Intervention Bandits

Blah blah

November 7, 2014

Abstract

An abstract.

1 Introduction

Useful references are: **?**.

2 Notation

Assume we have a known causal model with binary variables $\boldsymbol{X} = \{X_1..X_K\}$ that independently cause a target variable of interest Y. We can run sequential experiments on the system, where at each timestep t we can select a variable on which to intervene and then we observe the complete result, $(\boldsymbol{X}_t, Y_t)$ This problem can be viewed as a variant of the multi-armed bandit problem.

Let $p \in [0,1]^K$ be a fixed and known vector. In each time-step t:

1. The learner chooses an $I_t \in \{1, \ldots, K\}$ and $J_t \in \{0, 1\}$.
2. Then $X_t \in \{0,1\}^K$ is sampled from a product of Bernoulli distributions, $X_{t,i} \sim \text{Bernoulli}(p_i)$
3. The learner observes $\tilde{X}_t \in \{0,1\}^K$, which is defined by

$$\tilde{X}_{t,i} = \begin{cases} X_{t,i} & \text{if } i \neq I_t \\ J_t & \text{otherwise}\,. \end{cases}$$

4. The learner receives reward $Y_t \sim \text{Bernoulli}(q(\tilde{X}))$ where $q : \{0,1\}^K \to [0,1]$ is unknown and arbitrary.

The expected reward of taking action i, j is $\mu_{i,j} = \mathbb{E}[q(X)|X_i = j]$. The optimal reward and action are μ^* and (i^*, j^*) respectively, where $(i^*, j^*) = \arg\max_{i,j} \mu_{i,j}$ and $\mu^* = \mu(i^*, j^*)$. The n-step cumulative expected regret is

$$R_n = \mathbb{E} \sum_{t=1}^{n} (\mu^* - \mu_{I_t, J_t})\,.$$

3 Estimating $\mu_{i,j}$

The most natural way to estimate $\mu_{i,j}$ is to compute an empirical estimate based on samples when that action was taken. This approach would lead directly to the UCB algorithm with $2K$ actions and a regret bound that depended linearly on K. In this instance we can significantly outperform this approach by exploiting the extra structure in the problem.

Fix some time-step t and $i \in \{1, \ldots, K\}$ and $j \in \{0, 1\}$. We define estimator $\hat{\mu}_t$ of $\mu_{i,j}$ by

$$
\begin{aligned}
m_{a,b} &= \sum_{s=1}^{t} \mathbb{1}\{I_t = i, J_t = j, X_a = b, Y_t = 1\} \\
n_{a,b} &= \sum_{s=1}^{t} \mathbb{1}\{I_t = i, J_t = j, X_a = b\} \\
\hat{\mu}_t &= \eta_i \frac{m_{i,j}}{n_{i,j}} + \sum_{a \neq i} \eta_a \left[p_a \frac{m_{a,1}}{n_{a,1}} + (1 - p_a) \frac{m_{a,0}}{n_{a,0}} \right] \\
\eta_a &= \frac{n_a}{\sum_{a=1}^{K} n_a} \\
n_a &= \begin{cases} n_{i,j} & \text{if } a = i \\ \frac{1}{2} \min\left\{ \frac{n_{a,1}}{p_a}, \frac{n_{a,0}}{1-p_a} \right\} & \text{otherwise} \end{cases}
\end{aligned}
$$

Theorem 1. *With probability at least* $1 - \delta$ *we have that:* $|\hat{\mu}_t - \mu| \leq \sqrt{\frac{\beta}{\sum_a n_a} \log \frac{1}{\delta}}$*, where* $\beta > 0$ *is some constant.*

Proof. First note that $n_{a,b}$ is a random variable that is bounded by t for all a, b. We use the short-hand $\mu_{i,j}^{a,b} = \mathbb{E}[q(X)|X_i = j, X_a = b]$. Then

$$
\mu_{i,j} = p_a \mu_{i,j}^{a,1} + (1 - p_a) \mu_{i,j}^{a,0} \,.
$$

Now we can apply Hoeffding's bound and the union bound to show that

$$
\mathrm{P}\left\{ \left| \frac{m_{a,b}}{n_{a,b}} - \mu_{i,j}^{a,b} \right| \geq \sqrt{\frac{1}{2n_{a,b}} \log \frac{4t}{\delta}} \right\} \leq \frac{\delta}{2} \,.
$$

Therefore by the union bound

$$
\mathrm{P}\left\{ \left| p_a \frac{m_{a,1}}{n_{a,1}} + (1 - p_a) \frac{m_{a,0}}{n_{a,0}} - \mu_{i,j} \right| \geq p_a \sqrt{\frac{1}{2n_{a,1}} \log \frac{4t}{\delta}} + (1 - p_a) \sqrt{\frac{1}{2n_{a,0}} \log \frac{4t}{\delta}} \right\} \leq \delta
$$

Now by Jensen's inequality

$$
\begin{aligned}
p_a \sqrt{\frac{1}{2n_{a,1}} \log \frac{4t}{\delta}} + (1 - p_a) \sqrt{\frac{1}{2n_{a,0}} \log \frac{4t}{\delta}} &\leq \sqrt{\left(\frac{p_a}{2n_{a,1}} + \frac{1 - p_a}{2n_{a,0}} \right) \log \frac{4t}{\delta}} \\
&\leq \sqrt{\max\left\{ \frac{p_a}{n_{a,1}}, \frac{1 - p_a}{n_{a,0}} \right\} \log \frac{4t}{\delta}} \\
&= \sqrt{\frac{1}{2n_a} \log \frac{4t}{\delta}} \,.
\end{aligned}
$$

