A Latent Factor Model For Instructor Content Preference Analysis

Jack Wang
June 25, 2017
Why a Model for Analyzing *Instructors’* Preferences?

students

Prior work

This work

Instructors
Why a Model for Analyzing Instructors’ Preferences?

Leverage instructors’ experience and knowledge to enhance personalized learning systems.

This work
Context of “Content Preference”
Context of “Content Preference”

- What Content?
  - Students’ homework assignment questions that instructors exclude
Context of “Content Preference”

- What Content?
  - Students’ homework assignment questions that instructors exclude

- Why this content?
  - OpenStax Tutor automatically select homework questions from question corpus
  - Instructors can opt out questions they don’t like, so OpenStax system will NEVER select those questions for students
  - But there are too many questions!
Objective

**Predict** and **understand** instructors’ question exclusion preferences
Objective
Understand and predict instructors’ question exclusion preferences

Contribution
● Developed a latent factor model to predict instructors’ question preference
● Utilize Bloom’s Taxonomy tag of each question to explain instructors’ question preferences
Objective

Understand and predict instructors’ question exclusion preferences

Contribution

● Developed a latent factor model to predict instructors’ question preference
● Utilize Bloom’s Taxonomy tag of each question to explain instructors’ question preferences

Implication

● Short term: Automatically exclude questions for instructors (so that they don’t need to do this manually)
● Long term: Improve personalized question selection for students, taking into account instructors’ preferences
Bloom’s Taxonomy: Short Intro

Increasing cognitive complexity

- Remember
- Understand
- Apply
- Analyze
- Evaluate
- Create
Bloom’s Taxonomy: Short Intro

- Hierarchically classify each question into **ONE of SIX** levels
- Each question belongs to **ONE** Bloom’s level
Bloom’s Taxonomy: Short Intro

Hypothesis: instructors’ preferences for questions’ Bloom level tell us about their preferences for excluding certain questions

- Hierarchically classify each question into ONE of SIX levels
- Use a smaller set of factors to explain preferences for a lot more questions

Increasing cognitive complexity
Latent Factor Model: An Analogy

**Netflix**

<table>
<thead>
<tr>
<th></th>
<th>Matilda</th>
<th>Harry Potter</th>
<th>Home Alone</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>User 1</strong></td>
<td>⭐⭐⭐⭐⭐</td>
<td>⭐⭐⭐⭐⭐</td>
<td>?</td>
</tr>
<tr>
<td><strong>User 2</strong></td>
<td>⭐⭐⭐⭐⭐</td>
<td>?</td>
<td>⭐⭐⭐⭐⭐</td>
</tr>
<tr>
<td><strong>User 3</strong></td>
<td>⭐⭐⭐⭐⭐</td>
<td>⭐⭐⭐⭐⭐</td>
<td>⭐⭐⭐⭐⭐</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Latent Factor Model: An Analogy

**Netflix**

**Movie Genre**

<table>
<thead>
<tr>
<th>User 1</th>
<th>User 2</th>
<th>User 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>★★★★★</td>
<td>★★★★</td>
<td>★★★★★</td>
</tr>
<tr>
<td>★★★★</td>
<td>★????</td>
<td>★★★★★</td>
</tr>
<tr>
<td>★★★★</td>
<td>★★★★</td>
<td>★★★★★</td>
</tr>
</tbody>
</table>
Latent Factor Model: An Analogy

Netflix

Movie Genre

<table>
<thead>
<tr>
<th>User 1</th>
<th>User 2</th>
<th>User 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>★★★★☆</td>
<td>★★★☆</td>
<td>★★★★★</td>
</tr>
<tr>
<td>★★★☆</td>
<td>★★★☆</td>
<td>★★★★★</td>
</tr>
<tr>
<td>★★★★☆</td>
<td>★★★☆</td>
<td>★★★★★</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Instructor 1</th>
<th>question 1</th>
<th>question 2</th>
<th>question 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>?</td>
</tr>
<tr>
<td>Instructor 2</td>
<td>0</td>
<td>?</td>
<td>0</td>
</tr>
<tr>
<td>Instructor 3</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Latent Factor Model: An Analogy

