Unbiased Offline Evaluation of Contextual-bandit-based News Article Recommendation Algorithms

Lihong Li
Wei Chu
John Langford
Xuanhui Wang

Yahoo! Labs

WSDM 2011, Hong Kong
Yahoo-User Interaction

**CONTRIBUTION**

Unbiased offline evaluation for this interactive process

**CONTEXT**

Unbiased offline evaluation for this interactive process

**ACTION**

Unbiased offline evaluation for this interactive process

**REWARD**

Unbiased offline evaluation for this interactive process

**POLICY**

Unbiased offline evaluation for this interactive process
Outline

- News recommendation as a contextual bandit
- Unbiased offline evaluation
- Experiments
- Conclusions
Today Module @ Yahoo! Front Page

A small pool of articles chosen by editors

“Featured Article”

Few drugs developed for super bacteria

Doctors are struggling to fight a lethal bacteria that is "resistant to virtually every antibiotic." ➤ Where it's found

- Acinetobacter baumannii
- Do flu vaccines work?
- H1N1 still worrisome

1 - 4 of 32
Challenge: “Explore or Exploit?”

- Objective: to maximize click-through rate (CTR)
- Only displayed articles have user click feedback

**EXPLOIT** (choose *good* articles to maximize CTR)

**EXPLORE** (choose *novel* articles to improve CTR est.)

How to trade off?

Same in advertising, search, …
Contextual Bandit Formulation

K-armed “contextual bandit” [Langford & Zhang, 2008]

Observe $K$ arms $A$ and “context” $x_t$

Select $a_t \in A$

Receive reward $r_t \in [0,1]$

Goal: maximize $\sum_{t=1}^{T} r_t$

In Today Module:

$A$: available articles

$x_t$: user features

$a_t$: displayed article

$r_t$: 1 for click, 0 for no click

Key Challenge

No reward feedback

For unselected arms
Outline

- News recommendation as a contextual bandit

- Unbiased offline evaluation

- Experiments

- Conclusions
Offline Evaluation of Bandit Algorithms

• Want to estimate:  \( V(\pi) := \mathbb{E}_x [r(x, \pi(x))] \)

• Why offline evaluation
  • Cheap and risk-free!
  • Avoid frequent bucket tests
  • Replicable / fair comparisons

\( V(\pi) :\mathbb{E}_x [r(x, \pi(x))] \)
Common/Prior Evaluation Approaches

\[
\begin{align*}
\langle x_1, a_1, r_1 \rangle & \\
\langle x_L, a_L, r_L \rangle & \\
M & \\
\end{align*}
\]

Reward simulator:
\[
\hat{r}(x,a) \approx E[r|x,a]
\]

this (difficult) step is often biased

unreliable evaluation

In contrast, our approach
- avoids explicit user modeling \(\Rightarrow\) simple
- gives unbiased evaluation results \(\Rightarrow\) reliable
Our Evaluation Method: “Replay”

Want to estimate \( V(\pi) := \mathbb{E}_x \left[ r(x, \pi(x)) \right] \)

Key requirement for data collection: \( a_i \sim \text{unif}(A) \)

For \( i = 1, 2, \ldots, L \):

- reveal \( x_i \)
- choose \( \hat{a}_i = \pi(x_i) \)
- reveal \( r_i \) only if \( \hat{a}_i = a_i \) (a "match")

Finally, output \( \hat{V} = \frac{K}{L} \sum_{i=1}^{L} r_i \cdot I(\hat{a}_i = a_i) \)
Theoretical Guarantees

- **Thm 1**: Our estimator is unbiased
  - Mathematically, \( V(\pi) = \mathbb{E}[\hat{V}] \)
  - So on average \( \hat{V} \) reflects real, online performance
Outline

- News recommendation as a contextual bandit
- Unbiased offline evaluation
- Experiments
- Conclusions
Case Study in Today Module

- Data:
  - Large volume of real user traffic in Today Module

- Policies being evaluated:
  - EMP [Agarwal et al. 2009]
  - SEMP/CEMP: personalized models
  - Use policies’ online bucket CTR as “truth”

- Random bucket data for evaluation:
  - 40M visits, K ~ 20 on average
  - Use it to offline-evaluate policies’ CTR

Are they close?
Unbiasedness (Article nCTR)

The offline estimate is indeed unbiased!
Unbiasedness *(Daily nCTR)*

The offline estimate is indeed unbiased!

Estimated nCTR

Recorded Online nCTR

Ten Days in November 2009
Recall our theoretical error bound:

**Thm 2 (error bound):** \( V(\pi) - \hat{V} = O(\sqrt{K/L}) \)
When Business Rules Exist

- Human editors may overrule bandit algorithm’s recommendations
- Have roughly the same multiplicative impacts on algorithm’s CTR

\[
\frac{\text{offlineCTR}(\pi_{\text{EMP}})}{\text{onlineCTR}(\pi_{\text{EMP}})}
\]

Can still reveal relative performance!
Outline

- News recommendation as a contextual bandit
- Unbiased offline evaluation
- Experiments
- Conclusions
Extensions

- What if we don’t have uniformly random data?
  - Cost constraints, system constraints, etc.
  - Can use importance reweighting [Strehl, Langford, Li, Kakade, 2011]
  - “Doubly robust” tech. for variance reduction [Li, Dudik, Langford, 2011]

- What if $K$ is too large?
  - Pre-filter unpromising candidates [Moon et al. 2010]

- Open issues
  - incorporating history-dependent constraints
Take-Home Messages

- Interactive machine learning is common on Web
- We investigated an offline evaluation method that
  - gives unbiased result (with low variance)
  - enjoys fast error decay rate (with more data)
  - is shown accurate using Y! Today Module traffic
  - avoids frequent bucket tests that are risky/costly