Similarly,

$$
\mathrm{P}\left\{ \left| \frac{m_{i,j}}{n_{i,j}} - \mu_{i,j} \right| \geq \sqrt{\frac{1}{2n_a} \log \frac{4t}{\delta}} \right\} \leq \mathrm{P}\left\{ \left| \frac{m_{i,j}}{n_{i,j}} - \mu_{i,j} \right| \geq \sqrt{\frac{1}{2n_a} \log \frac{2t}{\delta}} \right\} \leq \delta \,.
$$

□

Algorithm 1 UCB

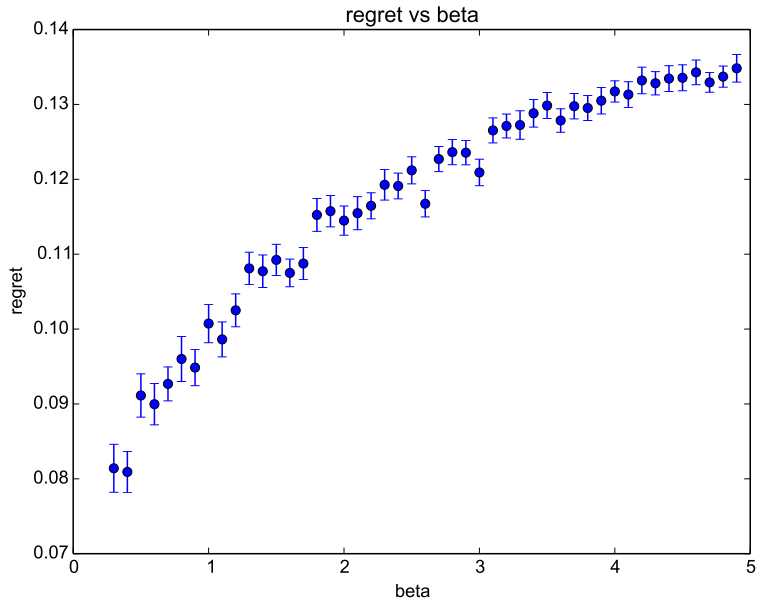
Input: Number of variables K, vector $p \in [0,1]^K$, horizon n
for $t \in 1, \ldots, n$ **do**
  for $i \in 1, \ldots, K$ **do**
    for $j \in \{0, 1\}$ **do**
      Compute $\tilde{\mu}_{i,j} = \hat{\mu}_{i,j} + \sqrt{\frac{\alpha}{\sum_a n_a} \log n}$
    end for
  end for
  Choose $I_t, J_t = \arg\max_{i,j} \tilde{\mu}_{i,j}$
end for

4 Algorithm

5 Theorems

6 Experiments

Figure 1: Testing values of beta - 100 experiments for each value of beta, with a 4 armed bandit.



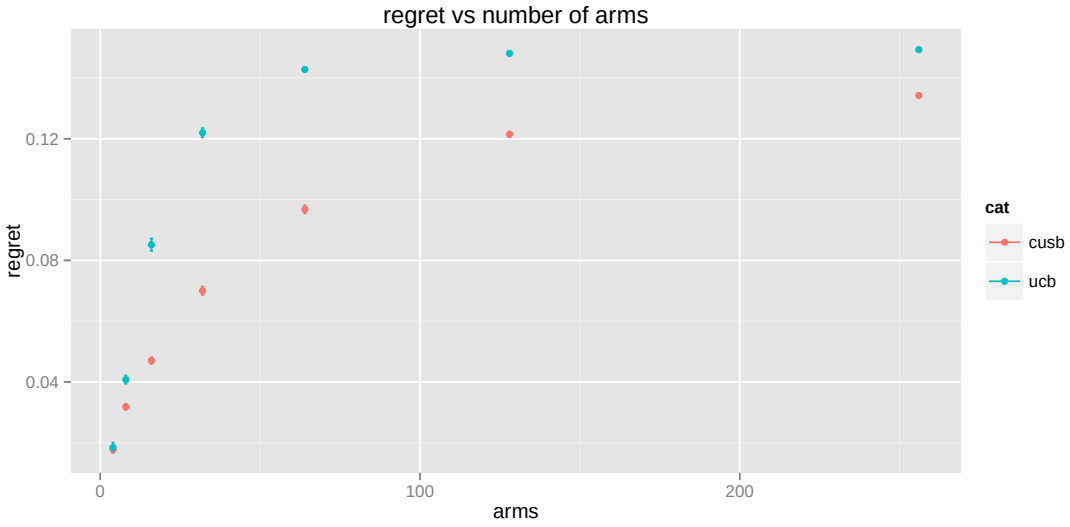
Simululations to compare the performance of standard UCB with our modified algorithm. For each number of arms, 100 bandits of each type were created and run upto to a horizon of 1000 timesteps. The mean regret and its standard error from these simulations is plotted in figure 2 The true data was generated from a model where:

$$p = [0.5]^K$$

$$q(\boldsymbol{X}) = \begin{cases} 0.5 & \text{if } X_1 = 0 \\ 0.6 & \text{otherwise} \end{cases}$$

For small number of arms there is not much difference. As the number of arms increases, cusb shows significantly lower regret. As the number of arms increases above 100 the regret of the two algorithms begins to converge again. I suspect this is due to the fixed time horizon of 1000 steps. With a time horizon of 100 steps the two algorithms begin to converge again at a smaller number of arms.

Figure 2: Shows relationship between the regret and the number of bandit arms $2K$ for the standard UCB algorithm and our modified algorithm.



7 Conclusion