CS 6780 Research Project: Multi-armed Bandits with Dependent Arms

Bangrui Chen BC496@CORNELL.EDU
Saul Toscano Palmerin ST684@CORNELL.EDU
Zhengdi Shen ZS267@CORNELL.EDU

1. Motivation

We are interested in the multi armed bandits problem with correlated arms. Theoretically this problem can be solved using dynamic programming, however it usually suffers from the curse of dimensionality when the dimension of the arm is high. There are two well known heuristic algorithms for this problem which are Exponential Gradient algorithm and the upper confidence bound algorithm. In this project, we hope to investigate the combination of these two different algorithms.

2. Problem Formulation

We have a finite set $\mathcal{U}_r = \{\mathbf{u}_1, \cdots, \mathbf{u}_m\} \subset \mathbb{R}^r$ that corresponds to the set of arms, where $r \geq 2$. For any time $t = 1, 2, \cdots, T$, we are asked to pick one arm X_t. The reward Y_t of playing arm $X_t \in \mathcal{U}_r$ in period t is given by

$$Y_t = \theta \cdot X_t + \epsilon_t,$$

where $\epsilon_t \sim N(0, \sigma^2)$ is the measurement error with σ known. Here θ is an unknown random vector, which is drawn from a multivariate normal distribution with mean μ and variance Σ. We further assume μ and Σ are known.

For a fixed time period T, the goal of this problem is to find a strategy π to maximize the following expression

$$E^{\pi}\left[\sum_{t=1}^{T} Y_t\right]. \tag{1}$$

Or equivalently, we are trying to find a policy that can minimize the Bayes risk under π:

$$\text{Risk}(T, \pi) = E\left[\text{Regret}(\theta, T, \pi)\right], \tag{2}$$

where the cumulative regret is defined as the following:

$$\text{Regret}(\theta_0, T, \pi) = \sum_{t=1}^{T} E\left[\max_{X \in \mathcal{U}_r} X \cdot \theta_0 - X_t \cdot \theta_0 | \theta = \theta_0\right]. \tag{3}$$

3. PEGE

PEGE algorithm

Algorithm 1 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
2. **Exploitation (c periods)**

4. Heuristic Algorithms

There are two well known algorithms for this problem, which are Exponential Gradient algorithm and the upper confidence bound algorithm.

Notations:

x_i:	feature of the recommended restaurant at step i, binary vector
y_i:	rate given by the user at step i, range is $(0, 5)$
θ:	user's preference
μ_0:	prior knowledge of $E[\theta]$
Σ_0:	prior knowledge of Cov(θ)

At each step, we assume the user's rating $y_i = \theta \cdot x_i + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

4.1. Upper Confidence Bound

At step $t+1$, we have knowledge $\mu_0,\ \Sigma_0,\ (x_1, y_1), \cdots, (x_t, y_t)$. The distribution

$$\theta | [\mu_0,\ \Sigma_0,\ (x_1, y_1), \cdots, (x_t, y_t)] \sim \mathcal{N}(\mu_t, \Sigma_t)$$

where

$$\Sigma_t^{-1} = \Sigma_0^{-1} + \frac{1}{\sigma^2} X_t^T X_t \qquad = \Sigma_t^{-1} + \frac{1}{\sigma^2} x_t x_t^T \tag{4}$$

$$\mu_t = \Sigma_t \left(\Sigma_0^{-1}\mu_0 + \frac{1}{\sigma^2} X_t^T y\right) \quad = \Sigma_t \left(\Sigma_{t-1}\mu_{t-1} + \frac{1}{\sigma^2} x_t y_t\right) \tag{5}$$

where

$$X_t = \begin{pmatrix} x_1^T \\ \vdots \\ x_t^T \end{pmatrix}$$

Then, for each restaurant r, we suppose its feature vector is $x^{(r)}$. And its expected rating and variance are

$$E[x^{(r)} \cdot \theta] = x^{(r)} \cdot \mu_t, \quad Var(x^{(r)} \cdot \theta) = (x^{(r)})^T \Sigma_t x^{(r)}$$

The restaurant we will recommend at step $t+1$ is

$$r_t = \arg\max_r E[x^{(r)} \cdot \theta] + 1.96\sqrt{Var(x^{(r)} \cdot \theta)}$$

5. Numerical Experiment

In this simulation, we use the yelp academic dataset. The goal of this simulation is to find the favorite restaurant categories for a new user. There are 4596 restaurants in the dataset and each restaurant belongs to one or multiple categories. We first find the top twenty categories that has most restaurants, which are Pizza, Sandwiches, Food etc, and use those 20 categories as our feature. For each restaurant, if it belongs to certain category, then the corresponding element of its feature vector is 1 and 0 otherwise. So the feature vector of each restaurant is a 20 dimensional binary vector.

For each user, we calculated his user preference vector based on his rating and the restaurants' feature vectors that he rated using ridge regression (since there are not too many ratings, ordinary linear regression doesn't work here due to singularity). Then we calculated the sample mean and the sample variance of all users' preference vector and denote them as (μ, Σ). We further assume that for each user's preference vector $\theta \sim N(\mu, \Sigma)$ and generate new user from this distribution.

6. Possible Applications

Recommender systems want to find the preference that a user would give to a subset of a finite set of items. They're widely applied to different problems. For example, they're used in Netflix where there are thousands of movies and TV episodes. The biggest challenge of these problems is that there are millions of objets and hundreds of million of users, and so it's necessary to find a model that performs well and be sufficiently fast.

References

1] X Zhao, P Frazier, *Exploration vs. Exploitation in the Information Filtering Problem*