**Movie Genre**

**NETFLIX**

1 => instructor opts out the question
0 => no preference

<table>
<thead>
<tr>
<th>User 1</th>
<th>User 2</th>
<th>User 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>⭐⭐⭐⭐⭐</td>
<td>?</td>
<td>⭐⭐⭐⭐⭐</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Instructor 1</th>
<th>question 1</th>
<th>question 2</th>
<th>question 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>?</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Instructor 2</th>
<th>question 1</th>
<th>question 2</th>
<th>question 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>?</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Instructor 3</th>
<th>question 1</th>
<th>question 2</th>
<th>question 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Latent Factor Model: An Analogy

Movie Genre

![Netflix posters](image)

Question characteristics (Bloom’s Taxonomy tag)

<table>
<thead>
<tr>
<th></th>
<th>question 1</th>
<th>question 2</th>
<th>question 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>User 1</td>
<td>[ratings]</td>
<td>?</td>
<td></td>
</tr>
<tr>
<td>User 2</td>
<td>[ratings]</td>
<td>?</td>
<td>[ratings]</td>
</tr>
<tr>
<td>User 3</td>
<td>[ratings]</td>
<td>[ratings]</td>
<td>[ratings]</td>
</tr>
<tr>
<td>Instructor 1</td>
<td>1</td>
<td>0</td>
<td>?</td>
</tr>
<tr>
<td>Instructor 2</td>
<td>0</td>
<td>?</td>
<td>0</td>
</tr>
<tr>
<td>Instructor 3</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Instructors’ Question Preference Data

- Biology textbook
- 20 instructors (represented by i)
- 896 questions (represented by j)
- Binary data matrix ($Y$) of dimension 20 x 896
  - $Y_{i,j} = 1$ => instructor opts out the question
  - $Y_{i,j} = 0$ => no preference
Latent Factor Model With Bloom’s Taxonomy Tag

\[ Y_{ij} \sim \text{Ber}(\phi(p_i^T a_j + g_i^T h_j)) \]
Latent Factor Model With Bloom’s Taxonomy Tag

\[ Y_{ij} \sim \text{Ber}\left(\phi(p_i^T a_j + g_i^T h_j)\right) \]

\( i \rightarrow \) instructor

\( j \rightarrow \) question
Latent Factor Model With Bloom’s Taxonomy Tag

\[ Y_{ij} \sim \text{Ber}(\phi(p_i^T a_j + g_i^T h_j)) \]

- \( a_j \): Question - Bloom’s tag association vector (obtained from question’s metadata)
- \( p_i \): Instructor - Bloom’s tag association vector

i \rightarrow \text{instructor}
j \rightarrow \text{question}
Latent Factor Model With Bloom’s Taxonomy Tag

\[ Y_{ij} \sim \text{Ber} \left( \phi \left( p_i^T a_j \right) + g_i^T h_j \right) \]

- \( a_j \): Question - Bloom’s tag association vector (obtained from question’s metadata)
- \( p_i \): Instructor - Bloom’s tag association vector

\( i \rightarrow \text{instructor} \)
\( j \rightarrow \text{question} \)
Latent Factor Model With Bloom’s Taxonomy Tag

\[ Y_{ij} \sim Ber\left(\phi(p_i^T a_j + g_i^T h_j)\right) \]

- \( Y_{ij} \): Response (0 or 1)
- \( a_j \): Question - Bloom’s tag association vector (obtained from question’s metadata)
- \( p_i \): Instructor - Bloom’s tag association vector

Increasing cognitive complexity:
- Remember
- Understand
- Apply
- Analyze
- Evaluate
- Create
We can now use instructors’ preferences for Bloom’s Taxonomy to account for their question preferences.

\[
Y_{ij} \sim \text{Ber}(\phi(p_i^T a_j + g_i^T h_j))
\]

- \(Y_{ij}\): Probability of instructor \(i\) choosing question \(j\).
- \(p_i\): Instructor’s Bloom’s tag association vector.
- \(a_j\): Question’s Bloom’s tag association vector.
- \(g_i\): Other factors affecting the choice.

Instructor - Bloom’s tag association vector

Question - Bloom’s tag association vector

We can now use instructors’ preferences for Bloom’s Taxonomy to account for their question preferences.
Latent Factor Model With Bloom’s Taxonomy Tag

We can now use instructors’ preferences for Bloom’s Taxonomy to account for their question preferences.

Expect: the larger the value in entry $P_{ij}$, the stronger instructor $i$ prefers to exclude question of Bloom $j$.

\[
Y_{ij} \sim \text{Ber}\left(\phi(p_i^T a_j + g_i^T h_j)\right)
\]

$a_j$ - Question - Bloom’s tag association vector (obtained from question’s metadata)

$p_i$ - Instructor - Bloom’s tag association vector
Evaluation

Prediction: How well does the model perform?

