Correlated Multi-armed Bandits

CS 6780 Advanced Machine Learning

Zhengdi Shen
Bangrui Chen
Saul Toscano Palmerin

April 23, 2015

Motivation

For a new user on Yelp, what restaurants should Yelp recommend at each time to maximize the expected average rating of the user?

- Each restaurant is represented with a r dimensional binary vector, corresponding to the categories it belongs to (e.g. Pizza, Sandwiches, Mexican, Chinese, Italian).
- Each user has an unknown preference vector θ.
- Multi-Armed Bandits. (At most 2^r arms!)
- Dependent arms.

Motivation

$r = 5$	Pizza	Sandwiches	Mexican	Chinese	Italian
Restaurant 1	1	1	1	0	1
Restaurant 2	1	1	0	0	1
Restaurant 3	0	0	0	1	0
Restaurant 4	1	0	0	0	1
...	...	...	...	...	...

Problem Formulation

- The reward of choosing a restaurant with features $X \in \{0,1\}^r$ at time t is defined by

$$Y_t = X \cdot \theta + W_t$$

where $W_t \sim N\left(0, \sigma^2\right)$ is a measurement error.

- We place a Gaussian prior distribution on the preference vector: $\theta \sim N\left(\mu_0, \Sigma_0\right)$

Regret

Definition of Regret

For any policy, we define the T-period regret cumulative as

$$\text{Regret}\left(\theta_0, T\right) = \sum_{t=1}^{T} \mathbb{E}\left[\max_{X \in \{0,1\}^r} X \cdot \theta_0 - X_t \cdot \theta_0 \mid \theta = \theta_0\right]$$

where X_t is the feature vector of the restaurant selected at stage t.

Lower Bound for Regret

Lower Bound for Regret

For an arbitrary policy, the regret is at least $\Omega\left(r\sqrt{T}\right)$ under some regularity conditions, where the set of arms is compact in $\mathbb{R}^r$.

"Optimal Algorithm"

The Phased Exploration and Greedy Exploitation (PEGE) algorithm has regret $\Omega\left(r\sqrt{T}\right)$ under some regularity conditions.

PEGE [Linearly Parameterized Bandits, P. R., J. T., 2010]

Find r arms $X_{b_1}, \cdots, X_{b_r}$ which form a maximal linearly independent system.
In each cycle c:

1. Exploration (r periods): Play arm X_{b_k}, and observe the reward $Y^{X_{b_k}}(c)$. Compute the ordinary least squares estimate $\hat{\theta}(c)$.
2. Exploitation (c periods): Play the greedy arm $G(c) = \arg\max_X X \cdot \hat{\theta}(c)$ for c periods.

But it may have large constant in front of the order of its regret.

"Optimal Algorithm"

The Phased Exploration and Greedy Exploitation (PEGE) algorithm has regret $\Omega\left(r\sqrt{T}\right)$ under some regularity conditions.

PEGE [Linearly Parameterized Bandits, P. R., J. T., 2010]

Find r arms $X_{b_1}, \cdots, X_{b_r}$ which form a maximal linearly independent system.
In each cycle c:

1. Exploration (r periods): Play arm X_{b_k}, and observe the reward $Y^{X_{b_k}}(c)$. Compute the ordinary least squares estimate $\hat{\theta}(c)$.
2. Exploitation (c periods): Play the greedy arm $G(c) = \arg\max_X X \cdot \hat{\theta}(c)$ for c periods.

But it may have large constant in front of the order of its regret.

"Optimal Algorithm"

The Phased Exploration and Greedy Exploitation (PEGE) algorithm has regret $\Omega\left(r\sqrt{T}\right)$ under some regularity conditions.

PEGE [Linearly Parameterized Bandits, P. R., J. T., 2010]

Find r arms $X_{b_1}, \cdots, X_{b_r}$ which form a maximal linearly independent system.

