Bayesian Networks in Educational Assessment Tutorial

Session II: Bayes Net Applications
ACED: ECD in Action

Duanli Yan, Diego Zapata, ETS
Russell Almond, FSU
Roy Levy, ASU

Agenda

SESSION	TOPIC	PRESENTERS
Session 1:	Evidence Centered Design Bayesian Networks	Duanli Yan & Diego Zapata
Session 2:	Bayes Net Applications ACED: ECD in Action	Duanli Yan & Diego Zapata
Session 3:	Refining Bayes Nets with Data	Russell Almond & Roy Levy
Session 4:	Refining Bayes Nets with Data	Russell Almond & Roy Levy

1. Discrete Item Response Theory (IRT)

- Proficiency Model
- Task/Evidence Models
- Assembly Model
- Some Numbers

IRT Proficiency Model

- There is one proficiency varaible, θ. (Sometimes called an "ability parameter", but we reserve the term *parameter* for quantites which are not person specific.)
- θ takes on values {-2, -1, 0, 1, 2} with prior probabilities of (0.1, 0.2, 0.4, 0.2, 0.1) (Triangular distribution).
- Observable outcome variables are all independent given θ
- Goal is to draw inferences about θ
  - Rank order students by θ
  - Classify students according to θ above or below a cut point

IRT Task/Evidence Model

- Tasks yield an work product which can be unambiguously scored right/wrong.
- Each task has a *single* observable outcome variable.
- *Tasks* are often called *items*, although the common usage often blurs the distinction between the presentation of the item and the outcome variable.

IRT (Rasch) Evidence Model

- Let X_j be observable outcome variable from Task j
- $P(X_j = right \mid \theta, \beta_j) = \frac{1}{1+e^{-(\theta-\beta_j)}}$
  β_j is the *difficulty* of the item.
- Can crank through the formula for each of the five values of θ to get values for Conditional Probability Tables (CPT)

IRT Assembly Model

- 5 items
- Increasing difficulty:
  $\beta \in \{-1.5, -0.75, 0, 0.75, 1.5\}$.
- Adaptive presentation of items

Conditional Probability Tables

θ	Prior	Item 1	Item 2	Item 3	Item 4	Item 5
-2	0.1	0.3775	0.2227	0.1192	0.0601	0.0293
-1	0.2	0.6225	0.4378	0.2689	0.1480	0.0759
0	0.4	0.8176	0.6792	0.5000	0.3208	0.1824
1	0.2	0.9241	0.8520	0.7311	0.5622	0.3775
2	0.1	0.9707	0.9399	0.8088	0.7773	0.6225

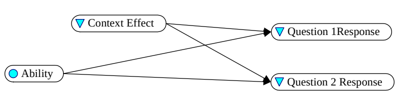
Problems Set 1

1. Assume θ=1, what is expected score (sum X_j)
2. Calculate $P(\theta\,|X_1$=<u>right</u>$)$, $E(\theta\,|X_1$=<u>right</u>$)$
3. Calculate $P(\theta\,|X_5$=<u>right</u>$)$, $E(\theta\,|X_5$=<u>right</u>$)$
4. Score three students who have the following observable patterns (Tasks 1--5):
   1,1,1,0,0
   1,0,0,1,1
   1,1,1,0,1
5. Suppose we have observed for a given student X_2=<u>right</u> and X_3=<u>right</u>, what is the next best item to present (hint, look for expected probabilities closest to .5,.5
6. Same thing, with X_2=<u>right</u> and X_3=<u>wrong</u>
7. Same thing, with X_2=<u>wrong</u> and X_3=<u>wrong</u>

2. "Context" effect --Testlets

- Standard assumption of conditional independence of observable variables given Proficiency Variables
- Violation
  - Shared stimulus
  - Context
  - Special knowledge
  - Shared Work Product
  - Sequential dependencies
  - Scoring Dependencies (Multi-step problem)
- Testlets (Wainer & Kiely, 1987)
- Violation results in overestimating the evidential value of observables for Proficiency Variables