Interpretability: Why instructors exclude certain questions?
Evaluation

Prediction:
How well does the model perform?

Interpretability:
Why instructors exclude certain questions?
How well does the model perform?
How well does the model perform?

ACC = fraction of correct predictions
How well does the model perform?

**ACC** = fraction of correct predictions

**UBCF** = user-based collaborative filtering

**IBCF** = item-based collaborative filtering

**FSVD** = Funk singular value decomposition

(Popular methods used in recommendation systems, such as those used in Netflix, Amazon, etc.)
How well does the model perform?

**ACC** = fraction of correct predictions

**UBCF** = user-based collaborative filtering

**IBCF** = item-based collaborative filtering

**FSVD** = Funk singular value decomposition

(Popular methods used in recommendation systems, such as those used in Netflix, Amazon, etc.)
Evaluation

Prediction:
How well does the model perform?

Interpretability:
Why instructors exclude certain questions?
Why instructors exclude certain questions?

An explanation using the instructors’ Bloom’s Tag preferences:

Example of Bloom 1 (Remembering)
Question:

\[\beta \, 1-4 \text{ glycosidic linkages are found in } \underline{\text{_____}}.\]

A. cellulose
B. disaccharides
C. RNA
D. starch
Why instructors exclude certain questions?

An explanation using the instructors’ Bloom’s Tag preferences:

Example of **Bloom 1 (Remembering)**
Question:

β 1-4 glycosidic linkages are found in ____.
A. cellulose
B. disaccharides
C. RNA
D. starch

<table>
<thead>
<tr>
<th>Instructor ID</th>
<th>Value for Bloom 1 (first value in ( p_i ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.0583</td>
</tr>
<tr>
<td>5</td>
<td>0.4408</td>
</tr>
<tr>
<td>9</td>
<td>0.8255</td>
</tr>
</tbody>
</table>
Why instructors exclude certain questions?

An explanation using the instructors’ Bloom’s Tag preferences:

Example of Bloom 1 (Remembering)
Question:

\(\beta\) 1-4 glycosidic linkages are found in ____.
A. cellulose  
B. disaccharides  
C. RNA  
D. starch

<table>
<thead>
<tr>
<th>Instructor ID</th>
<th>Value for Bloom 1 (first value in (p_i))</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.0583</td>
</tr>
<tr>
<td>5</td>
<td>0.4408</td>
</tr>
<tr>
<td>9</td>
<td>0.8255</td>
</tr>
</tbody>
</table>

- Weak exclusion preference
- Moderate exclusion preference
- Strong exclusion preference
Why instructors exclude certain questions?

An explanation using the instructors’ Bloom’s Tag preferences:

Example of Bloom 5 (evaluating) Question:

Which of the following statements about chemiosmosis is false?

A. Free energy from redox reactions is used to pump protons across a membrane.
B. The integral membrane protein cytochrome synthase produces ATP.
C. Chemiosmosis generates 90% of the ATP made during aerobic glucose catabolism.

<table>
<thead>
<tr>
<th>Instructor ID</th>
<th>Value for Bloom 1 (first value in $p_i$)</th>
<th>Actually excluded?</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.0583</td>
<td>no</td>
</tr>
<tr>
<td>5</td>
<td>0.4408</td>
<td>no</td>
</tr>
<tr>
<td>9</td>
<td>0.8255</td>
<td>yes</td>
</tr>
</tbody>
</table>
Why instructors exclude certain questions?

An explanation using the instructors’ Bloom’s Tag preferences:

Example of Bloom 5 (evaluating)
Question:

*Which of the following statements about chemiosmosis is false?*

A. Free energy from redox reactions is used to pump protons across a membrane.

B. The integral membrane protein cytochrome synthase produces ATP.

C. Chemiosmosis generates 90% of the ATP made during aerobic glucose catabolism.
Why instructors exclude certain questions?

An explanation using the instructors’ Bloom’s Tag preferences:

Example of **Bloom 5 (evaluating)**

Question:

*Which of the following statements about chemiosmosis is false?*

A. Free energy from redox reactions is used to pump protons across a membrane.

B. The integral membrane protein cytochrome synthase produces ATP.

C. Chemiosmosis generates 90% of the ATP made during aerobic glucose catabolism.

<table>
<thead>
<tr>
<th>Instructor ID</th>
<th>Value for Bloom 5 (fifth value in $p_i$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.0746</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>0.5827</td>
</tr>
</tbody>
</table>
### Why instructors exclude certain questions?