In each cycle c:

1. Exploration (r periods): Play arm X_{b_k}, and observe the reward $Y^{X_{b_k}}(c)$. Compute the ordinary least squares estimate $\hat{\theta}(c)$.
2. Exploitation (c periods): Play the greedy arm $G(c) = \arg\max_X X \cdot \hat{\theta}(c)$ for c periods.

But it may have large constant in front of the order of its regret.

"Optimal Algorithm"

PHASED EXPLORATION AND GREEDY EXPLOITATION (PEGE)

Description: For each cycle $c \geq 1$, complete the following two phases.

1. **Exploration (r periods):** For $k = 1, 2, \ldots, r$, play arm $\mathbf{b}_k \in \mathcal{U}_r$ given in Assumption 1(b), and observe the reward $X^{\mathbf{b}_k}(c)$. Compute the OLS estimate $\widehat{\mathbf{Z}}(c) \in \mathbb{R}^r$, given by

$$\widehat{\mathbf{Z}}(c) = \frac{1}{c}\left(\sum_{k=1}^{r}\mathbf{b}_k\mathbf{b}_k'\right)^{-1}\sum_{s=1}^{c}\sum_{k=1}^{r}\mathbf{b}_k X^{\mathbf{b}_k}(s) = \mathbf{Z} + \frac{1}{c}\left(\sum_{k=1}^{r}\mathbf{b}_k\mathbf{b}_k'\right)^{-1}\sum_{s=1}^{c}\sum_{k=1}^{r}\mathbf{b}_k W^{\mathbf{b}_k}(s) ,$$

where for any k, $X^{\mathbf{b}_k}(s)$ and $W^{\mathbf{b}_k}(s)$ denote the observed reward and the error random variable associated with playing arm $\mathbf{b}_k$ in cycle s. Note that the last equality follows from Equation (1) defining our model.

2. **Exploitation (c periods):** Play the greedy arm $\mathbf{G}(c) = \arg\max_{\mathbf{v}\in\mathcal{U}_r} \mathbf{v}'\widehat{\mathbf{Z}}(c)$ for c periods.

Exponentiated Gradient Algorithm (EGA)

EGA

- Initialize $w_1 = \left(\frac{1}{N}, \ldots, \frac{1}{N}\right), \gamma = \min\left\{1, \sqrt{\frac{N \log N}{(e-1)\Delta T}}\right\}$
- FOR t from 1 to T
  - Algorithm randomly picks i_t with probability $P_t(i_t) = (1-\gamma)w_{t,i} + \gamma/N$
  - Arms incur losses $\Delta_{t,1} \ldots \Delta_{t,N}$
  - Algorithm observes and incurs loss Δ_{t,i_t}
  - Algorithm updates w for bandit i_t as
    $$w_{t+1,i_t} = w_{t,i_t} \exp\left(-\eta \Delta_{t,i_t}/P(i_t)\right)$$
    Then normalize w_{t+1} so that $\sum_j w_{t+1,j} = 1$.

Upper Confidence Bound (UCB)

UCB

Given $\theta \sim N(\mu_0, \Sigma_0)$, for t from 1 to T:

1. Play arm $X_{i_t} = \arg\max \left\{ \mu_{t-1} \cdot X_i + 1.96 X_i' \sum_{t-1} X_i \right\}$

2. Calculate μ_t and Σ_t based on reward Y_t, arm X_{i_t}, μ_{t-1}, and Σ_{t-1}.

$$
\begin{aligned}
P(w|y) &\propto P(y|w)P(w) \\
P(w|y) &\sim N(\mu, S) \\
S^{-1} &= S_0^{-1} + \frac{1}{\sigma^2} X^T X \\
\mu &= S\left(S_0^{-1}\mu_0 + \frac{1}{\sigma^2} X^T y\right)
\end{aligned}
$$

Our Goal

- Evaluate the performance of the existing approaches.
- Develop hybrid methods for specific conditions.
- Find a way to map the user's rating to a compact set, say integers from 0 to 5.

Thanks!!

Any Questions?