Abstract

Introduction

Explore/Exploit

Many decisions that people are faced with require finding a balance between exploiting known information for short-term gain and exploring new sources of information that may lead to future reward. For example, a person might choose to go to a restaurant that they know well and have had multiple satisfying meals at, or choose to try a new restaurant that could result in a better or worse experience. This is known as the explore-exploit trade-off.

MAB

Many tasks requiring this kind of trade-off can be characterized as multi-arm bandit (MAB) problems. A MAB problem consists of A possible actions, where the a-th action yields the random reward r_t drawn from an unknown distribution. Over T trials, the agent must choose $a_{t:T}$ such that $\sum r_{t:T}$ is maximized. The policy $\tilde{\pi}$ maps the observed history of actions and rewards $h_{1:t-1} = [(a_1, r_1), ..., (a_{t-1}, r_{t-1})]$ to the next action a_t for all possible histories $h_{1:t-1} \in \mathcal{H}$. The goal in a MAB problem is to find the optimal policy such that the expected reward over T trials, $\mathbf{E}[r_{1:T}|\tilde{\pi}(h_{1:t-1})]$, is maximized. This is the solution to the Bellman equation

$$\mathbf{E}[r_{1:T}|\tilde{\pi}(h_{1:t-1})] = \mathbf{E}[r_t|\tilde{\pi}(h_{1:t-1})] + \mathbf{E}[r_{t+1:T}|\tilde{\pi}(h_{1:t})]$$

where the first term yields the expected reward on trial t and the second term recursively defines the expected reward on all subsequent trials up to T. This second term can be evaluated for trial $t+1$ by weighing the expected reward given all possible histories $\mathcal{H}_{1:t} = [(a_1, r_1), ..., (a_{t-1}, r_{t-1}), (a_t, r_t)]$ for $r_t \in \mathcal{R}$ by the probability of that history occurring:

$$\mathbf{E}[r_{t+1:T}|\tilde{\pi}(h_{1:t})] = \sum_{r_t \in \mathcal{R}} p(r_t|a_t, h_{1:t-1})\mathbf{E}[r_{1:T}|\tilde{\pi}(h_{1:t})]$$

If the reward r observed after choosing action a is drawn from the distribution $H(\theta_a)$, an agent must employ a strategy that balances choosing the action that maximizes the reward on trial t, $a_t^* = \text{argmax}_{a_t \in A}\mathbf{E}[H(\theta_{a_t})]$, and learning more about θ_A by choosing novel actions to maximize reward on future trials. As the problem of evaluating all possible histories is intractable, one of two general classes of approximations can be used. The first, simulation-based methods, involves using a subset of possible histories to evaluate the long-term reward of an action, and includes Thompson sampling (Thompson, 1933) and Monte Carlo tree search (Coulom, 2007). Rather than relying on simulations, myopic strategies define a value function for approximating long-term reward. Common myopic strategies include upper confidence bound (UCB) (Agrawal, 1995) and epsilon-greedy (Sutton & Barto, 1998) algorithms.

CMAB/Function Learning

Contextual multi-armed bandit (CMAB) problems introduce additional information into the standard MAB problem by way of a set of features associated with the set of possible examples (Langford & Zhang, 2008). For example, a standard MAB formulation of the problem of choosing which restaurant to eat at assumes that the reward yielded by any two restaurants will be uncorrelated. However, it might be the case that these restaurants share a set of features (e.g. size, location, menu items) such that choosing similar restaurants can be assumed to yield similar rewards. Rather than having to execute an action to be able to evaluate its expected reward, considering shared features allows one to learn a function, $R : x_t \rightarrow r_t$ that maps features of the action a_t, x_{a_t} to that action's expected reward, r_t.

With the inclusion of context, CMAB problems require the additional step of learning the function R between updating the history, $h_{1:t-1} = [(a_1, x_1, r_1), ..., (a_{t-1}, x_{t-1}, r_{t-1})]$, and choosing $a_t = \tilde{\pi}(h_{t-1})$. Theories of function learning typically follow either a rule-based or similarity-based approach.