An explanation using the instructors’ Bloom’s Tag preferences:

Example of **Bloom 5 (evaluating)**

**Question:**

*Which of the following statements about chemiosmosis is false?*

A. Free energy from redox reactions is used to pump protons across a membrane.

B. The integral membrane protein cytochrome synthase produces ATP.

C. Chemiosmosis generates 90% of the ATP made during aerobic glucose catabolism.

### Instructor ID

<table>
<thead>
<tr>
<th>Instructor ID</th>
<th>Value for Bloom 5 (fifth value in $p_i$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.0746</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>0.5827</td>
</tr>
</tbody>
</table>

- **Weak exclusion preference**
- **Strong exclusion preference**
- **Moderate exclusion preference**
Why instructors exclude certain questions?

An explanation using the instructors’ Bloom’s Tag preferences:

Example of **Bloom 5 (evaluating)**

**Question:**

*Which of the following statements about chemiosmosis is false?*

- A. Free energy from redox reactions is used to pump protons across a membrane.
- B. The integral membrane protein cytochrome synthase produces ATP.
- C. Chemiosmosis generates 90% of the ATP made during aerobic glucose catabolism.

<table>
<thead>
<tr>
<th>Instructor ID</th>
<th>Value for Bloom 5 (fifth value in $p_i$)</th>
<th>Actually excluded?</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.0746</td>
<td>no</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>9</td>
<td>0.5827</td>
<td>yes</td>
</tr>
</tbody>
</table>
Project summary

- Developed statistical model to characterize instructors’ question exclusion preference
  - Fairly accurately predicts whether an instructor prefers to exclude a question
  - Explains instructor’s question preferences in terms of Bloom’s Taxonomy
Project summary

- Developed statistical model to characterize instructors’ question exclusion preference
  - Fairly accurately predicts whether an instructor prefers to exclude a question
  - Explains instructor’s question preferences in terms of Bloom’s Taxonomy

Future work

- Include additional factors other than Bloom’s Taxonomy
- Extend the analysis to additional learning resources other than question exclusions
Acknowledgements

Richard Baraniuk
Andrew S. Lan
Phillip Grimaldi
Backup Slides
Bloom’s Taxonomy

- **Remember**
  - Recall facts and basic concepts
    - define, duplicate, list, memorize, repeat, state

- **Understand**
  - Explain ideas or concepts
    - classify, describe, discuss, explain, identify, locate, recognize, report, select, translate

- **Apply**
  - Use information in new situations
    - execute, implement, solve, use, demonstrate, interpret, operate, schedule, sketch

- **Analyze**
  - Draw connections among ideas
    - differentiate, organize, relate, compare, contrast, distinguish, examine, experiment, question, test

- **Evaluate**
  - Justify a stand or decision
    - appraise, argue, defend, judge, select, support, value, critique, weigh

- **Create**
  - Produce new or original work
    - design, assemble, construct, conjecture, develop, formulate, author, investigate

Vanderbilt University Center for Teaching
Latent Factor Model With Bloom’s Taxonomy Tag

\[ Y_{ij} \sim \text{Ber}(\phi(p_i^T a_j + g_i^T h_j)) \]
Latent Factor Model With Bloom’s Taxonomy Tag

\[ Y_{ij} \sim \text{Ber} \left( \phi \left( p_i^T a_j + g_i^T h_j \right) \right) \]

The binary data matrix of instructor \( i \)'s preference for question \( j \)

\( i \rightarrow \text{instructor} \)
\( j \rightarrow \text{question} \)
Motivated by the fact that data matrix is in binary form

\[ Y_{ij} \sim \text{Ber}(\phi(p_i^T a_j + g_i^T h_j)) \]

The binary data matrix of instructor \( i \)'s preference for question \( j \)

\( i \rightarrow \text{instructor} \)
\( j \rightarrow \text{question} \)
Latent Factor Model With Bloom’s Taxonomy Tag

\[ Y_{ij} \sim \text{Ber} \left( \phi \left( \mathbf{p}_i^T \mathbf{a}_j + \mathbf{g}_i^T \mathbf{h}_j \right) \right) \]

Motivated by the fact that data matrix is in binary form

The binary data matrix of instructor \( i \)'s preference for question \( j \)

Latent factors

\( i \rightarrow \text{instructor} \)
\( j \rightarrow \text{question} \)
The binary data matrix of instructor $i$'s preference for question $j$ is motivated by the fact that the data matrix is in binary form.

$$Y_{ij} \sim \text{Ber}(\phi(p_i^T a_j + g_i^T h_j))$$

The binary data matrix of instructor $i$'s preference for question $j$.

