contextual: Simulating Contextual Multi-Armed Bandit Problems in R

Robin van Emden
JADS

Eric Postma
Tilburg University

Maurits Kaptein
Tilburg University

Abstract

Many statistical and reinforcement learning decision problems can be framed as (contextual) multi-armed bandit problems. A comprehensiveness that has made them particularly useful in the optimization of many real-world sequential decision problems. This relevancy has led to a proliferation of Bandit algorithms or policies, and a large body of analytical research. At the same time, there are, as of yet, few options to compare these Bandit policies in a standardized way through simulation and real-life offline data. This article therefore introduces the R package contextual, which provides an integrated and easily extendable approach to the comparison of different Bandit policies and generalizations.

Keywords: contextual multi-armed bandits, simulation, sequential experimentation, R.

A vignette for the van Emden, Kaptein, and Postma (2018) paper.

1. Introduction

In a Multi-Armed Bandit (MAB) problem, named after the one-armed slot machines of yore, an agent makes use of an algorithm or "policy" to optimize the reward they receive over time in a sequential decision problem with limited information. That is, a MAB policy advices an agent when to explore new options and when to exploit known ones – where, importantly, for each decision, at each time step t, the only information the agent acquires is the reward for that decision. The agent remains in the dark about the potential rewards of the unchosen option - or any other information but their current and past rewards and choices, for that matter.

As a matter of fact, MAB problems reflect dilemmas we all encounter on a daily basis: should you stick to what you know and get an expected result ("exploit") or do you choose something you don't know all that much about and potentially learn something new ("explore")?

- Do you feed your next coin to the one-armed bandit that paid out last time, or do you test your luck on another arm, on another machine?
- When going out to dinner, do you explore new restaurants, or do you exploit familiar ones?
- Do you stick to your current job, or explore and hunt around?
- Do I keep my current stocks, or change my portfolio and pick some new ones?

- As an online marketer, do you try a new ad, or keep the current one?
- As a doctor, do you treat your patients with tried and tested medication, or do you prescribe a new and promising experimental treatment?

Though MAB models are already powerful of their own accord, a recent generalization, known as the contextual Multi-Armed Bandit (cMAB), adds one important element to the equation: in addition to past decisions and their rewards, cMAB agents are able to make use of side information on the state of the world at each t, right before making their decision. In other words, an agentthat follows the advice of a cMAB policy may decide differently in different contexts.

This access to side information has proven to make cMAB algorithms an even better fit to many real-life decision problems. Do you show a certain add to returning customers, to new ones, or both? Do you prescribe a different treatment to male patients, female patients, or both? In the real world, it appears almost no choice exists without its context. So it may be no surprise that cMAB algorithms have found many uses: from recommender systems and advertising to health apps and personalized medicine. This practical applicability has led to a multitude of policies, each with their own strengths and weaknesses.

Still, though cMAB algorithms have gained much traction in both research and industry, they have mostly been studied mathematically and analytically – comparisons on simulated, and, importantly, real-life large-scale offline "partial label" data sets have been lacking. To this end, the current paper introduces the contextual R package. A package that aims to facilitate the simulation, offline comparison, and evaluation of (Contextual) Multi-Armed bandit policies. Though there exists one R package for basic MAB analysis, there is, as of yet, no extensible and widely applicable R package that is able to analyze and compare, respectively, basic K-armed, Continuum, Adversarial and Contextual Multi-Armed Bandit Algorithms on either simulated or online data.

In section 2, this paper will continue with a more formal definition of MAB and a CMAB problems. In section 3, we will continue with an overview of contextual's general implementation. In section 4, we list our implemented polices, and simulate a MAB and a cMAB policy. In section 5, we demonstrate how easy it is to add and simulate your own custom policy. In section 6, we replicate two papers, thereby demonstrating how to test policies on offline data sets. Finally, in section 7, we will go over some of the additional features in the package, and conclude with some comments on the current state of the package and possible enhancements.

2. Contextual Multi-Armed Bandits

In the canonical multi-armed bandit (MAB) problem a gambler faces a number of slot machines, each with a potentially different payoff. It is the gamblers goal to make as much profit (or, in the case of gambling, as little loss) as possible by sequentially choosing which machine to play, learning from the observations as she goes along.

3. Implementation of the contextual R package

In the canonical multi-armed bandit (MAB) problem a gambler faces a number of slot machines, each with a potentially different payoff. It is the gamblers goal to make as much profit

(or, in the case of gambling, as little loss) as possible by sequentially choosing which machine to play, learning from the observations as she goes along.

4. A basic example

```
library("contextual")

bandit      <- BasicBandit$new()

bandit$set_weights(c(0.1, 0.9))

policy      <- EpsilonGreedyPolicy$new()

agent       <- Agent$new(policy, bandit)

simulation  <- Simulator$new(agent,
                             horizon = 100L,
                             simulations = 100L,
                             worker_max = 1)

history     <- simulation$run()

Plot$new()$grid(history)
```

For results, see Figure 1 on page 4.

5. Object orientation: extending contextual

The R6 package allows the creation of classes with reference semantics, similar to R's built-in reference classes. Compared to reference classes, R6 classes are simpler and lighter-weight, and they are not built on S4 classes so they do not require the methods package. These classes allow public and private members, and they support inheritance, even when the classes are defined in different packages.

One R6 class can inherit from another. In other words, you can have super- and sub-classes.

Subclasses can have additional methods, and they can also have methods that override the superclass methods. In this example of a custom **contextual** bandit, we'll extend BasicBandit and override the initialize() method..

6. Special features

For instance, quantifying variance..

