Exploration Scavenging

John Langford, Alexander L. Strehl, Jennifer Wortman
jl@yahoo-inc.com, strehl@yahoo-inc.com, wortmanj@seas.upenn.edu

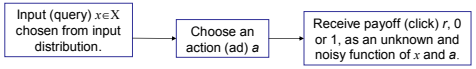
Yahoo! Research
(http://www.research.yahoo.com)
111 W 40th Street, 17th Floor, New York, NY

Department of Computer and Information Science,
University of Pennsylvania,
Philadelphia, PA, 19104

Abstract:
We examine the problem of evaluating a policy in the contextual bandit setting using only observations collected during the execution of another policy. We propose and prove the correctness of a principled method for policy evaluation which works even when the exploration policy is deterministic, as long as each action is explored sufficiently often. Although our theoretical results hold only when the exploration policy chooses ads independent of side information, an assumption that is typically violated by commercial systems, we show how clever uses of the theory provide non-trivial and realistic applications.

The Offline Policy Learning Problem

The Setting: Bandit Supervised Learning



Offline Policy evaluation (or maximization): given data $(x_1, a_1, r_1), \ldots, (x_m, a_m, r_m)$ from old policy π, how do we evaluate a new policy h:

$$V(h) = \text{Probability of click for policy h} = E_{(x,r|h)}[r].$$

Potential applications to medical treatments, robotics, etc...

Importance-weighted Approach

Basic observation: Only the reward of the displayed ad is observed $\Rightarrow$ Can't use supervised learning.

One approach to consider is importance sampling. It relies on the logging policy being explicitly **randomized**. The key observation is:

$$\text{Value(h)} = E_{(x,r)\sim D, a\sim\pi}\left[\frac{r(a)\cdot I[h(x)=a]}{\pi(a|x)}\right],$$

where *h* = old policy, π = new policy, *a* = action, *r(a)* = reward for action *a*.

Unfortunate problem: The logging policies are **not randomized**. What can we do?

Offline Policy Estimator

From data $\{(x_1, a_1, r_1), (x_2, a_2, r_2), \ldots, (x_T, a_T, r_T)\}$, we form the estimator:

$$\hat{V}(h) = \frac{1}{T}\sum_{t=1}^{T} \frac{I(h(x_t)=a_t)(\text{reward at time t})}{(\#\text{ times action } a_t \text{ chosen}/T)}.$$

Main Result

Theorem 2.1. *For any policy h and $\delta \in (0,1)$, if the each action a_t is chosen independently of input x_t, then with probability $1-\delta$,*

$$|V(h) - \hat{V}(h)| \leq \sum_{a=1}^{k} \sqrt{\frac{2\ln(2kT/\delta)}{|\{t : a_t = a\}|}}.$$

Multiple Exploration Policies

Current policy changes over time. Denote the set of historical policies $\Pi = \{\pi_1, \cdots, \pi_K\}$. Redefining the action space to be Π, we evaluate policies of the form $h : X \to \Pi$.

Corollary 2.1. *Let H be a set of policies of the form $h : X \to \Pi$ and δ a number between 0 and 1. Let $h^* = \text{argmax}_{h\in H}\{\text{Value}(h)\}$ and $\hat{h}^* = \text{argmax}_{h\in H}\{\hat{V}(h)\}$. Then, with probability at least $1-\delta$,*

$$\text{Value}(\hat{h}^*) \geq \text{Value}(h^*) - \sum_{a=1}^{k} \sqrt{\frac{2\ln(2kT|H|/\delta)}{|\{t : a_t = a\}|}}.$$

In this result, the historical policies π_i as well as the new policies h depend on input. Our Theorem applies because the choice of which π_i to follow does not.

Impossibility Result

Under no assumptions, the problem is unsolvable.

2 inputs, 0 and 1.

2 action, 0 and 1.

Old policy $\pi(x) = x$, is determinstic.

We cannot evaluate the new policy h(x) = 1-x.

Conclusion

We have provided a new estimator for the policy evaluation problem that converges rapidly.

We have demonstrated that it provides good estimates of reordering policies in online advertising.