Decision making in 2-arm bandit problem

Hannah Chen
hechen@eng.ucsd.edu

Sonali Rahagude
srahagud@ucsd.edu

Sneha Venkatesh Yelimeli
svyelime@eng.ucsd.edu

Chun Fan
c9fan@ucsd.edu

Yashodhan Karandikar
ykarandi@eng.ucsd.edu

Abstract

1 Introduction

1.1 Motivation

2-arm bandit problems, markov decision process

2 Background

2.1 Related Work

Shuannan's work

The stochastic multi-armed bandit problem is an important model for studying the exploration-exploitation trade-off in reinforcement learning.The goal of the decision-maker here is to obtain the maximum number of rewards over a number of trials. This report presents an empirical study of the some popular 2-armed bandit algorithms. Our work is motivated from the paper "Psychological models of human and optimal performance in bandit problems." The paper presents a variety of existing heuristic algorithms for finite horizon multi-arm bandit problems. It then builds on the idea of latent state modeling and presents a new model for the decision-making process in the bandit problem. It uses a Bayesian model in which how people balance exploration with exploitation depends on their assumptions about the distribution of reward rates.

3 Design

3.1 Formulation of MDP

states definition, recursive formulation, decision rule

3.2 Environments

3.3 Generation of Optimal data

3.3.1 Algorithm

3.4 Inference of Parameters for Heuristics

For inferring parameters for different heuristic methods, we use the optimal data and do a grid search.

3.4.1 Inference from Optimal Data

In inferring from the optimal decision data, we assume that the optimal decision-maker has perfect knowledge of the environment, i.e. the α and β values are known. This makes sense since our goal of inference from the optimal data is to characterize the decision making process for given heuristics and see how well they fit the best possible decision. For each of the heuristic methods, we infer the parameters of method using grid search on the parameter. Thus, we infer the parameter ϵ for $\epsilon - greedy$ and $\epsilon - decreasing$ methods, parameter γ for WSLS method and parameter τ for $\tau - switch$ model. Once we have inferred the parameters, we then use these parameters along with the same reward rates to general the decision. A comparison of these decisions with the optimal ones gives us an idea of how well the different methods perform in the setting. Algorithm below describes the entire process,

3.4.2 Inference from Human Data

3.4.3 Common Algorithm Input : (alpha, beta, decisions, rewards), output: (parameter value)

Algorithm 1 LDA Generative process with collapsed Gibbs Sampling

```
Input: words $\boldsymbol{w} \in$ documents $\boldsymbol{d} \in [1, D]$
randomly initialize $\boldsymbol{z}$ and increment counters
for iteration $i \in [1, epoch]$ do
    for document $d \in [1, D]$ do
        for word $\in [1, N_d]$ do
            $topic \leftarrow z[word]$
            decrement counters according to document $d$, $topic$ and $word$
            for $k \in [1, K]$ do
                calculate $p(z = k|.)$ using Gibbs equation
            end for
            $newTopic \leftarrow$ sample from $p(z|.)$
            $z[word] \leftarrow newTopic$
            decrement counters according to document $d$, $newTopic$ and $word$
        end for
    end for
end for
```

4 Experimental Setup

The goals of our experimental setup are thus 3-fold: generate optimal data for the bandit problem given different environment settings (plentiful, neutral and sparse), infer parameter values for different bandit algorithms from both optimal as well as human data and test the inferred models for decision-making given the same reward rates as in the optimal setting. We perform 50 experiments on 3 different environments plentiful (α = 4,β = 2), neutral(α = 1, β = 1) and scarce (α = 2,β = 4). Each experiment consists of 8 trials and the reward rates for each arm θ_k in a given experiment is drawn from the beta distribution i.e. $\theta_k \sim Beta(\alpha, \beta)$. Thus, we draw a total of 100 reward rates in the setting.

For the i^{th} trial in the t^{th} experiment, if the arm k is chosen, the reward R_i^t is determined by the outcome of a Bernoulli using the reward rate θ_k^t, so $R_i^t \sim Bernoulli(\theta_k^t)$

5 Lessons learnt

6 Results

6.1 Parameters by inference from optimal data

add table for decision vectors and reward vectors for different values of alpha, beta

6.2 Parameters by inference from human data

6.3 Match Percentage

6.3.1 Comparison with optimal data

a simple comparison of match percentage of different heuristics with optimal data

6.3.2 Comparison with human data

do we want to include comparison with human data?

7 Conclusions

A Appendix

References

[1] Charles Elkan. Clustering documents with an exponential-family approximation of the dirichlet compound multinomial distribution. In *Proceedings of the 23rd international conference on Machine learning*, pages 289–296. ACM, 2006.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[3] Lindsay I Smith. A tutorial on principal components analysis. *Cornell University, USA*, 51:52, 2002.

[4] Gregor Heinrich, Jörg Kindermann, Codrina Lauth, Gerhard Paaß, and Javier Sanchez-Monzon. Investigating word correlation at different scopes—a latent concept approach. In *Workshop Lexical Ontology Learning at Int. Conf. Mach. Learning*, 2005.

[5] Marina Meilă. Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, 98(5):873–895, 2007.

[6] Gregor Heinrich. Parameter estimation for text analysis. Technical report, 2004.