Personalization and bandits with applications in health.

Prof. dr. M.C.Kaptein
Professor Data Science & Health
Principal Investigator @ JADS

24 May 2018

About me

- Academic profile:
  - MSc. Economic Psychology, Tilburg University
  - PdEng. User System Interaction, University of Eindhoven
  - Ph.D. Industrial Design, TU/e & Stanford University
  - Post Doc. Marketing, Aalto School of Economics, Helsinki
  - Assistant Professor Artificial Intelligence, Radboud University
  - Assistant Professor Statistics, Tilburg University (Tenured)
  - Professor, Data Science & Health, JADS
- Author:
  - “Persuasion Profiling; how the internet knows what makes you tick” (2014)
  - “Modern Statistical Methods for HCI” (2016, w. J. Robertson)
  - “Hello World — Hello Computer” (2018)

The Multi-Armed Bandit Problem

- Deciding between competing treatments
- Studying the performance of allocation policies
- Discussion

Personalization: the contextual MAB problem

- Personalization
- Solutions
- Approaches in online marketing
- Discussion

Offline policy evaluation

- What would have happened if???
- Offline policy evaluation
- Observational data and causal effect estimation
- Discussion

Conclusions

Section 1

The Multi-Armed Bandit Problem

Deciding between two treatments

- Problem: We have k possible treatments (pills, lifestyle advices, communication, radiation, etc.) with uncertain outcomes. We want to choose the best treatment.
- Standard solution: Try out the treatments for some period of time (or for some number of units n), estimate the outcome, choose the treatment with the highest outcome afterwards.
- Very general problem that we encounter in multiple fields.

Sequentially decide between two treatments

- Effectively we often decide between treatments over a period of time.
- One by one we select a treatment for a unit.
- Our **aim** is to maximize the outcome (e.g., health, survival, etc.) over all treated units.
- **Challenge:** Balance exploration and exploitation.

Sequentially decisions made formal

For $t = 1, \ldots, t = T$

- We select and action a_t. (Often actions $k = 1, \ldots, k = K$, not always).
- Observe reward r_t

Aim: Maximize (expected) cumulative reward $\sum_{t=1}^{T} r_t$

Select actions according to some *policy*
$\pi : \{a_1, \ldots, a_{t-1}, r_1, \ldots, r_{t-1}\} \mapsto a_t$

Or broader: **Study the properties of allocation policies.**

Disclaimer

The abstract problem admittedly ignores a number of important factors:

- Rewards are not immediately observed
- Units are not the same
- Some rewards should be avoided at all costs
- Unethical to use suboptimal treatment

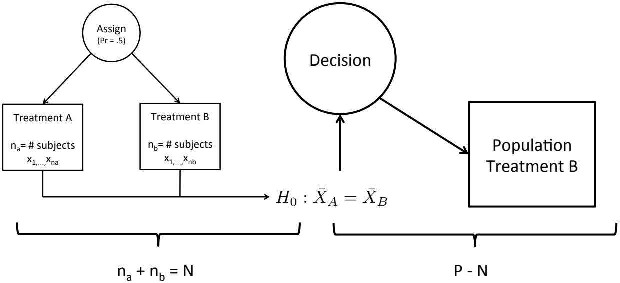
But: It is still useful to study decision policies to get an intuition for the problem. We can always re-introduce the difficulties.

The outcome: expected cumulative regret

Focus on studying expected, cumulative regret: $R = \sum_{t=1}^{T} r_t^* - r_t$ (with expectation over multiple runs and where r_t^* is the reward obtained when playing the optimal policy).

- Optimal policy has regret of 0.
- Policies with non-zero, non-decreasing probability of choosing the wrong action have (asymptotic) linear regret
- Prefer policies with low regret for fixed T, or with specific asymptotic behavior.

ϵ-first



Assign

(Pr = .5)

