Multi-armed Bandit Problem Summary

LIU Weizhi[*]

2014-09-12

Contents

1 Multi-armed Bandit Problem

1.1 Input

1.1.1 Definition

1. Multiarmed-Bandit Problem You are supposed to **decide which order to play, how many times to play** for K slot machines with

[*] weizhiliu@nus.edu.sg

diffirent reward/loss distribution for each in order to **maximize your final reward** in **limited times**.

2. Regret The expected difference between the reward sum associated with an optimal strategy and the sum of the collected rewards

3. Online/Offline Algorithm "In computer science, an online algorithm is one that can process its input piece by piece in a serial fashion, i.e., in the order that the input is fed to the algorithm, without having the entire input available from the start. In contrast, offline algorithm is given the whole problem data from the begin and is required to output an answer which solves the problem at hand."[1]

1.1.2 Variations

- Reward Generation
  1. Bernoulli multi-armed bandit (reward ˜ Bernoulli)
  2. Restless bandit problem (reward ˜ Markov machine)
- Contextual
- Expert Opinions

1.1.3 Bandit Strategies

- Optimal solutions
- Approximate solutions
  - Semi-uniform strategies
    - Epsilon-greedy strategy
    - Epsilon-first strategy
    - Epsilon-decreasing strategy
    - Adaptive epsilon-greedy strategy based on value differences
    - Contextual-Epsilon-greedy strategy
  - Probability matching strategies (Thompson sampling / Bayesian Bandits)
    - Algorithm #Bayesian Bandits Algorithm Description
      1. Sample the selection probabilities for all bandit

[1] Please refer to http://en.wikipedia.org/wiki/Online_algorithm

        2. Select the bandit with largest selection probability
        3. Observe the result of pulling this bandit, and update your prior on this bandit.
        4. return to 1
    - Pricing strategies
    - Strategies with ethical constranits

1.1.4 Applications

1. Clinical trials
2. Adaptive routing
3. Online Advertisement

1.1.5 Useful Resources

1.2 Process

1.2.1 Why is this problem hard to solve?

- Your resources are limited which means you have to effectively utilize your every try.
- You are facing stochastic reward which means risk and profit.
- You are supposed to **analyze history performance** of each slot machine to guide **your next try**.
    - Scenairo 1 - simulation times are small (not significant statistically)
        - On the one hand, slot machine performed best may get lucky or there might be much better alternative to be explored.
        - On the other hand, those with bad performance may get unlucky which means statistically, they might give you more reward in the future.
        - **You might need more tries to evaluate the variance of reward for each slot machine.**
    - Scenairo 2- simulation times are large (significant statistically for some slot machine)

  - If you insist on those with statistically significant reward slot machine, you might loss the opportunity to explore a better alternative.
  - Nevertheless, you have to bear with risk when you try to explore those slot machine with statistically insignificant reward.
  - **exploration and exploitation dilemma**
- Conclusion
  - exploration in Scenairo 1
  - exploiation in Scenairo 2
  - Should we treat the relationship between exploration and exploiation in a absolutely discrete case or mixture case with preference in different time?
    - Discrete Case
    - Mixture Case

1.2.2 How to transform the search process from exploration to exploiation based on the history performance?

Feynman's restaurant problem

1.3 Output

1.3.1 Bandit Strategies

- Bayesian Bandits
  - Simple case with bernoulli multi-armed bandit [2]
    - Assumption
      1. There are K bandits with each's reward obeys Bernoulli(p_k) distribution
      2. The initial prior distribution of each p_k is Uniform[0,1]
    - Algorithm
      please see here.
    - Analyze

[2] Please refer to this website http://camdp.com/blogs/multi-armed-bandits

1. The initial priors are Beta($\alpha = 1$, $\beta = 1$)(a uniform distribution), and the observed reward R (0 or 1) is Binomial, the posterior is a Beta($\alpha = 1 +$ R, $\beta = 1 + 1$ - R). (please see Conjugate prior)

   - · Conjugate prior According to Bayesian formula

     $$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

     , if we can find a prior $P(\theta)$ and likelihood function $P(X|\theta)$ such that the formula of prior $P(\theta)$ and $P(\theta|X)$ has the same distribution (with different parameteres), then $P(\theta)$ and likelihood function $P(X|\theta)$ are called conjugate.

2. Evaluation of strategy performance

   - · Total Regret of a strategy

     $$Regret = Tp^* - \sum_{t=1}^{n} R_{i,t}$$

   - · Expected total regret (Bound for any sub-optimal strategy's expected regret)

     $$E[Regret] = \Omega(\log(T))$$

1.3.2 IE5504 Project

- Research Proposal
  - Problem
    - ∗ Maximize your reward with limited tries given K slot machines whose reward distribution obeys Bernoulli distribution with different p$_k$.
  - Strategy
    1. Initialize the success rate for all bandits and generate a success/failure (1/0) sequence with length n for each bandit.
    2. Calculate the APCS (Approximate Probability of Correct Selection [3]) of all bandits based on their prior distribution.

[3] Please refer to **Stochastic Simulation Optimization - An Optimal Computing Budget Allocation** P37

3. Find the bandit B with highest APCS (if not unique, then uniformly randomly choose a bandit)
4. Observe the reward of bandit B
5. Update the posterior distribution of bandit B
6. Stop if tries run out otherwise return to 1

- Evaluation

  1. Pseudo Regret (given $\mathrm{R}_{\mathrm{i,t}}$ for all i and t)

$$Pseudo\ Regret = \max_i \sum_{t=1}^{n} R_{i,t} - \sum_{t=1}^{n} R_{I_t,t}$$

  where I_t is the bandit selected at time t.

- Todo

  1. Simulation Pattern - Batch simulation or Sequence simulation
  2. Selection criteria - Stochastic Ordering (Probability of correct selection - allocation rule [4]) or Sample Ordering (The largest sample) or Probability of correct selection for each bandit
  3. 2 * 2 experiments according to different scenairo of simulation pattern and selection criteria
  4. Apart from bayesian bandit, find more strategy to solve this multi-armed bandit problem.
  5. Visualization of posterior distribution against simulation time

[4] Please refer to **Stochastic Simulation Optimization - An Optimal Computing Budget Allocation** P46