"Context" effect -- Variables



- Context variable – A parent variable introduced to handle conditional dependence among observables (testlet)
  - Consistent with Stout's (1987) 'essential n-dimensionality'
  - Wang, Bradlow & Wainer (2001) SCORIGHT program for IRT
  - Patz & Junker (1999) model for multiple ratings

"Context" effect -- example

- Suppose that Items 3 and 4 share common presentation material
- Example: a word problem about "Yacht racing" might use nautical jargon like "leeward" and "tacking"
- People familiar with the content area would have an advantage over people unfamiliar with the content area.
- Would never us this example in practice because of DIF (Differential Item Functioning)

Adding a context variable

- Group Items 3 and 4 into a single task with two observed outcome variables
- Add a person-specific, task-specific latent variable called "context" with values <u>familiar</u> and <u>unfamiliar</u>
- Estimates of θ will "integrate out" the context effect
- Can use as a mathematical trick to force dependencies between observables.

IRT Model with Context Variable

Problem Set 2

- Compare the following quantities in the context and no context models:
  1. $P(X_2)$, $P(X_3)$, $P(X_4)$
  2. $P(\theta|X_2=\underline{right})$, $P(\theta|X_3=\underline{right})$
  3. $P(X_4|X_2=\underline{right})$, $P(X_4|X_3=\underline{right})$
  4. $P(\theta|X_3=\underline{wrong}, X_4=\underline{wrong})$, $P(\theta|X_3=\underline{right}, X_4=\underline{wrong})$,
  5. $P(\theta|X_3=\underline{wrong}, X_4=\underline{right})$, $P(\theta|X_3=\underline{right}, X_4=\underline{right})$

Context Effect Postscript

- If Context effect is generally construct-irrelevant variance, if correlated with group membership this is bad (DIF)
- When calibrating using 2PL IRT model, can get similar joint distribution for θ, X_3, and X_4 by decreasing the discrimination parameter

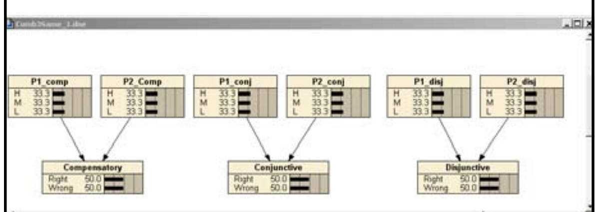
3. Combination Models

Consider a task which requires two Proficiencies:

Three different ways to combine those proficiencies:

- **Compensatory**: More of Proficiency 1 compensates for less of Proficiency 2. Combination rule is *sum*.
- **Conjunctive**: Both proficiencies are needed to solve the problem. Combination rule is *minimum*.
- **Disjunctive**: Two proficiencies represent alternative solution paths to the problem. Combination rule is *maximum*.

Combination Model Graphs



Common Setup for All Three Models

- There are two parent nodes, and both parents are conditionally independent of each other. The difference among the three models lies in the third term below:

$$P(P_1, P_2, X) = P(P_1) \bullet P(P_2) \bullet P(X \mid P_1, P_2)$$

- The priors for the parent nodes are the same for the three models with 0.3333 of probability at each of the H, M, and L states.
- The initial marginal probability for X is the same for the three models (50/50).

Conditional Probability Tables

This table contains the conditional probabilities for the parent nodes (P1 and P2) and the combination model for the three models.

Table 3 – Part 2

Conditional Problems for Compensatory, Conjunctive, and Disjunctive

P1	P2	Compensatory "Right"	Conjunctive "Right"	Disjunctive "Right"
H	H	0.9	0.9	0.7
H	M	0.7	0.7	0.7
H	L	0.5	0.3	0.7
M	H	0.7	0.7	0.7
M	M	0.5	0.7	0.3
M	L	0.3	0.3	0.3
L	H	0.5	0.3	0.7
L	M	0.3	0.3	0.3
L	L	0.1	0.3	0.1