Treatment A

n_a= # subjects

$x_{1,\dots,}x_{na}$

Treatment B

n_b= # subjects

$x_{1,\dots,}x_{nb}$

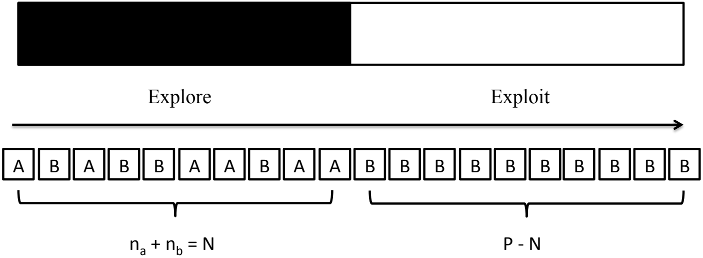
Decision

Population Treatment B

$H_0 : \bar{X}_A = \bar{X}_B$

$n_a + n_b = N$

P - N



Explore

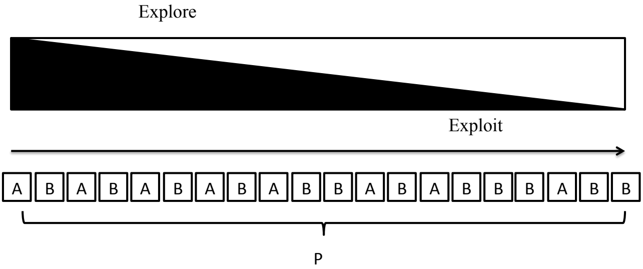
Exploit

A B A B B A A B A A B B B B B B B B B B

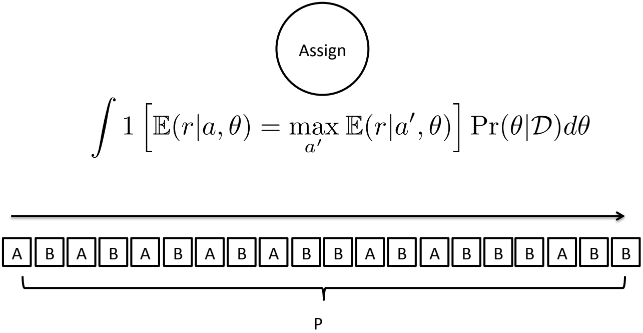
$n_a + n_b = N$

P - N

ϵ-greedy; changing ϵ



Thompson sampling



$$\int 1\left[\mathbb{E}(r|a,\theta)=\max_{a'}\mathbb{E}(r|a',\theta)\right]\Pr(\theta|\mathcal{D})d\theta$$

$$\underbrace{\boxed{A}\boxed{B}\boxed{A}\boxed{B}\boxed{A}\boxed{B}\boxed{A}\boxed{B}\boxed{A}\boxed{B}\boxed{B}\boxed{A}\boxed{B}\boxed{A}\boxed{B}\boxed{B}\boxed{B}\boxed{A}\boxed{B}\boxed{B}}_{\text{P}}$$

Alternatives:

- UCB methods (frequentist Thompson Sampling)
- Bayes Optimal (hard to do for large T)
- Many others in the literature . . .

Comparing policies

- Both ϵ-first and ϵ-greedy have (asymptotic) linear regret: guaranteed outperformed by Thompson sampling & UCB methods in the long-run.
- This is even true with accurate power calculations / “optimal” experiment length in ϵ-first
- For small T we need to be more greedy (explore less)

Intuitions

- Interpret RCT's as one possible policy of choosing treatments in a sequential allocation problem
- Many other policies exist, some demonstrably better
- Balancing exploration exploitation (expected reward vs. variance): RCT effectively over exploits asymptotically
    - Deterministic choices lead to linear regret
    - RCT relatively ok when relatively large T, large n, and $n << T$; poor otherwise.

Note on the length of the RCT

I have skipped the question of the length of the RCT: how to determine n?

- Power calculations a-priori?
  - Need effect size estimates: often not known.
  - These do not include T!
- Lot of recent work on adaptive designs: decide on stopping as data is collected while maintaining (frequentist) characteristics.

Section 2

Personalization: the contextual MAB problem

Personalization: include a context

For $t = 1, \dots, t = T$

- **We observe the context** x_t.
- We select and action a_t.
- Observe reward r_t

Aim remains the same, but problem more challenging: the best action might depend on the context.

Possible solution: alternative problems

One way of solving the contextual bandit problem is to see each context as a new problem, and just solve the bandit problem within that context.

- Explosion of the number of problems.
- No way to “borrow strength” from other context(s) (totally independent)

Is this our current approach to personalization in healthcare?

Better solution: pool information across problems

- Estimate some model to predict $\mathbb{E}(r_t) = f(a_t, x_t)$.
    - Exploitation: choose the action that maximizes $\mathbb{E}(r_t)$.
    - Exploration: choose the action that has a high uncertainty.

Note: personalization as an optimization problem

If $\mathbb{E}(r_t) = f(a_t, x_t)$ was known, we would just pick the best action accordingly.

However, it is not known; thus we need learn it efficiently (without wasting too many trials on exploration)

Is the RCT an effective search strategy? (pairwise evaluation of 2 points in space vs. imposing more structure).

