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
Unpublished work © 2002-2017

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 Bayesian Networks</td>
<td>Duanli Yan &amp; Diego Zapata</td>
</tr>
<tr>
<td>Session 2:</td>
<td>Bayes Net Applications ACED: ECD in Action</td>
<td>Duanli Yan &amp; Diego Zapata</td>
</tr>
<tr>
<td>Session 3:</td>
<td>Refining Bayes Nets with Data</td>
<td>Russell Almond &amp; Roy Levy</td>
</tr>
<tr>
<td>Session 4:</td>
<td>Refining Bayes Nets with Data</td>
<td>Russell Almond &amp; Roy Levy</td>
</tr>
</tbody>
</table>

1. Discrete Item Response Theory (IRT)

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

IRT Proficiency Model

- There is one proficiency variable, \( \theta \) (Sometimes called an “ability parameter”, but we reserve the term parameter for quantities which are not person specific.)
- \( \theta \) 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 \( \theta \)
- Goal is to draw inferences about \( \theta \)
  - Rank order students by \( \theta \)
  - Classify students according to \( \theta \) 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 \( j \)
- \[ P(X_j = \text{right} \mid \theta, \beta_j) = \frac{1}{1 + e^{-\beta_j}}, \]
  \( \beta_j \) is the difficulty of the item.
- Can crank through the formula for each of the five values of \( \theta \) to get values for Conditional Probability Tables (CPT)

Unpublished Work (c) 2002-2017 ETS
IRT Assembly Model

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

Problems Set 1

1. Assume \( \theta = 1 \), what is the expected score (sum \( X_j \))
2. Given \( \theta \) and \( X_i = \text{right} \), calculate \( P(\theta | X_i = \text{right}) \)
3. Given \( \theta \) and \( X_i = \text{right} \), calculate \( P(\theta | X_i = \text{right}) \)
4. Score three students who have the following observable patterns (Tasks 1--5):  
   1,1,1,0,0  
   1,0,0,1,1  
   1,1,0,1,1
5. Suppose we have observed for a given student \( X_5 = \text{right} \) and \( X_4 = \text{right} \), what is the next best item to present (hint, look for expected probabilities closest to .5, .5)
6. Same thing, with \( X_5 = \text{right} \) and \( X_4 = \text{wrong} \)
7. Same thing, with \( X_5 = \text{wrong} \) and \( X_4 = \text{wrong} \)

Conditional Probability Tables

<table>
<thead>
<tr>
<th>( \theta )</th>
<th>Prior</th>
<th>Item 1</th>
<th>Item 2</th>
<th>Item 3</th>
<th>Item 4</th>
<th>Item 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2</td>
<td>0.1</td>
<td>0.3775</td>
<td>0.2227</td>
<td>0.1192</td>
<td>0.0601</td>
<td>0.0293</td>
</tr>
<tr>
<td>-1</td>
<td>0.2</td>
<td>0.6225</td>
<td>0.4378</td>
<td>0.2689</td>
<td>0.1480</td>
<td>0.0759</td>
</tr>
<tr>
<td>0</td>
<td>0.4</td>
<td>0.8176</td>
<td>0.6792</td>
<td>0.5000</td>
<td>0.3208</td>
<td>0.1824</td>
</tr>
<tr>
<td>1</td>
<td>0.2</td>
<td>0.9241</td>
<td>0.8520</td>
<td>0.7311</td>
<td>0.5622</td>
<td>0.3775</td>
</tr>
<tr>
<td>2</td>
<td>0.1</td>
<td>0.9707</td>
<td>0.9399</td>
<td>0.8088</td>
<td>0.7773</td>
<td>0.6225</td>
</tr>
</tbody>
</table>

“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 use 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 familiar and unfamiliar
- Estimates of 0 will “integrate out” the context effect
- Can use as a mathematical trick to force dependencies between observables.