Problem Set 3

1. Verify that $P(P_1)$, $P(P_2)$, and $P(Obs)$ are the same for all three models. (*Obs* represents either the node *Compensatory*, *Conjunctive*, or *Disjunctive*)
2. Assume *Obs*=right, Calculate $P(P_1)$ and $P(P_2)$ for all three models.
3. Assume *Obs*=wrong, Calculate $P(P_1)$ and $P(P_2)$ for all three models.
4. Assume *Obs*=right, and P_1 = H. Calculate $P(P_2)$ for all three models.
5. Assume *Obs*=right, and P_1 = M. Calculate $P(P_2)$ for all three models.
6. Assume *Obs*=right, and P_1 = L. Calculate $P(P_2)$ for all three models.
7. Explain the differences

Activity 3

- Go back to the Driver's License Exam you built in Session I and add some numbers
- Now put in some observed outcomes
  - How did the probabilities change?
  - Is that about what you expected?

ACED Background

- ACED (Adaptive Content with Evidence-based Diagnosis)
- Val Shute (PD), Aurora Graf, Jody Underwood, Eric Hansen, Peggy Redman, Russell Almond, Larry Casey, Waverly Hester, Steve Landau, Diego Zapata
- Domain: Middle School Math, Sequences
- Project Goals:
  - Adaptive Task Selection
  - Diagnostic Feedback
  - Accessibility

ACED Features

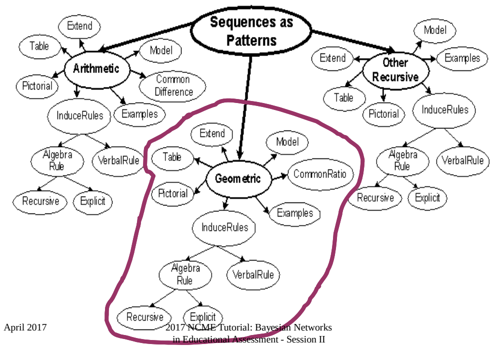
Valid Assessment. Based on evidence-centered design (ECD).

Adaptive Sequencing. Tasks presented in line with an adaptive algorithm.

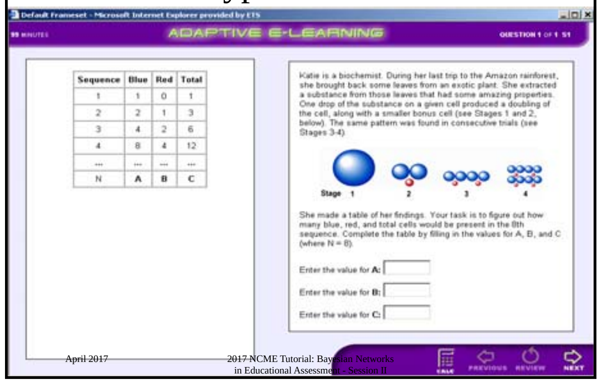
Diagnostic Feedback. Feedback is immediate and addresses common errors and misconceptions.

Aligned. Assessments aligned with (a) state and national standards and (b) curricula in current textbooks.

ACED Proficiency Model



Typical Task



ACED Design/Build Process

- Identify Proficiency variables
- Structure Proficiency Model
- Elicit Proficiency Model Parameters
- Construct Tasks to target proficiencies at Low/Medium/High difficulty
- Build Evidence Models based on difficulty/Q-Matrix

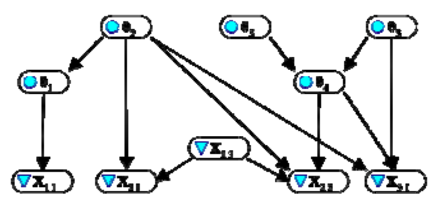
Parameterization of Network

- Proficiency Model:
  - Based on Regression model of child given parent
  - SME provided correlation and intercept
  - SME has low confidence in numeric values
- Evidence Model Fragment
  - Tasks Scored Right/Wrong
  - Based on IRT model
  - High/Medium/Low corresponds to θ = +1/0/-1
  - Easy/Medium/Hard corresponds to difficulty -1/0/+1
  - Discrimination of 1
  - Used Q-Matrix to determine which node is parent

PM-EM Algorithm for Scoring

- Master Bayes net with just proficiency model(PM)
- Database of Bayes net fragments corresponding to evidence models (EMs), indexed by task ID
- To score a task:
  - Find EM fragment corresponding to task
  - Join EM fragment to PM
  - Enter Evidence
  - Absorb evidence from EM fragment into network
  - Detach EM fragment

An Example



- Five proficiency variables
- Three tasks, with observables $\{X_{11}\}$, $\{X_{21}, X_{22}, X_{23}\}$, $\{X_{31}\}$.

