Bayesian Networks in Educational Testing

Jiří Vomlel

Laboratory for Intelligent Systems Prague
University of Economics

This presentation is available at:
Contents:

• Educational testing is a “big business”.
• What is a fixed test and an adaptive test?
• An example: a test of basic operations with fractions.
• Optimal and myopically optimal tests.
• Construction of a myopically optimal fixed test.
• Results of experiments.
• An example showing that modeling dependence between skills is important.
• Conclusions.
Educational Testing Service (ETS)

- Educational Testing Service is the world’s largest private educational testing organization with 2,300 regular employees.

- Volumes for ETS’s Largest Exams in 2000-2001:
  
  3,185,000 SAT I Reasoning Test and SAT II: Subject Area Tests  
  (the SAT test is the standard college admission test in US)

  2,293,000 PSAT: Preliminary SAT/National Merit Scholarship Qualifying Test

  1,421,000 AP: Advanced Placement Program

  801,000 The Praxis Series: Professional Assessments for Beginning Teachers and Pre-Professional Skills Tests

  787,000 TOEFL: Test of English as a Foreign Language

  449,000 GRE: Graduate Record Examinations General Test

  etc.
Fixed Test vs. Adaptive Test
Computerized Adaptive Testing (CAT)

Objective: An optimal test for each examinee

Two basic steps: (1) examinee’s knowledge level is estimated
(2) questions appropriate for the level are selected.

R. Almond and R. Mislevy from ETS proposed to use graphical models in CAT.

- one student model (relations between skills, abilities, etc.)
- several evidence models, one for each task or question.
CAT for basic operations with fractions

Examples of tasks:

\[ T_1: \left( \frac{3}{4} \cdot \frac{5}{6} \right) - \frac{1}{8} = \frac{15}{24} - \frac{1}{8} = \frac{5}{8} - \frac{1}{8} = \frac{4}{8} = \frac{1}{2} \]

\[ T_2: \frac{1}{6} + \frac{1}{12} = \frac{2}{12} + \frac{1}{12} = \frac{3}{12} = \frac{1}{4} \]

\[ T_3: \frac{1}{4} \cdot 1\frac{1}{2} = \frac{1}{4} \cdot \frac{3}{2} = \frac{3}{8} \]

\[ T_4: \left( \frac{1}{2} \cdot \frac{1}{2} \right) \cdot \left( \frac{1}{3} + \frac{1}{3} \right) = \frac{1}{4} \cdot \frac{2}{3} = \frac{2}{12} = \frac{1}{6} \].
# Elementary and operational skills

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CP</strong></td>
<td>Comparison (common numerator or denominator)</td>
<td>$\frac{1}{2} &gt; \frac{1}{3}, \quad \frac{2}{3} &gt; \frac{1}{3}$</td>
</tr>
<tr>
<td><strong>AD</strong></td>
<td>Addition (common denom.)</td>
<td>$\frac{1}{7} + \frac{2}{7} = \frac{1+2}{7} = \frac{3}{7}$</td>
</tr>
<tr>
<td><strong>SB</strong></td>
<td>Subtract. (common denom.)</td>
<td>$\frac{2}{5} - \frac{1}{5} = \frac{2-1}{5} = \frac{1}{5}$</td>
</tr>
<tr>
<td><strong>MT</strong></td>
<td>Multiplication</td>
<td>$\frac{1}{2} \cdot \frac{3}{5} = \frac{3}{10}$</td>
</tr>
<tr>
<td><strong>CD</strong></td>
<td>Common denominator</td>
<td>$(\frac{1}{2}, \frac{2}{3}) = (\frac{3}{6}, \frac{4}{6})$</td>
</tr>
<tr>
<td><strong>CL</strong></td>
<td>Cancelling out</td>
<td>$\frac{4}{6} = \frac{2 \cdot 2}{2 \cdot 3} = \frac{2}{3}$</td>
</tr>
<tr>
<td><strong>CIM</strong></td>
<td>Conv. to mixed numbers</td>
<td>$\frac{7}{2} = \frac{3 \cdot 2 + 1}{2} = 3 \frac{1}{2}$</td>
</tr>
<tr>
<td><strong>CMI</strong></td>
<td>Conv. to improp. fractions</td>
<td>$3 \frac{1}{2} = \frac{3 \cdot 2 + 1}{2} = \frac{7}{2}$</td>
</tr>
</tbody>
</table>
## Misconceptions

