating algorithm for individual variables in cliques.
The Lauritzen-Spiegelhalter Algorithm

1. Recursive representation of the joint distribution of variables.
2. Directed graph representation of (1).
4. Determination of cliques and clique intersections.
5. Join tree representation.
7. Updating scheme (message passing)
Two-Skills Example
(adapted from Andreassen, Jensen, & Olesen)

- Two skills: $\theta_A$ and $\theta_B$;
  each could be High (H) or Low (L).
- Two tasks: observable variables $X_1$ and $X_2$;
  each could be Right (1) or Wrong (0).
- The skills are modeled as independent,
- The responses as conditionally independent given skill states.
Aside: Medical diagnosis with observable symptoms of “latent” disease states has many parallels to measurement modeling in assessment

• State is a “construct,” inferred from theory & experience; proposed to organize our knowledge

• Conditional independence of observations given (possibly complex) underlying disease state

• Persistent interest in the underlying state

• Observations mainly of transitory interest

• States & relationships general knowledge, meant to aid thinking about unique cases but surely oversimplified

• State is the level at which treatment & prognosis is discussed, although there is often therapeutic/educational value in addressing specifics from observational setting
1) Recursive representation of joint distribution

\[
P(X_1, X_2, \Theta_A, \Theta_B) = P(X_1 | X_2, \Theta_A, \Theta_B) P(X_2 | \Theta_A, \Theta_B) \]

* \[P(\Theta_A | \Theta_B) P(\Theta_B)\]

\[
= P(X_1 | \Theta_A, \Theta_B) P(X_2 | \Theta_A, \Theta_B) P(\Theta_A) P(\Theta_B).
\]
2) DAG representation
2) DAG representation

Aside: A look ahead toward cognitive diagnosis

Good differential diagnosis value for “neither” vs. “at least one of the two”

Good differential diagnosis value for “ThetaB” vs. “not ThetaB”
2) DAG representation

Aside: A look ahead toward cognitive diagnosis

No differential diagnosis value for “which of the two?”

Good differential diagnosis value for “which of the two?”
“Marry” parents: Look at the set of parents of each variable. If they are not already connected, connect them. (Direction doesn’t matter, since we’ll drop it in the next step.) Rationale: If variables are all parents of the same variable, then even if they were independent otherwise, learning the value of their common child generally introduces dependence among them. We will need to include this possibility in our computational machinery.
3b) Undirected graph

Drop the directionality of the edges. Although the conditional probability directions were important for constructing the graph and will be important for building potential tables, we want a structure for computing that can go in any direction.
3c) Triangulated graph

Triangulation means looking at the undirected graph for cycles from a variable to itself going through a sequence of other variables. There should be no cycle with length greater than three. Whenever there is, add undirected edges so there are not cycles. The two skills moral graph is already triangulated, so it is not an issue here. A different example is shown above. Why do we do this? It is essential to producing cliques of variables that are trees. Can be many ways to do this; finding “best” one is NP-hard. People develop heuristic approaches.
4) Determine cliques and clique intersections

From the triangulated graph, one determines *cliques*, or subsets of variables that are all linked pairwise to one another. Cliques overlap, with sets of overlapping variables called clique intersections. The two cliques here are {GenA, GenB, X1} and {GenA, GenB, X2}. The clique intersection is {GenA, GenB}.
4) Determine cliques and clique intersections

• Cliques and intersections are the structure for local updating.

• The amount of computation grows roughly geometrically with clique size, as measured by the number of possible configurations of all values of all variables in a clique.
  
  ▪ A clique representation with many small cliques is therefore preferred to a representation with a few larger cliques.

• Strategies for increased efficiency include …
  
  ▪ defining “collector” variables,
  ▪ adding variables to break loops, and
  ▪ dropping associations when the consequences are benign.
5) Join tree representation

A join-tree representation depicts the *singly-connected structure* of cliques and clique intersections (i.e., no loops).

A join tree has the *running intersection property*: If a variable appears in two cliques, it appears in all cliques and clique intersections in the single path connecting them.
6) Potential tables

- Local calculation is carried out with tables that convey the joint distributions of variables within cliques, or *potential tables*.
- Basically brute force calculation within cliques. That’s why we want small cliques (low treewidth).
- Similar tables for clique intersections are used to pass updating information from one clique to another.
6) Potential tables

For each clique, determine the joint probabilities for all the possible combinations of values of all variables. For convenience, we have written them as matrices. These potential tables indicate the initial status of the network in our example—before specific knowledge of a particular individual's skills or item responses is known.