Sigmoid function to shrink latent factors to values in $[0, 1]$.

$$\phi(x) = \frac{1}{1 + e^{-x}}$$

$i \rightarrow$ instructor

$j \rightarrow$ question

Latent Factor Model With Bloom’s Taxonomy Tag
Latent Factor Model With Bloom’s Taxonomy Tag

Motivated by the fact that data matrix is in binary form

The binary data matrix of instructor $i$'s preference for question $j$

Factors associated with Bloom’s Taxonomy tag

Latent factors

Sigmoid function to shrink latent factors to values in $[0, 1]$

$$Y_{ij} \sim \text{Ber} \left( \phi \left( p_i^T a_j + g_i^T h_j \right) \right)$$

$$\phi(x) = \frac{1}{1 + e^{-x}}$$

$i \rightarrow$ instructor

$j \rightarrow$ question
Latent Factor Model With Bloom’s Taxonomy Tag

\[ Y_{ij} \sim \text{Ber} \left( \phi \left( p_i^T a_j + g_i^T h_j \right) \right) \]

Motivated by the fact that data matrix is in binary form

Factors associated with Bloom’s Taxonomy tag

The binary data matrix of instructor \( i \)’s preference for question \( j \)

Latent factors

All other factors

Sigmoid function to shrink latent factors to values in \([0, 1]\)

\[ \phi(x) = \frac{1}{1 + e^{-x}} \]

i \rightarrow \text{instructor}

j \rightarrow \text{question}
\[
\minimize_{P, G, H} \quad f(P, G, H)
\]
\[
f(P, G, H) = \sum_{i=1}^{N} \sum_{j=1}^{Q} \log \left( 1 + \exp \left( -(p_i^T a_j + g_i^T h_j) \right) \right)
\]
\[
+ \frac{\lambda}{2} \sum_{i=1}^{N} \|p_i\|_2^2 + \frac{\gamma}{2} \sum_{i=1}^{N} \|g_i\|_2^2 + \frac{\eta}{2} \sum_{j=1}^{Q} \|h_j\|_2^2
\]

\[
p_i^{\text{new}} = p_i^{\text{old}} - \delta \frac{\partial}{\partial p_i} f(p_i^{\text{old}}, g_i^{\text{old}}, h_j^{\text{old}})
\]
\[
g_i^{\text{new}} = g_i^{\text{old}} - \delta \frac{\partial}{\partial g_i} f(p_i^{\text{new}}, g_i^{\text{old}}, h_j^{\text{old}})
\]
\[
h_j^{\text{new}} = h_j^{\text{old}} - \delta \frac{\partial}{\partial h_j} f(p_i^{\text{new}}, g_i^{\text{new}}, h_j^{\text{old}})
\]

\[
\frac{\partial}{\partial p_i} f(p_i, g_i, h_j) = -\sum_{j=1}^{Q} \frac{a_j}{1 + e^{-(p_i^T a_j + g_i^T h_j)}} + \lambda p_i
\]
\[
\frac{\partial}{\partial g_i} f(p_i, g_i, h_j) = -\sum_{j=1}^{Q} \frac{h_j}{1 + e^{-(p_i^T a_j + g_i^T h_j)}} + \gamma g_i
\]
\[
\frac{\partial}{\partial h_j} f(p_i, g_i, h_j) = -\sum_{i=1}^{N} \frac{g_i}{1 + e^{-(p_i^T a_j + g_i^T h_j)}} + \eta h_j
\]
Why instructors exclude certain questions?

Because they have various preferences for Bloom’s Taxonomy tags

\[ Y_{ij} \sim \text{Ber}(\phi(p_i^T a_j + g_i^T h_j)) \]

Each point: \( p_i \) Instructor - Bloom’s tag association vector

Project each vector into 2D space using multidimensional scaling
Why instructors exclude certain questions?

Because they have various preferences for Bloom’s Taxonomy tags

\[ Y_{ij} \sim \text{Ber}(\phi(p_i^T a_j + g_i^T h_j)) \]

Each point: \( p_i \) Instructor - Bloom’s tag association vector

Project each vector into 2D space using multidimensional scaling