Q: Which observables depend on which proficiency variables?

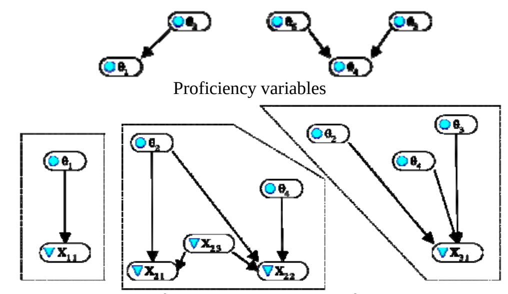
A: See the Q-matrix (Fischer, Tatsuoka).

	θ_1	θ_2	θ_3	θ_4	θ_5	X_{23}
X_{11}	1	0	0	0	0	--
X_{21}	0	1	0	0	0	1
X_{22}	0	1	0	1	0	1
X_{23}	0	0	0	0	0	N/A
X_{31}	0	1	1	1	0	--

Proficiency Model / Evidence Model Split

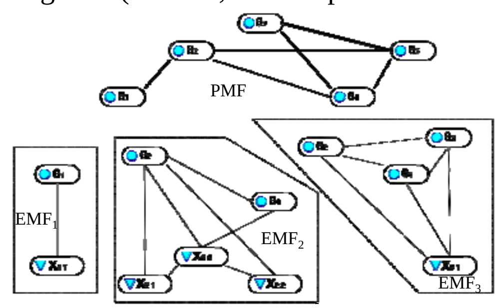
- Full Bayes net for proficiency model and observables for all tasks can be decomposed into fragments.
  - Proficiency model fragment(s) (PMFs) contain proficiency variables.
  - An evidence model fragment (EMF) for each task.
  - EMF contains observables for that task and all proficiency variables that are parents of any of them.
- Presumes observables are conditionally independent between tasks, but can be dependent within tasks.
- Allows for adaptively selecting tasks, docking EMF to PMF, and updating PMF on the fly.

On the way to PMF and EMFs…

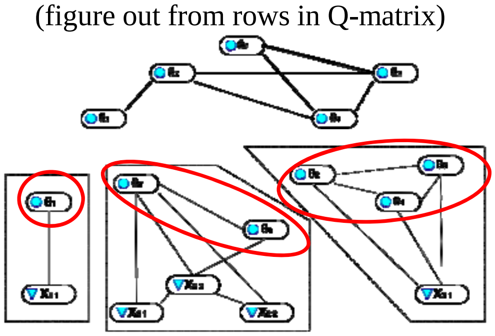


Observables and proficiency variable parents for the tasks

Marry parents, drop directions, and triangulate (in PMF, with respect to all tasks)



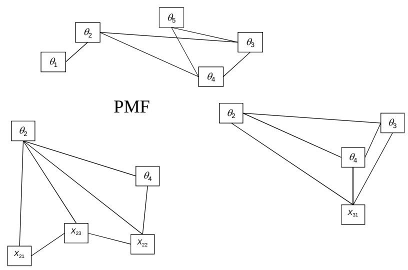
Footprints of tasks in proficiency model (figure out from rows in Q-matrix)



Result:

- Each EMF implies a join tree for Bayes net propagation.
  - Initial distributions for proficiency variables are uniform.
- The footprint of the PM in the EMF is a clique intersection between that EMF and the PMF.
- Can "dock" EMFs with PMF one-at-a-time, to …
  - absorb evidence from values of observables to that task as updated probabilities for proficiency variables, and
  - predict responses in new tasks, to evaluate potential evidentiary value of administering it.

