Aggressive Learning for Contextual Bandits

Miroslav Dudik (mdudik@yahoo-inc.com)
Daniel Hsu (danielhsu@gmail.com)
Satyen Kale (skale@yahoo-inc.com)
Nikos Karampatziakis (nk@cs.cornell.edu)
John Langford (jl@yahoo-inc.com)
Lev Reyzin (lev.reyzin@gmail.com)
Tong Zhang (tongz@rci.rutgers.edu)

April 6, 2011

1 Introduction

The contextual bandit setting consists of the following basic loop repeated indefinitely:

1. The world presents context information as features x.
2. The learning algorithm chooses an action a.
3. The world presents a reward r for the action.

The key difference between the contextual bandit setting and standard supervised learning is that *only* the reward of the chosen action is revealed. For example, after always choosing the same action several times in a row, the feedback given provides no basis to prefer the chosen action over another action. In essence, the contextual bandit setting captures the difficulty of exploration while avoiding the difficulty of credit assignment as in more general reinforcement learning settings.

Many natural problems fit well into this setting. For example, the problem of choosing interesting news articles or ads for users by internet companies can be naturally modeled as a contextual bandit setting, since the company gets feedback about articles presented, but not about articles not presented. Another setting is the medical domain where discrete treatments are tested before approval, the process of deciding which patients are eligible for a treatment takes context into account.

Until now, only the EXP4 Auer et al. (2002) algorithm was known to optimally guide exploration for the purpose of learning over a large arbitrarily structured policy space. This algorithm has two core (and apparently irremovable) drawbacks which make it quite awkward to apply in practice.

1. Computational Tractability. The computational complexity is $O(N)$ in general where N is number of policies being competed with. This problem appears extremely difficult to address, as it has not yet been addressed for this style of algorithm even in a fully supervised setting, and even with an optimization oracle. In contrast, for example, in an IID supervised learning setting empirical risk minimization (otherwise known as follow the leader) requires only $O(\log N)$ computation given an optimization oracle.

2. Faster Learning. The EXP4 algorithm is optimized for a deeply adversarial world. Often, it's possible to learn much faster in less adversarial settings. The structural form of the EXP4 proof precludes simple sound modifications to make it more aggressive.

We present a new core algorithm with new analysis techniques working in any policy space with any data source satisfying an IID assumption. In the IID setting, the world draws $(x, \vec{r})$ from some unknown distribution D, revealing x in step 1 and the reward r_a of the chosen action a in step 3. Given a set of policies Π, the goal is to create an algorithm for step 2 which competes with the set of policies. We can measure our regret by comparing the algorithm's cumulative reward to the expected cumulative reward of the best policy in the set.

We have a new algorithm (really, a class of algorithms), satisfying the following basic result:

Theorem 1. *For all distributions D over K actions and features, for all sets of policies Π, with probability $1 - \delta$, the new algorithm has regret at most:*

$$14\sqrt{TK \ln \frac{2T(T+1)|\Pi|}{\delta}}.$$

This result can easily be extended to deal with VC classes, and yields faster learning rates than EXP4 in various special cases. For settings with delayed feedback, where the observed reward is not known immediately, we can easily modify the algorithm to learn much faster than EXP4.

The best previous (known suboptimal) result for the IID setting is the epoch greedy algorithm Langford & Zhang (2007) which has regret scaling as $O(T^{2/3})$ in the worst case. All other bandit-related results are for specialized settings where policy spaces have constant, linear, or gaussian structure, typically with additional constraints on the reward structure.

References

Auer, Peter, Cesa-Bianchi, Nicolò, Freund, Yoav, and Schapire, Robert E. The nonstochastic multiarmed bandit problem. *SIAM Journal of Computing*, 32(1):48–77, 2002.

Langford, John and Zhang, Tong. The epoch-greedy algorithm for contextual multi-armed bandits. In *Neural Information Processing Systems (NIPS)*, 2007.