Exploring Bandit Algorithms for Automatic Content Selection

Dai Shi
I want money!
Which one to choose?
Exploration and Exploitation Dilemma

Exploration: Explore an unknown or suboptimal choice to improve the knowledge about the problem.

Exploitation: Exploit the current knowledge to guess the best choice.
Reinforcement learning

First, the agent gets some perception from environment.

Then it makes an action based on its judgment.

After that the environment would return a reward.

This reward would facilitate the agent making future actions.
Bandit Algorithms

Old Website
Conversion: 0.2%
Traffic: 1200 visitors
Site Revenue: $

Optimized Website
Conversion: 1.2%
Traffic: 2800 visitors
Site Revenue: $\$\$\$
Epsilon-greedy
Softmax

• There’s obvious drawback of epsilon-greedy, which is in the exploration part, it’s haphazard. In order to tackle this problem, we can choose arm based on the reward rate.

\[
P(A) = \frac{r_A}{\tau e^{-r_A/\tau}} + \frac{r_B}{\tau e^{-r_B/\tau}}
\]

Boltzmann Distribution
UCB 1

• Upper confidence bounds
• No randomness
• No configurable parameters

Number of successes of arm \( i \)
Number of pulls of arm \( i \)
Observed success rate
Factor representing uncertainty

Total number of pulls of all arms

Priority \( i \) = \( \frac{c_i}{n_i} + \sqrt{\frac{2 \cdot \log n}{n_i}} \)
Stochastic and Adversarial

• In the stochastic model we assume that the rewards of a given arm is an i.i.d sequence of random variables.

• While, in the adversarial model, there’s no restriction on the sequence of rewards.
Exp3

• Exponential-weight algorithm for Exploration and Exploitation.

\[ p_i(t) \leftarrow (1 - \gamma) \frac{w_i(t)}{\sum_{j=1}^{K} w_i(t)} + \frac{\gamma}{K} \]

• It contains uniform distribution, which ensures the exploration of all the arms.
Contextual Bandit Problem

Query, IP address, browser properties, etc.

result (ie. ad, news story)

click or not

Source: http://www.levreyzin.com/presentations/CMU_bandits.pdf
## Methods

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>LinUCB</td>
<td>Expected payoff + confidence interval According to context matrix</td>
</tr>
<tr>
<td>LinearBayes</td>
<td>$P(a</td>
</tr>
<tr>
<td>Exp4</td>
<td>Exponential weight of each policy, which can be defined by supervised learning</td>
</tr>
</tbody>
</table>
Synthetic: 10000 trials
Synthetic: 100000 trials
Synthetic: 1000000 trials
Product Data: Results

![Graph showing CTR vs Data Volume for different algorithms: UCB1, LinearBayes, LinUCB, Exp3, Softmax, Epsilon.](image)
Exploration
Exploitation
General Approach

• 1. Collect unbiased data for testing.

• 2. Set up tailored solution, referenced to bandit algorithms.

• 3. Test and implement it on real product with comparison test.

• 4. Full implementation.
Summary

• Gap between theory and practice

• Can be really useful in many applications

• Need to be customized case by case

• Help us to make decisions in real life