Basic CMAB Strategies

(eps greedy, mean/var greedy, ucb/pmi/mi, entropy search)

Once an agent has a mapping between features and reward, the policy $\tilde{\pi}$ must be selected that chooses the action a_t given the history of observed actions, features, and rewards $h_{1:t-1}$. Given that this policy does not take into account all possible trajectories, it must include a value function that approximates the long-term reward of any given action.

Alternative Goals

In addition to CMABs, the explore/exploit tradeoff can also be observed in other common context-dependent problems. In active learning (AL) Bramley et al. (2016), participants assume some degree of control over the contexts that they observe, rather than passively observing a predetermined set of examples. Optimization problems Rachlin et al. (1981) require finding the set of features x that maximizes some reward $f(x)$; for example, finding the number of hours to dedicate to work and to leisure respectively that maximizes satisfaction.

While these tasks might all share the same mapping between action and reward, their goal-specific reward, and thus the function used to approximate long-term reward, are distinct. For any particular action reward pair (a_t, r_t), there exists the goal-specific action reward pair (a_t, r_t'). Since the goal in the CMAB task is to maximize cumulative reward over time, the goal-specific reward is the same as the reward, that is:

$$\mathbf{E}_{cmab}[r_t'|a_t, h_{1:t-1}] = \mathbf{E}[r_t|a_t, h_{1:t-1}]$$

In contrast, the goal of the optimization problem is simply to find the maximum possible reward withing T steps. As such, the goal specific reward is defined as:

$$\mathbf{E}_{opt}[r'_t|a_t,h_{1:t-1}] = max(\mathbf{E}[r_t|a_t,h_{1:t-1}] - \sum_{i=1}^{t-1} r'_i, 0)$$

That is, on trial if trial t yields an increase in reward over the previous maximum reward, r'_t is the difference. If not, r'_t is 0. Since active learning is concerned with learning the reward function rather than the magnitude of the rewards themselves, its goal-specific reward can be described as the sum of the decrease in variance across all possible actions:

$$\mathbf{E}_{al}[r'_t|a_t,h_{1:t-1}] = \sum_{a\in A} \mathbf{Var}[r|a,h_{1:t-1}] - \mathbf{Var}[r|a,h_{1:t}]$$

References

Agrawal, R. (1995). Sample mean based index policies with o(log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, *27*(4), 1054-1078. Retrieved from http://www.jstor.org/stable/1427934

Bramley, N., Gerstenberg, T., & Tenenbaum, J. B. (2016). *Natural science: Active learning in dynamic physical microworlds.* 38th Annual Meeting of the Cognitive Science Society.

Coulom, R. (2007). Efficient selectivity and backup operators in monte-carlo tree search. In H. J. van den Herik, P. Ciancarini, & H. H. L. M. J. Donkers (Eds.), *Computers and games: 5th international conference, cg 2006, turin, italy, may 29-31, 2006. revised papers* (pp. 72–83). Berlin, Heidelberg: Springer Berlin Heidelberg.

Langford, J., & Zhang, T. (2008). The epoch-greedy algorithm for multi-armed bandits with side information. In J. C. Platt, D. Koller, Y. Singer, & S. T. Roweis (Eds.), *Advances in neural information processing systems 20* (pp. 817–824). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/3178-the-epoch-greedy-algorithm-for-multi-armed-bandits-with-side-information.pd

Rachlin, H., Battalio, R., Kagel, J., & Green, L. (1981). Maximization theory in behavioral psychology. *Behavioral and Brain Sciences*, *4*(3), 371388. doi: 10.1017/S0140525X00009407

Sutton, R. S., & Barto, A. G. (1998). *Introduction to reinforcement learning* (1st ed.). Cambridge, MA, USA: MIT Press.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, *25*(3/4), 285-294. Retrieved from http://www.jstor.org/stable/2332286