<table>
<thead>
<tr>
<th>$\theta_A$</th>
<th>$\theta_B$</th>
<th>$X_1:1$</th>
<th>$X_1:0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>H</td>
<td>.012</td>
<td>.000</td>
</tr>
<tr>
<td>H</td>
<td>L</td>
<td>.088</td>
<td>.010</td>
</tr>
<tr>
<td>L</td>
<td>H</td>
<td>.088</td>
<td>.010</td>
</tr>
<tr>
<td>L</td>
<td>L</td>
<td>.008</td>
<td>.784</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\theta_A$</th>
<th>$\theta_B$</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>H</td>
<td>.012</td>
</tr>
<tr>
<td>H</td>
<td>L</td>
<td>.098</td>
</tr>
<tr>
<td>L</td>
<td>H</td>
<td>.098</td>
</tr>
<tr>
<td>L</td>
<td>L</td>
<td>.792</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\theta_A$</th>
<th>$\theta_B$</th>
<th>$X_2:1$</th>
<th>$X_2:0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>H</td>
<td>.011</td>
<td>.001</td>
</tr>
<tr>
<td>H</td>
<td>L</td>
<td>.005</td>
<td>.093</td>
</tr>
<tr>
<td>L</td>
<td>H</td>
<td>.088</td>
<td>.010</td>
</tr>
<tr>
<td>L</td>
<td>L</td>
<td>.008</td>
<td>.784</td>
</tr>
</tbody>
</table>
6) Potential tables

The potential table for the clique intersection is the marginal distribution of ThetaA and ThetaB.

<table>
<thead>
<tr>
<th>( \theta_A )</th>
<th>( \theta_B )</th>
<th>( X_1:1 )</th>
<th>( X_1:0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>H</td>
<td>.012</td>
<td>.000</td>
</tr>
<tr>
<td>H</td>
<td>L</td>
<td>.088</td>
<td>.010</td>
</tr>
<tr>
<td>L</td>
<td>H</td>
<td>.088</td>
<td>.010</td>
</tr>
<tr>
<td>L</td>
<td>L</td>
<td>.008</td>
<td>.784</td>
</tr>
</tbody>
</table>

Marginal probs for ThetaA & ThetaB

<table>
<thead>
<tr>
<th>( \theta_A )</th>
<th>( \theta_B )</th>
<th>\text{Probability}</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>H</td>
<td>.012</td>
</tr>
<tr>
<td>H</td>
<td>L</td>
<td>.098</td>
</tr>
<tr>
<td>L</td>
<td>H</td>
<td>.098</td>
</tr>
<tr>
<td>L</td>
<td>L</td>
<td>.792</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( \theta_B )</th>
<th>( X_2:1 )</th>
<th>( X_2:0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>.011</td>
<td>.001</td>
</tr>
<tr>
<td>L</td>
<td>.005</td>
<td>.093</td>
</tr>
<tr>
<td>H</td>
<td>.088</td>
<td>.010</td>
</tr>
<tr>
<td>L</td>
<td>.008</td>
<td>.784</td>
</tr>
</tbody>
</table>
6) Potential tables

The potential table for Clique 1 is calculated using the prior probabilities of .11 for both ThetaA and ThetaB, the assumption that they are independent, and the conditional probabilities of X1 for each ThetaA * ThetaB combination.

<table>
<thead>
<tr>
<th>$\theta_A$</th>
<th>$\theta_B$</th>
<th>$X_1:1$</th>
<th>$X_1:0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>H</td>
<td>.012</td>
<td>.000</td>
</tr>
<tr>
<td>H</td>
<td>L</td>
<td>.088</td>
<td>.010</td>
</tr>
<tr>
<td>L</td>
<td>H</td>
<td>.088</td>
<td>.010</td>
</tr>
<tr>
<td>L</td>
<td>L</td>
<td>.008</td>
<td>.784</td>
</tr>
</tbody>
</table>

Conditional probs for X1 given ThetaA & ThetaB
6) Potential tables

Similar calculation for the other clique: Marginal probabilities of ThetaA & ThetaB, times the conditionals for X2.

Note that the implied distributions for ThetaA * ThetaB are consistent across both clique potential tables and the clique intersection table.

From these, we can reconstruct a coherent joint distribution for the entire set of variables.
7) Updating scheme

- Absorbing new evidence about a single variable is effected by re-adjusting the appropriate margin in a potential table that contains that variable, then propagating the resulting change to the clique to other cliques via the clique intersections.
- This process continues outward from the clique where the process began, until all cliques have been updated.
- The single-connectedness and running intersection properties of the join tree assure that coherent probabilities result.
7) Updating scheme

Suppose we learn $X_1=1$. Go to any clique where $X_1$ appears (actually there’s just one in this example). Zero out the entries for $X_1=0$. The remaining values express our new beliefs about the proportional chances that the other variables in that clique take their respective joint values.
7) Updating scheme

Propagate the new beliefs about \{\Theta_A, \Theta_B\} to the clique intersection. You could normalize these if you wanted to, but the proportional information is what matters.
7) Updating scheme

Propagate the new beliefs about \{ThetaA, ThetaB\} to the next clique. Divide each row by the old weight for that combination of clique-intersection variables and multiply it by the new one. I.e., the adjustment factor for each row is New Weight / Old Weight.