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. \( P(\theta | X_2 = \text{right}) \), \( P(\theta | X_3 = \text{right}) \)
  3. \( P(X_4 | X_2 = \text{right}) \), \( P(X_4 | X_3 = \text{right}) \)
  4. \( P(\theta | X_3 = \text{wrong}, X_4 = \text{wrong}) \), \( P(\theta | X_3 = \text{right}, X_4 = \text{wrong}) \)
  5. \( P(\theta | X_3 = \text{wrong}, X_4 = \text{right}) \), \( P(\theta | X_3 = \text{right}, X_4 = \text{right}) \)

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 \( \theta \), \( X_3 \), 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 \( \text{sum} \).
- **Conjunctive:** Both proficiencies are needed to solve the problem.
  Combination rule is \( \text{minimum} \).
- **Disjunctive:** Two proficiencies represent alternative solution paths to the problem.
  Combination rule is \( \text{maximum} \).
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:
  \[ P(P_1, P_2, X) = P(P_1) \cdot P(P_2) \cdot P(X \mid P_1, P_2) \]
- 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 all 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>
<thead>
<tr>
<th>P1</th>
<th>P2</th>
<th>Compensatory</th>
<th>Conjunctive</th>
<th>Disjunctive</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>H</td>
<td>“Right”</td>
<td>“Right”</td>
<td>“Right”</td>
</tr>
<tr>
<td>H</td>
<td>M</td>
<td>0.9</td>
<td>0.9</td>
<td>0.7</td>
</tr>
<tr>
<td>H</td>
<td>L</td>
<td>0.7</td>
<td>0.7</td>
<td>0.7</td>
</tr>
<tr>
<td>M</td>
<td>H</td>
<td>0.5</td>
<td>0.5</td>
<td>0.3</td>
</tr>
<tr>
<td>M</td>
<td>M</td>
<td>0.7</td>
<td>0.7</td>
<td>0.3</td>
</tr>
<tr>
<td>M</td>
<td>L</td>
<td>0.5</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>L</td>
<td>H</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>L</td>
<td>M</td>
<td>0.5</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>L</td>
<td>L</td>
<td>0.1</td>
<td>0.3</td>
<td>0.1</td>
</tr>
</tbody>
</table>

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 \( P(P_1) \), \( P(P_2) \), and \( P(Obs) \) for all three models.
3. Assume Obs = wrong. Calculate \( P(P_1) \), \( P(P_2) \), and \( P(Obs) \) 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

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 $\theta = +1/0/-1$
  - 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).

<table>
<thead>
<tr>
<th>θ₁</th>
<th>θ₂</th>
<th>θ₃</th>
<th>θ₄</th>
<th>θ₅</th>
<th>X₂₃</th>
</tr>
</thead>
<tbody>
<tr>
<td>X₁₁</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>X₂₁</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>X₂₂</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>X₂₃</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>N/A</td>
</tr>
<tr>
<td>X₃₁</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

**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…**

Proficiency variables

Observables and proficiency variable parents for the tasks

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

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

PMF

EMF₁

EMF₂

EMF₃

**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.
Weight of Evidence

- Good (1985)
- $H$ is binary hypothesis, e.g., $\text{Proficiency} > \text{Medium}$
- $E$ is evidence for hypothesis
- Weight of Evidence (WOE) is

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

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
Expected Weight of Evidence

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

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

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

Calculating EWOE

Madigan and Almond (1996)

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

Related Measures

• Value of Information

$$Vol(T) = E_T \left[ \max_d E_{S,U}(d, S) - \max_d E_{S,U}(d, \overline{S}) \right]$$

• $S$ is proficiency state
• $d$ is decision
• $u$ is utility

Related Measures (2)

• Mutual Information

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

• Kullback-Liebler distance between joint distribution and independence

$$\sum_x \sum_y P(x)P(y) \log \frac{P(y|x)}{P(y)}$$

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

<table>
<thead>
<tr>
<th>Proficiency (EAP)</th>
<th>Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solve Geometric Sequences (SGS)</td>
<td>0.88</td>
</tr>
<tr>
<td>Find Common Ratio</td>
<td>0.90</td>
</tr>
<tr>
<td>Generate Examples</td>
<td>0.92</td>
</tr>
<tr>
<td>Extend Sequence</td>
<td>0.86</td>
</tr>
<tr>
<td>Model Sequence</td>
<td>0.80</td>
</tr>
<tr>
<td>Use Table</td>
<td>0.82</td>
</tr>
<tr>
<td>Use Pictures</td>
<td>0.82</td>
</tr>
<tr>
<td>Induce Rules</td>
<td>0.78</td>
</tr>
<tr>
<td>Number Right</td>
<td>0.88</td>
</tr>
</tbody>
</table>

- 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