Docking evidence model fragments



Scoring Exercise

Outcome	Task Name	Proficiency Variable	Difficulty
Wrong	tCommonRatio1a.xml	CommonRatio	Easy
Right	tCommonRatio2b.xml	CommonRatio	Medium
Wrong	tCommonRatio3b.xml	CommonRatio	Hard
Wrong	tExplicitGeometric1a.xml	ExplicitGoemetric	Easy
Right	tExplicitGeometric2a.xml	ExplicitGoemetric	Medium
Wrong	tExplicitGeometric3b.xml	ExplicitGoemetric	Hard
Wrong	tRecursiveRuleGeometric1a.xml	RecursiveRuleGeometric	Easy
Wrong	tRecursiveRuleGeometric2b.xml	RecursiveRuleGeometric	Medium
Wrong	tRecursiveRuleGeometric3a.xml	RecursiveRuleGeometric	Hard
Right	tTableExtendGeometric1a.xml	TableGeometric	Easy
Right	tTableExtendGeometric2b.xml	TableGeometric	Medium
Right	tTableExtendGeometric3a.xml	TableGeometric	Hard
Wrong	tVerbalRuleExtendModelGeometric1a.xml	VerbalRuleGeometric	Easy
Wrong	tVerbalRuleExtendModelGeometric1b.xml	VerbalRuleGeometric	Easy
Right	tVerbalRuleExtendModelGeometric2a.xml	VerbalRuleGeometric	Medium
Wrong	tVisualExtendGeometric1a.xml	VisualGeometric	Easy
Wrong	tVisualExtendGeometric2a.xml	VisualGeometric	Medium
Wrong	tVisualExtendGeometric3a.xml	VisualGeometric	Hard

Weight of Evidence

- Good (1985)
- *H* is binary hypothesis, e.g., *Proficiency* > Medium
- *E* is evidence for hypothesis
- Weight of Evidence (WOE) is

$$W(H:E)=\log\frac{\Pr(E|H)}{\Pr(E|\overline{H})}=\log\frac{\Pr(H|E)}{\Pr(\overline{H}|E)}-\log\frac{\Pr(H)}{\Pr(\overline{H})}$$

Properties of WOE

- "Centibans" (log base 10, multiply by 100)
- Positive for evidence supporting hypothesis, negative for evidence refuting hypothesis
- Movement in tails of distribution as important as movement near center
- Bayes theorem using log odds

Conditional Weight of Evidence

- Can define Conditional Weight of Evidence

$$W(H:E_2|E_1)=\log\frac{\Pr(E_2|H,E_1)}{\Pr(E_2|\overline{H},E_1)}$$

- Nice Additive properties

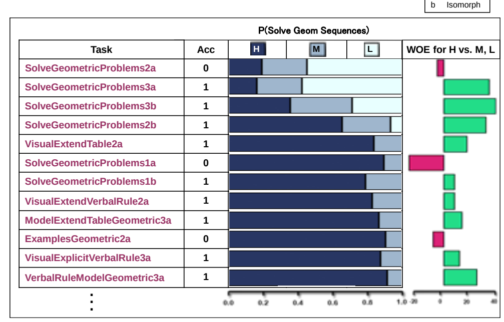
$$W(H:E_1,E_2)=W(H:E_1)+W(H:E_2|E_1)$$

- Order sensitive
- WOE Balance Sheet (Madigan, Mosurski & Almond, 1997)

Evidence Balance Sheet

63 tasks total

1 Easy
2 Medium
3 Hard
a Item type
b Isomorph



Expected Weight of Evidence

When choosing next "test" (task/item) look at expected value of WOE where expectation is taken wrt *P(E|H)*.

$$EW(H:E) = \sum_{j=1}^{n} W(H:e_j)\Pr(e_j \mid H)$$

where $\{e_j, j=1,...,n\}$ represent the possible results.

