People-focused Reinforcement Learning

Emma Brunskill
Want autonomous agents that act well (make good sequences of decisions)
Most decision making algorithms developed for robots*
Want algorithms to enable:
Agents Making Decisions as They Interact with People
Agents Making Decisions as Interact with People
Data = People
The John Kasich record
Yahoo Politics looks at what's behind the launch of the 16th GOP presidential primary camp ... Read More »
User news article recommendation

The John Kasich record
Yahoo Politics looks at what's behind the launch of the 16th GOP presidential primary camp ...
Read More »

Kerry says Iran vow to defy U.S. is 'very disturbing'

Toshiba CEO quits over accounting scandal
Recommending to a New User
Set of Articles Could Recommend

a new user
Set of Articles Could Recommend

- Sports: [1,0,0]
- Politics: [0,1,0]
- Science: [0,0,1]

a new user
Set of Articles Could Recommend

Generally Could Be Many Features

- [1,0,0] Sports
- [0,1,0] Politics
- [0,0,1] Science

a new user
Recommend an Article

$t = 1$

a new user
Sports
User Clicks

\[ t=1 \]

a new user  Sports

click!
Recommend Another Article

a new user

Sports

t=1

click!

Science

t=2

Recommend Another Article

a new user

Sports

t=1

click!

Science

t=2

no click
(or swipe past)
Recommend An Article

a new user

Sports

click!

t=1

Science

no click

t=2

Sports

click!

t=3
Recommend An Article
Goal: Maximize Expected Sum of Clicks

a new user
Sports
click!

Science
no click

Sports
click!
Recommend An Article
Goal: Maximize Expected Sum of Clicks
If Knew User Interests Choose Article Most Likely to Click

a new user

<table>
<thead>
<tr>
<th>t=1</th>
<th>t=2</th>
<th>t=3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sports</td>
<td>Science</td>
<td>Sports</td>
</tr>
<tr>
<td>click!</td>
<td>no click</td>
<td>click!</td>
</tr>
</tbody>
</table>
Cold Start Recommendation
Goal: Maximize Expected Sum of Clicks
Don’t Know User Interests

a new user

Sports

click!

t=1

Science

no click

t=2

Sports

click!

t=3
Cold Start Recommendation
Contextual Multi-armed Bandit
Exploration vs Exploitation

a new user

Sports

t=1

click!

Science

t=2

no click

Sports

t=3

click!
Problem: Trying Out Things is Expensive!
Poor Recommendations Might Cause Users to Leave

a new user

Sports

t=1

Science

t=2

Sports

t=3

click!

no click

click!
But Doing Cold-start for Many New Users: Sequence of contextual bandit problems

new user A  click!  no click  click!
But Doing Cold-start for Many New Users: Sequence of contextual bandit problems

new user A

new user B
But Doing Cold-start for Many New Users: Sequence of contextual bandit problems

new user A

new user B

new user C

click!

no click

no click

no click

click!

click!

no click

click!
Transfer / Lifelong / Multitask Learning

new user A
click!
no click
click!

new user B
no click
no click
click!

new user C
click!
no click
click!
But Doing Cold-start for Many New Users: Sequence of contextual bandit problems

Can we use sequence of users’ data to accelerate the cold-start process of future new users?
Cold-start for More Than One User: Learn Separately Per User

Fully Personalized Models
Cold-start for More Than One User: Share Data Across All User

- Fully Personalized Models
- Single Group Model
New Idea: Latent Contextual Bandits
1 of M “Types” of Users
Improving Cold Start Performance

Train/re-train a set of latent models using users’ data → Leverage the learned latent models to make recommendations for a new user
Cluster Users & Learn Policies Per Cluster

Model 1 \rightarrow Policy 1 \quad \text{pick sports!}

Model 2 \rightarrow Policy 2 \quad \text{pick politics!}

Model 3 \rightarrow Policy 3 \quad \text{pick science!}
Leverage the Learned Latent Models for New Users

<table>
<thead>
<tr>
<th>Policy</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.33</td>
</tr>
<tr>
<td>2</td>
<td>0.33</td>
</tr>
<tr>
<td>3</td>
<td>0.33</td>
</tr>
</tbody>
</table>

a new user
Leverage the Learned Latent Models for New Users

<table>
<thead>
<tr>
<th>Policy</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.33</td>
</tr>
<tr>
<td>2</td>
<td>0.33</td>
</tr>
<tr>
<td>3</td>
<td>0.33</td>
</tr>
</tbody>
</table>

Policy 1
Policy 2
Policy 3

a new user
No click
Leverage the Learned Latent Models for New Users

a new user  No click

<table>
<thead>
<tr>
<th>Policy</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.33</td>
</tr>
<tr>
<td>2</td>
<td>0.33</td>
</tr>
<tr>
<td>3</td>
<td>0.33</td>
</tr>
</tbody>
</table>
Updated Probability over Types

- **a new user**
- **No click**

<table>
<thead>
<tr>
<th>Policy</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td>2</td>
<td>0.4</td>
</tr>
<tr>
<td>3</td>
<td>0.4</td>
</tr>
</tbody>
</table>
Make Another Recommendation

Policy 1
Policy 2
Policy 3

Policy | Weight
---|---
1 | 0.2
2 | 0.4
3 | 0.4

a new user | No click | click
Updated Probability over Types

<table>
<thead>
<tr>
<th>Policy</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td>2</td>
<td>0.4</td>
</tr>
<tr>
<td>3</td>
<td>0.4</td>
</tr>
</tbody>
</table>

a new user  No click  click
Updated Probability over Types

<table>
<thead>
<tr>
<th>Policy</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1</td>
</tr>
<tr>
<td>2</td>
<td>0.2</td>
</tr>
<tr>
<td>3</td>
<td>0.7</td>
</tr>
</tbody>
</table>
Can Much More Quickly Narrow in On User Type!

<table>
<thead>
<tr>
<th>Policy</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1</td>
</tr>
<tr>
<td>2</td>
<td>0.2</td>
</tr>
<tr>
<td>3</td>
<td>0.7</td>
</tr>
</tbody>
</table>
Turning a Bandit Problem Into an Active Classification

<table>
<thead>
<tr>
<th>Policy</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1</td>
</tr>
<tr>
<td>2</td>
<td>0.2</td>
</tr>
<tr>
<td>3</td>
<td>0.7</td>
</tr>
</tbody>
</table>
Use Tensor Decomposition or Dirichlet Processes to Learn Latent Models

Train/re-train a set of latent models using users’ data

Leverage the learned latent models to make recommendations for a new user
Experiments

- Front page news dataset provided by Yahoo! as part of CMU-Yahoo! Inmind project
- 500,000 users’ click data
- 21 news categories as article features
Empirical results of the latent contextual bandit algorithm

- Consider first 20 interactions per user
- After that not really a new user
Relative CTR of LCB is about 10% higher than baseline
Learned Latent Models
*Required Counterfactual/Offline Estimation

- Didn’t actually get to run experiment with 500,000 users
- But had access to those
- Leveraged that data to evaluate offline to get unbiased estimate of how an alternative algorithm would have done
- See papers (AAAI 2015, AAAI 2016, ICML 2016)
Summary

- Use partial personalization to speed cold start problem (IJCAI 2016)

- Sequential decision making & RL to maximize lifetime benefits (to customers, patients, etc)

- ebrun@cs.stanford.edu