<table>
<thead>
<tr>
<th>Label</th>
<th>Description</th>
<th>Occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAD</td>
<td>$\frac{a}{b} + \frac{c}{d} = \frac{a+c}{b+d}$</td>
<td>14.8%</td>
</tr>
<tr>
<td>MSB</td>
<td>$\frac{a}{b} - \frac{c}{d} = \frac{a-c}{b-d}$</td>
<td>9.4%</td>
</tr>
<tr>
<td>MMT1</td>
<td>$\frac{a}{b} \cdot \frac{c}{b} = \frac{a\cdot c}{b}$</td>
<td>14.1%</td>
</tr>
<tr>
<td>MMT2</td>
<td>$\frac{a}{b} \cdot \frac{c}{b} = \frac{a+c}{b\cdot b}$</td>
<td>8.1%</td>
</tr>
<tr>
<td>MMT3</td>
<td>$\frac{a}{b} \cdot \frac{c}{d} = \frac{a\cdot d}{b\cdot c}$</td>
<td>15.4%</td>
</tr>
<tr>
<td>MMT4</td>
<td>$\frac{a}{b} \cdot \frac{c}{d} = \frac{a\cdot c}{b+d}$</td>
<td>8.1%</td>
</tr>
<tr>
<td>MC</td>
<td>$\frac{a \cdot b}{c} = \frac{a\cdot b}{c}$</td>
<td>4.0%</td>
</tr>
</tbody>
</table>
Process that lead to the student model

- decision on what skills will be tested, preparation of paper tests
- paper tests given to students at Brønderslev high school, 149 students did the test.
- analysis of results, finding misconceptions, summarizing results into a data file,
- learning a Bayesian network model using the PC-algorithm and the EM-algorithm,
- attempts to explain some relations between skills and misconceptions using hidden variables,
- a new learning phase with hidden variables included, certain edges required to be part of the learned model.
Student model
Evidence model for task $T_1$

$$\left( \frac{3}{4} \cdot \frac{5}{6} \right) - \frac{1}{8} = \frac{15}{24} - \frac{1}{8} = \frac{5}{8} - \frac{1}{8} = \frac{4}{8} = \frac{1}{2}$$

$T_1 \iff MT \& CL \& ACL \& SB \& \neg MMT3 \& \neg MMT4 \& \neg MSB$
Student + Evidence model

[Diagram showing relationships between various elements labeled AD, SB, MAD, MSB, CMI, CIM, ACI, ACIM, ACL, ACD, CL, CD, MT, MMT1, MMT2, MMT3, MMT4, X1, and T1.]
Example of an adaptive test

Entropy of a probability distribution $P(S_i)$

$$H(P(S_i)) = - \sum_{s_i \in S} P(S_i = s_i) \cdot \log P(S_i = s_i)$$

Total entropy in a node $n$: $H(e_n) = \sum_{S_i \in S} H(P(S_i | e_n))$.

Expected entropy at the end of a test $t$ is

$$EH(t) = \sum_{\ell \in \mathcal{L}(t)} P(e_\ell) \cdot H(e_\ell).$$
Let $\mathcal{T}$ be the set of all possible tests. A test $t^*$ is **optimal** iff

$$t^* = \arg \min_{t \in \mathcal{T}} EH(t).$$

A **myopically optimal** test $t$ is a test where each question $X^*$ of $t$ minimizes the expected value of entropy after the question is answered:

$$X^* = \arg \min_{X \in \mathcal{X}} EH(t \downarrow X),$$

i.e. it works as if the test finished after the selected question $X^*$. 
Myopic construction of a fixed test

