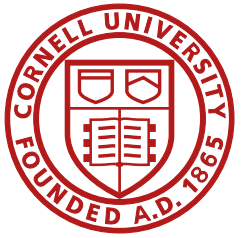

Correlated Multi-armed Bandits

CS 6780 Advanced Machine Learning

Zhengdi Shen
Bangrui Chen
Saul Toscano Palmerin

April 24, 2015

Bandits

- Pull arms sequentially so as to maximize the total expected reward, e.g, gambler faces at a row of slot machines.
  Robbins in 1952, realizing the importance of the problem.

Bandits

use Yelp example?

- Pull arms sequentially so as to maximize the total expected reward, e.g, gambler faces at a row of slot machines. Robbins in 1952, realizing the importance of the problem.
- The crucial tradeoff the gambler faces at each trial is between "exploitation" of the machine that has the highest expected payoff and "exploration" to get more information about he expected payoffs of the other machines.

Independent Arms

- Rewards for each arm are generally assumed to be independent of each other.

Independent Arms

- Rewards for each arm are generally assumed to be independent of each other.
- A theorem, the Gittins Index published first by John C. Gittins gives an optimal policy in the Markov setting for maximizing the expected discounted reward.

Independent Arms

- Rewards for each arm are generally assumed to be independent of each other.
- A theorem, the Gittins Index published first by John C. Gittins gives an optimal policy in the Markov setting for maximizing the expected discounted reward.
- This assumption enables us to consider each arm separately, but Lai and Robbins proved the regret under an arbitrary policy increases linearly with number of arms. Most policies that assume independence require each arm to be tried at least once, and are impractical in settings involving many arms.

Dependent Arms

- What if they are dependent? E.g., ads on similar topics, using similar text/phrases, should have similar rewards.

Dependent Arms

- What if they are dependent? E.g., ads on similar topics, using similar text/phrases, should have similar rewards.
- In such a setting, the information obtained from puling one arm can change our understanding of other arms. Here, we want a policy whose regret is independent of the number of arms.

Linearly Parameterized Bandits

- Mersereau proposed a simple model that the reward of each arm depends on a single random variable z, with a known prior distribution.

Linearly Parameterized Bandits

- Mersereau proposed a simple model that the reward of each arm depends on a single random variable z, with a known prior distribution.
- Since the reward of each arm depends on a single random variable, the mean rewards are perfectly correlated.

Linearly Parameterized Bandits

- Mersereau proposed a simple model that the reward of each arm depends on a single random variable z, with a known prior distribution.
- Since the reward of each arm depends on a single random variable, the mean rewards are perfectly correlated.
- Under certain assumptions, the cumulative Bayes risk over T periods under a greedy policy admits an $O(\log(T))$ upper bound, independent of the number of arms.

Linearly Parameterized Bandits

- Mersereau proposed a simple model that the reward of each arm depends on a single random variable z, with a known prior distribution.
- Since the reward of each arm depends on a single random variable, the mean rewards are perfectly correlated.
- Under certain assumptions, the cumulative Bayes risk over T periods under a greedy policy admits an $O(\log(T))$ upper bound, independent of the number of arms.
- In this project, we are going to consider a model that the reward of each arm depends linearly on a multivariate random variable, with a known prior distribution.

A More Practical Motivation

For a newly registered user on Yelp, how should yelp select the forwarding restaurants at each time so that it can maximize the expected average rating of this new user?

Problem Setting

- We represent each restaurant with a 20 dimensional binary vector, and the 20 features are Pizza, Sandwiches, Food, Nightlife, American(new), Bars, American(traditional), Mexican, Chinese, Italian, Japanese, Fast Food, Burgers, Breakfast and Brunch, Coffee and Tea, Delis, Indian, Thai, Sushi Bars, Mediterranean and Asian Fusion. Denote the set of arms as U^{20}.

Problem Setting

- We represent each restaurant with a 20 dimensional binary vector, and the 20 features are Pizza, Sandwiches, Food, Nightlife, American(new), Bars, American(traditional), Mexican, Chinese, Italian, Japanese, Fast Food, Burgers, Breakfast and Brunch, Coffee and Tea, Delis, Indian, Thai, Sushi Bars, Mediterranean and Asian Fusion. Denote the set of arms as U^{20}.
- We assume each user has a user preference vector θ corresponding to the 20 different features listed above, with $\theta \sim N(\mu, \Sigma)$ with μ and Σ known (This can be calculated from the historical data).