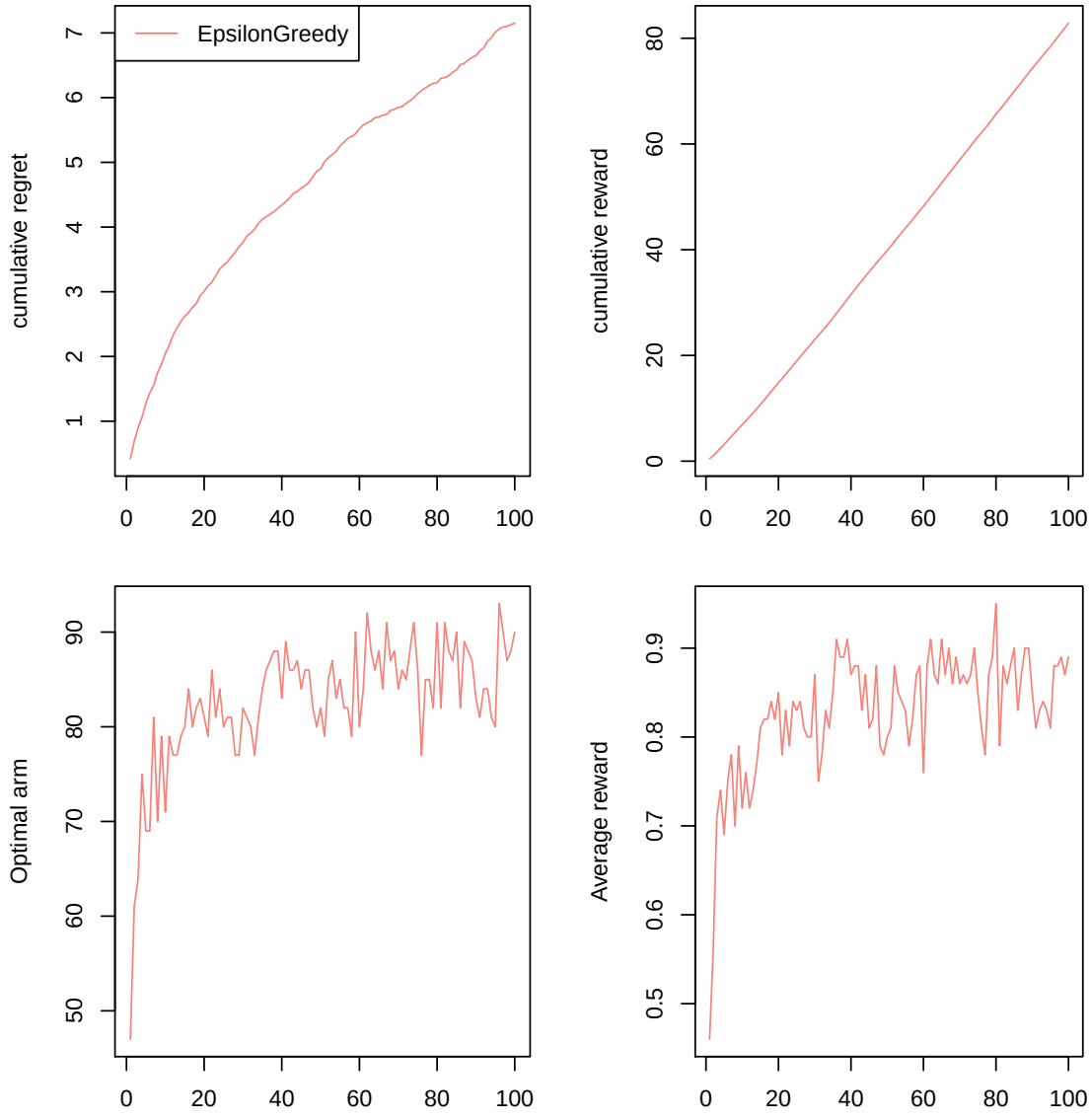

Figure 1: Epsilon Greedy

7. The art of optimal parallelisation

There is a very intersting trade of between the amount of parallelisation (how many cores, nodes used) the resources needed to compute a certain model, and the amount of data going to and fro the cores.

PERFORMANCE DATA ————————————————————

on 58 cores: k3*d3 * 5 policies * 300 * 10000 –> 132 seconds

on 120 cores: k3*d3 * 5 policies * 300 * 10000 –> 390 seconds

—

on 58 cores: k3*d3 * 5 policies * 3000 * 10000 –> 930 seconds

on 120 cores: k3*d3 * 5 policies * 3000 * 10000 –> 691 seconds

8. Extra greedy UCB

In the canonical multi-armed bandit (MAB) problem a gambler faces a number of slot machines, each with a potentially different payoff. It is the gamblers goal to make as much profit (or, in the case of gambling, as little loss) as possible by sequentially choosing which machine to play, learning from the observations as she goes along.

9. Conclusions

The goal of a data analysis is not only to answer a research question based on data but also to collect findings that support that answer. These findings usually take the form of a table, plot or regression/classification model and are usually presented in articles or reports.

10. Acknowledgments

Thanks go to CCC.

References

van Emden R, Kaptein M, Postma E (2018). *contextual: Simulating Contextual Multi-Armed Bandit Problems in R.* Jheronimus Academy of Data Science.

Affiliation:

Robin van Emden
Jheronimus Academy of Data Science
Den Bosch, the Netherlands
E-mail: robin@pwy.nl
URL: pavlov.tech

Eric O. Postma
Tilburg University
Communication and Information Sciences
Tilburg, the Netherlands
E-mail: e.o.postma@tilburguniversity.edu

Maurits C. Kaptein
Tilburg University
Statistics and Research Methods
Tilburg, the Netherlands
E-mail: m.c.kaptein@uvt.nl
URL: www.mauritskaptein.com