\[ e_{\text{list}} := [\emptyset]; \]
\[ \text{test} := []; \]
for \( i := 1 \) to \(|\mathcal{X}|\) do \( \text{counts}[i] := 0; \)
for position := 1 to test_length do
    \[ \text{new}_{e_{\text{list}}} := []; \]
    for all \( e \in e_{\text{list}} \) do
        \[ i := \text{most}_\text{informative}_X(e); \]
        \[ \text{counts}[i] := \text{counts}[i] + P(e); \]
        for all \( x_i \in X_i \) do
            \[ \text{append}(\text{new}_{e_{\text{list}}}, \{e \cup \{X_i = x_i\}\}); \]
        \[ e_{\text{list}} := \text{new}_{e_{\text{list}}}; \]
        \[ i^* := \text{arg}_i \text{max} \text{counts}[i]; \]
    \[ \text{append}(\text{test}, X_{i^*}); \]
    \[ \text{counts}[i^*] := 0; \]
\[ \text{return}(\text{test}); \]
Skill Prediction Quality

Quality of skill predictions vs. Number of answered questions

- Adaptive
- Average
- Descending
- Ascending
Total entropy of probability of skills

![Entropy on skills versus Number of answered questions graph with lines for adaptive, average, and descending/ascending order.]
Question Prediction Quality

![Chart showing the quality of question predictions vs. number of answered questions. The chart includes four lines representing adaptive, average, descending, and ascending methods.]
An example of a simple diagnostic task

Diagnosis of the absence or the presence of three skills

\[ S_1, S_2, S_3 \]

by use of a bank of three questions

\[ X_{1,2}, X_{1,3}, X_{2,3} \]

such that

\[
P(X_{i,j} = 1 | S_i = s_i, S_j = s_j) = \begin{cases} 
1 & \text{if } (s_i, s_j) = (1, 1) \\
0 & \text{otherwise.}
\end{cases}
\]

Assume answers to all questions from the item bank are wrong, i.e.

\[ X_{1,2} = 0, \; X_{1,3} = 0, \; X_{2,3} = 0. \]
Reasoning assuming skill independency

All skills are independent

\[ P(S_1, S_2, S_3) = P(S_1) \cdot P(S_2) \cdot P(S_3) \]

and \( P(S_i), i = 1, 2, 3 \) are uniform.

Then the probabilities for \( j = 1, 2, 3 \) are:

\[ P(S_j = 0 \mid X_{1,2} = 0, X_{1,3} = 0, X_{2,3} = 0) = 0.75, \]

i.e. we can not decide which skills are present and which are missing.
Modeling dependence between skills

with deterministic hierarchy

\[ S_1 \Rightarrow S_2, \ S_2 \Rightarrow S_3 \]

\[
\begin{align*}
P(S_1 = 0 \mid X_{1,2} = 0, X_{1,3} = 0, X_{2,3} = 0) &= 1 \\
P(S_2 = 0 \mid X_{1,2} = 0, X_{1,3} = 0, X_{2,3} = 0) &= 1 \\
P(S_3 = 0 \mid X_{1,2} = 0, X_{1,3} = 0, X_{2,3} = 0) &= 0.5
\end{align*}
\]

Observe, that for \( i = 1, 2, 3 \)

\[
P(S_i \mid X_{1,2} = 0, X_{1,3} = 0, X_{2,3} = 0) = P(S_i \mid X_{2,3} = 0), \ \text{i.e.}
\]

\( X_{2,3} = 0 \) gives the same information as \( X_{1,2} = 0, X_{1,3} = 0, X_{2,3} = 0 \).
Conclusions

- Empirical evidence shows that educational testing can benefit from application of Bayesian networks.
- Adaptive tests may substantially reduce the number of questions that are necessary to be asked.
- The new method for the design of a fixed test provided good results on tested data. It may be regarded as a good cheap alternative to computerized adaptive tests when they are not suitable.
- One theoretical problem related to application of Bayesian networks to educational testing is efficient inference exploiting deterministic relations in the model. This problem was addressed in our UAI 2002 paper.
... and this is the END.

It’s time to have a beer.

... or are there any questions?