Problem Setting

- Y_t: the reward of playing arm $\mathbf{u} \in U^{20}$ in period t, which is given by

$$Y_t = \mathbf{u}^{'}\theta + W_t$$

where $\mathbf{u}^{'}\theta$ is the inner product between the vector $\mathbf{u} \in R^{20}$ and the random vector $\mathbf{Z} \in \mathbb{R}^{20}$. $\{W_t : t \geq 1\}$ is i.i.d distributed with $N(0, \sigma^2)$.

Problem Setting

- Y_t: the reward of playing arm $\mathbf{u} \in U^{20}$ in period t, which is given by

$$Y_t = \mathbf{u}^{'}\theta + W_t$$

  where $\mathbf{u}^{'}\theta$ is the inner product between the vector $\mathbf{u} \in R^{20}$ and the random vector $\mathbf{Z} \in \mathbb{R}^{20}$. $\{W_t : t \geq 1\}$ is i.i.d distributed with $N(0, \sigma^2)$.
- $\mathscr{H}_t$: the set of possible histroies until the end of period t.

The Model

- $\psi = (\psi_1, \psi_2, \cdots)$: Policy ψ is a sequence of functions such that $\psi_t : \mathscr{H}_{t-1} \to \mathbb{U}_{20}$ selects an arm in period t based on the history until the end of period t-1.

The Model

- $\psi = (\psi_1, \psi_2, \cdots)$: Policy ψ is a sequence of functions such that $\psi_t : \mathscr{H}_{t-1} \to \mathbb{U}_{20}$ selects an arm in period t based on the history until the end of period t-1.
- For any policy ψ and $\mathbf{z} \in \mathbb{R}^{20}$, the T-period cumulative regret under ψ given $\theta = \theta_0$, denoted by $Regret(\theta_0, T, \psi)$, is defined by

$$Regret(\theta_0, T, \psi) = \sum_{t=1}^{T} E\left[\max_{\mathbf{v} \in \mathbb{U}_{20}} \mathbf{v}'\theta_0 - \mathbf{U}_t'\theta_0 | \theta = \theta_0\right]$$

where for any $t \geq 1$, $\mathbf{U}_t \in \mathbb{U}_r$ is the arm chosen under ψ in period t.

Our Goal

- Find an heuristic maximize the new user's rating.

Our Goal

- Find an heuristic maximize the new user's rating.
- Prove the regret bound of our heuristic algorithm.

EGA

Exponential Gradient Algorithm

-
-

UCB

Upper Confidence Bound

-
-

Assumption

Assumption 1

- There exists a positive constant σ_0 such that for any $r \geq 2$, $\mathbf{u} \in \mathbb{U}^r$, $t \geq 1$ and $x \in \mathbb{R}$, we have $E[e^{xW_t^u}] \leq e^{\frac{x^2\sigma_0^2}{2}}$.
- There exists positive constants $\bar{u}$ and λ_0 such that for any $r \geq 2$,

$$\max_{u \in \mathbb{U}^r} \|\mathbf{u}\| \leq \bar{u}$$

and the set of arms $\mathbb{U}_r \subset \mathbb{R}^r$ has r linearly independent elements $\mathbf{b}_1, \cdots, \mathbf{b}_r$ such that $\lambda_{\min}(\sum_{k=1}^r \mathbf{b}_k \mathbf{b}_k^{'}) \geq \lambda_0$.

PEGE

Phased Exploration and Greedy Exploitation

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

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

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

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

where for any k, $X^{b_k}(s)$, and $W^{b_k(s)}$ denote the observed reward and the error random variable associated with playing arm $\mathbf{b}_k$ in cycle s.

PEGE

Phased Exploration and Greedy Exploitation

- **Exploitation (c periods)**: Play the greedy arm $\mathbf{G}(c) = \arg\max_{v \in \mathbb{U}^r} \mathbf{v}' \hat{\mathbf{Z}}(c)$ for c periods.

Theorem

Under some regularity conditions, the regret of this algorithm is $O(r\sqrt{T})$, where r is the dimension of the arm.

Question?

Thanks for your time!