<table>
<thead>
<tr>
<th>Theta_A</th>
<th>Theta_B</th>
<th>X1: 1</th>
<th>X1: 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>H</td>
<td>.012</td>
<td>.000</td>
</tr>
<tr>
<td>H</td>
<td>L</td>
<td>.088</td>
<td>.010</td>
</tr>
<tr>
<td>L</td>
<td>H</td>
<td>.088</td>
<td>.010</td>
</tr>
<tr>
<td>L</td>
<td>L</td>
<td>.008</td>
<td>.784</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Theta_A</th>
<th>Theta_B</th>
<th>X2: 1</th>
<th>X2: 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>H</td>
<td>.011</td>
<td>.012</td>
</tr>
<tr>
<td>H</td>
<td>L</td>
<td>.088</td>
<td>.098</td>
</tr>
<tr>
<td>L</td>
<td>H</td>
<td>.088</td>
<td>.098</td>
</tr>
<tr>
<td>L</td>
<td>L</td>
<td>.008</td>
<td>.792</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>New Weight</th>
<th>Old Weight</th>
<th>Adjustment Factor = New/Old</th>
</tr>
</thead>
<tbody>
<tr>
<td>.012</td>
<td>.012</td>
<td>1.00</td>
</tr>
<tr>
<td>.088</td>
<td>.098</td>
<td>.90</td>
</tr>
<tr>
<td>.088</td>
<td>.098</td>
<td>.90</td>
</tr>
<tr>
<td>.008</td>
<td>.792</td>
<td>.01</td>
</tr>
</tbody>
</table>
7) Updating scheme

Clique 2

<table>
<thead>
<tr>
<th>New Weight</th>
<th>Old Weight</th>
<th>Adjustment Factor = New/Old</th>
</tr>
</thead>
<tbody>
<tr>
<td>.012</td>
<td>.012</td>
<td>1.00</td>
</tr>
<tr>
<td>.088</td>
<td>.098</td>
<td>.90</td>
</tr>
<tr>
<td>.088</td>
<td>.098</td>
<td>.90</td>
</tr>
<tr>
<td>.008</td>
<td>.792</td>
<td>.01</td>
</tr>
</tbody>
</table>

Belief updating (i.e., multiplication by adjustment factors)

<table>
<thead>
<tr>
<th>( \theta_A )</th>
<th>( \theta_B )</th>
<th>( X_2: 1 )</th>
<th>( X_2: 0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>H</td>
<td>.011 \times 1.00</td>
<td>.001 \times 1.00</td>
</tr>
<tr>
<td>H</td>
<td>L</td>
<td>.005 \times .90</td>
<td>.093 \times .90</td>
</tr>
<tr>
<td>L</td>
<td>H</td>
<td>.088 \times .90</td>
<td>.010 \times .90</td>
</tr>
<tr>
<td>L</td>
<td>L</td>
<td>.008 \times .01</td>
<td>.784 \times .01</td>
</tr>
</tbody>
</table>

After belief updating

<table>
<thead>
<tr>
<th>( \theta_A )</th>
<th>( \theta_B )</th>
<th>( X_2: 1 )</th>
<th>( X_2: 0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>H</td>
<td>.011</td>
<td>.001</td>
</tr>
<tr>
<td>H</td>
<td>L</td>
<td>.004</td>
<td>.084</td>
</tr>
<tr>
<td>L</td>
<td>H</td>
<td>.080</td>
<td>.009</td>
</tr>
<tr>
<td>L</td>
<td>L</td>
<td>.000</td>
<td>.008</td>
</tr>
</tbody>
</table>

After normalizing

<table>
<thead>
<tr>
<th>( \theta_A )</th>
<th>( \theta_B )</th>
<th>( X_2: 1 )</th>
<th>( X_2: 0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>H</td>
<td>.056</td>
<td>.005</td>
</tr>
<tr>
<td>H</td>
<td>L</td>
<td>.020</td>
<td>.426</td>
</tr>
<tr>
<td>L</td>
<td>H</td>
<td>.406</td>
<td>.046</td>
</tr>
<tr>
<td>L</td>
<td>L</td>
<td>.000</td>
<td>.041</td>
</tr>
</tbody>
</table>

Apply the adjustment factor for each row,

then renormalize with respect to all values.
7) Updating scheme

Apply the adjustment factor for each row,
then renormalize with respect to all values.

Predictive distribution for $X_2$
Comments

• Finding the optimal triangulation is NP-hard—you need heuristics
• Netica will tell you cliques if you ask it
• Computation depends on *tree width* of junction tree, or largest clique size (large potential tables)
• More conditional independence is generally better
Some Favorable & Unfavorable Structures

• Which graph is better?
  Why?
Some Favorable & Unfavorable Structures

- Multiple children are good (think IRT)
  *Multiple simple cliques*

- Multiple parents are not good. Why?
  *Moralization forces a clique containing all parents & the child.*
Key Points for Measurement Models

• Student model (SM) variables are of transcending interest.
  ▪ They characterize student knowledge, skill, strategies
  ▪ Cannot be directly observed
• Observable variables are means of getting evidence about SM variables.
  ▪ They characterize salient aspects of performance
• Observable variables from performances are modeled as conditionally independent across (but not necessarily within) tasks, given SM variables.