Calculating EWOE

Madigan and Almond (1996)

- Enter any observed evidence into net

1. Instantiate Hypothesis = True (may need to use virtual evidence if hypothesis is compound)
2. Calculate $\mathrm{P}(E_i|H)$ for each candidate item
3. Instantiate Hypothesis = False
4. Calculate $\mathrm{P}(E_i|\overline{H})$ for each candidate item

Related Measures

- Value of Information

$$\mathrm{VoI}(T) = \mathrm{E}_T\left[\max_d \mathrm{E}_{\mathbf{S}} u(d,\mathbf{S}) - \max_d \mathrm{E}_{\mathbf{S}|T} u(d,\mathbf{S})\right]$$

- **S** is proficiency state
- *d* is decision
- *u* is utility

Related Measures (2)

- Mutual Information
- Extends to non-binary hypothesis nodes

$$\sum_{x,y} \mathrm{P}(x,y) \log \frac{\mathrm{P}(x,y)}{\mathrm{P}(x)\mathrm{P}(y)}$$

- Kullback-Liebler distance between joint distribution and independence

$$\sum_{x} \mathrm{P}(x) \sum_{y} \mathrm{P}(y|x) \log \frac{\mathrm{P}(y|x)}{\mathrm{P}(y)}$$

Task Selection Exercise 1

- Use ACEDMotif1.dne
  - Easy, Medium, and Hard tasks for Common Ratio and Visual Geometric
- Use Hypothesis *SolveGeometricProblems* > Medium
- Calculate EWOE for six observables
- Assume candidate gets first item right and repeat
- Next assume candidate gets first item wrong and repeat
- Repeat exercise using hypothesis *SolveGeometricProblems* > Low

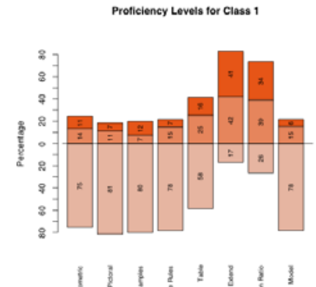
Task Selection Exercise 2

- Use Network ACEDMotif2.dne
- Select the *SolveGeometricProblems* node
- Run the program Network>Sensitivity to Findings
- This will list the Mutual information for all nodes
- Select the observable with the highest mutual information as the first task
- Use this to process a person who gets every task right
- Use this to process a person who gets every task wrong

ACED Evaluation

- Middle School Students
- Did not normally study geometric series
- Four conditions:
  - Elaborated Feedback/Adaptive (E/A; n=71)
  - Simple Feedback/Adaptive (S/A; n=75)
  - Elaborated Feedback/Linear (E/L; n=67)
  - Control (no instruction; n=55)
- Students given all 61 geometric items
- Also given pretest/posttest (25 items each)

ACED Scores



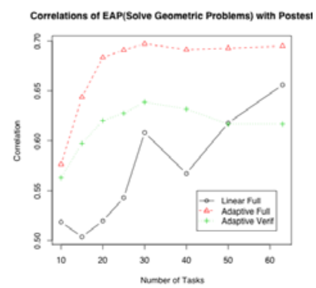
- For Each Proficiency Variable
  - Marginal Distribution
  - Modal Classification
  - EAP Score (High=1, Low=-1)

ACED Reliability

Proficiency (EAP)	Reliability
Solve Geometric Sequences (SGS)	*0.88*
Find Common Ratio	0.90
Generate Examples	0.92
Extend Sequence	0.86
Model Sequence	0.80
Use Table	0.82
Use Pictures	0.82
Induce Rules	0.78
Number Right	0.88

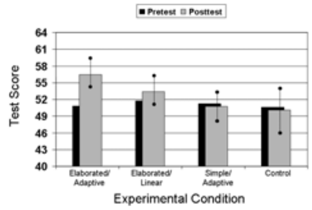
- Calculated with Split Halves (ECD design)
- Correlation of EAP score with posttest is 0.65 (close to reliability of posttest)
- Even with pretest forced into the equation, EAP score accounted for 17% unique variance
- Reliability of modal classifications was worse

Effect of Adaptivity



- For adaptive conditions, correlation with posttest seems to hit upper limit by 20 items
- Standard Error of Correlations is large
- Jump in linear case related to sequence of items

Effect of feedback



- E/A showed significant gains
- Others did not
- Learning and assessment reliability!!!!!

Acknowledgements

- Special thanks to Val Shute for letting us used ACED data and models in this tutorial.
- ACED development and data collection was sponsored by National Science Foundation Grant No. 0313202.
- Complete data available at: http://ecd.ralmond.net/ecdwiki/ACED/ACED