Meta-Learning Contextual Bandit Exploration

Amr Sharaf
University of Maryland
amr@cs.umd.edu

Hal Daumé III
Microsoft Research &
University of Maryland
me@hal3.name

Abstract
In contextual bandits, an algorithm must choose actions given observed contexts, learning from a reward signal that is observed only for the action chosen. This leads to an exploration/exploitation trade-off: the algorithm must balance taking actions it already believes are good with taking new actions to potentially discover better choices. We develop a meta-learning algorithm, $\hat{M}\hat{E}\hat{L}$, that learns an exploration policy based on simulated, synthetic contextual bandit tasks. $\hat{M}\hat{E}\hat{L}$ learns via imitation learning against these simulations to learn an exploration policy that can be applied to true contextual bandit tasks at test time. In hundreds of real-world datasets, $\hat{M}\hat{E}\hat{L}$ outperforms seven strong baseline algorithms on most datasets, and can leverage a rich feature representation for learning an exploration strategy.

1 Introduction
In a contextual bandit problem, an agent optimizes its behavior over a sequence of rounds based on limited feedback (Kaelbling, 1994; Auer, 2003; Langford & Zhang, 2008). In each round, the agent chooses an action based on a context (features) for that round, and observes a reward for that action but no others. Contextual bandit problems arise in many real-world settings like online recommendations and personalized medicine. As in reinforcement learning, the agent must learn to balance exploitation and exploration.

We present a meta-learning approach to automatically learn a good exploration mechanism from data—specifically, simulated, synthetic offline data. Based on these simulations, our algorithm, $\hat{M}\hat{E}\hat{L}$ (MEta LEarner for Exploration), learns a good heuristic exploration strategy that should ideally generalize to future contextual bandit problems. At training time ($\S\ 2$), $\hat{M}\hat{E}\hat{L}$ simulates many contextual bandit problems from fully labeled synthetic data. Thus, $\hat{M}\hat{E}\hat{L}$ is able to counterfactually simulate what would happen under all possible action choices. It then uses this information to compute regret estimates for each action, which can be optimized using the AggreVaTe imitation learning algorithm (Ross & Bagnell, 2014). Our imitation learning strategy mirrors that of the meta-learning approach of Bachman et al. (2017) in the active learning setting. Empirically, we use $\hat{M}\hat{E}\hat{L}$ to train an exploration policy on only synthetic datasets and evaluate the resulting bandit performance across three hundred (simulated) contextual bandit tasks ($\S\ 3$), comparing to a number of alternative exploration algorithms, and showing the efficacy of our approach ($\S\ 3$).

2 Approach: Learning and Effective Exploration Strategy
In order to have an effective approach to the contextual bandit problem, one must be able to both optimize a policy based on historic data and make decisions about how to explore. The exploitation/exploitation dilemma is fundamentally about long-term payoffs: is it worth trying something potentially suboptimal now in order to learn how to behave better in the future? A particularly simple and effective form of exploration is $\varepsilon$-greedy: given a function $f$ output by $POLOPT$, act according to $f(x)$ with probability $(1-\varepsilon)$ and act uniformly at random with probability $\varepsilon$. Intuitively, one would hope to improve on such a strategy by taking more (any!) information into account; for instance,
Can we learn to explore in contextual bandits?
Contextual Bandits: News Display
Contextual Bandits: News Display
Contextual Bandits: News Display
Contextual Bandits: News Display

Goal: Maximize Sum of Rewards
Training Mêlée by Imitation

Access to $\pi^*$ at train

Goal: learn $\pi$

Roll-out with $\pi^*$

Roll-in with $\pi$

Examples / Time
Generalization: Meta-Features

- No direct dependency on the contexts $x$.

- Features include:
  
  - Calibrated predicted probability $p(a_t \mid f_t, x_t)$;
  
  - Entropy of the predicted probability distribution;
  
  - A one-hot encoding for the predicted action $f_t(x_t)$;
  
  - Current time step $t$;
  
  - Average observed rewards for each action.
A representative learning curve
Win / Loss Statistics

Win statistics: each (row, column) entry shows the number of times the row algorithm won against the column, minus the number of losses.
Win / Loss Statistics

Win statistics: each (row, column) entry shows the number of times the row algorithm won against the column, minus the number of losses.
Theoretical Guarantees

- The no-regret property of Aggrevate can be leveraged in our meta-learning setting.

- We relate the regret of the learner to the overall regret of $\pi$.

- This shows that, if the underlying classifier improves sufficiently quickly, Mêlée will achieve sublinear regret.
Conclusion

- Q: Can we learn to explore in contextual bandits?

- A: Yes, by imitating an expert exploration policy;

- Generalize across bandit problems using meta-features;

- Outperform alternative strategies in most settings;

- We provide theoretical guarantees.