Model for $\mathbb{E}(r_t)$ should be a *causal* model

- Choose actions according to some policy π.
- If π chooses actions uniformly random, the average reward for a given action provides a unbiased estimate of the causal effect of that action.
- **If π depends on x_t this is not the case**
  - Compute propensity score $P(a_t|x_t, \dots) = p_t$ (often known)
  - Correct using inverse propensity score weighing

(Note: with deterministic assignment, IPS estimates not possible: $p_t = 0$ or $p_t = 1$)

Contemporary approach in online marketing

Problem structure:

- Sequentially interact with customers, described by feature vector x_t
- Select a piece of content (ad, news item, product, etc.)
- Observe respons (click, purchase, etc.)

Solution:

- Use as many input features as possible, fit (black-box) model for $\mathbb{E}(r_t) = f(a_t, x_t)$.
- Use (e.g.,) Thompson sampling for probabilistic assignment (and control for p_t!)

Never clear who receives what, no deterministic assignment, no interpretable rules, but **higher reward**.

Contemporary approach in online marketing 2

Combination of:

- Fixed (small) proportion of uniform random exploration.
- Fixed (large) proportion of “best” policy on yesterday’s random exploration dataset.

Software to do this (and previous) available `https://github.com/Nth-iteration-labs/streamingbandit`.

Lessons learned

- Personalization of treatments as a contextual bandit problem
- Efficient search to learn $\mathbb{E}(r_t) = f(a_t, x_t)$
- Cautious of causal effects!
- Modern machine learning / reinforcement learning methods available

RCT within subgroups of patients very poor strategy!

Section 3

Offline policy evaluation

What would have happened if???

Suppose we have data generated according to some policy π, what can we say about the performance of another policy π'?

1. Suppose we can know the data generating mechanism: easy, we just evaluate π' (simulation).
2. Suppose we know π: offline policy evaluation
3. Suppose we do not know π: observational data?

Simulation

- Specify data generating mechanism (including all potential outcomes)
- Examine empirical performance of policies using data generating mechanism
- Often too many assumptions, not externally valid

Software available:
`https://github.com/Nth-iteration-labs/contextual`

Offline policy evaluation using randomized data

If we have data D available from a policy with random uniform allocation we can:

For each t in T:

1. get suggested action $a'_t = \pi'(x_t)$
2. if $a'_t == a_t : R+ = r_t$
3. otherwise: ignore datapoint

Provides an unbiased estimate of the reward R of policy π' for an expected horizon of T/k.

Offline evaluation using propensity scores

The algorithm on the previous slide works because $p_1 = p_2 = \dots$: all actions have the same propensity.

This is not always the case; however, if we know $p_1, p_2, \dots$, we can correct for these propensities using IPS estimator.

Hence, we can effectively use historical data generated using policy π to evaluate π' as long as we know $p_1, \dots$.

(and, we can also use Doubly Robust Estimation methods; controlling for both the propensity as well as the possibly biased mean model.)

What about observational data?

- Observational data: data generated using $\pi^?$: we do not know why—or with what probability—actions are selected.
- Renders data potentially useless: (e.g., $P(a_t|x_t) = 0$)

But, we can try to estimate $P(a_t|x_t)$ using ML methods (or just simple logistic regression). Then we can use offline evaluation!

Relation to causal effect estimation

Question: We currently treat all cancer-type X patients with treatment A; what would happen if we personalize treatments and choose between treatment A, B, or C based on patient characteristics x?

Suppose:

- $R^{\pi_A} = \sum_{t=1}^{T} a_t = A$
- $R^{\pi_{pers}} = \sum_{t=1}^{T} f(x_t)$

Interested in $R^{\pi_A} - R^{\pi_{pers}}$.
However, data generated under $\pi^?$: we need p_t!

Lessons learned

- Collected observational data not necessary useful to evaluate alternative allocation schemes: p_t not known, or 0 or 1.
- But, if p_t is known, or can be estimated, we can use data generated using π to evaluate π'.
- Hence, we can use RCT data to evaluate personalized treatments schemes (given enough data)
- Potentially, if we can *estimate* p_t properly, we can even use observational data to do the same.

Note the *potentially* in the last bullet; this is only true under relatively strict assumptions (SUTVA, correct model for p_t, etc.). But on the other end, we always need to make assumptions...

Section 4

Conclusions

Conclusions

IKNL registry (observational) data potentially useful to evaluate alternative, even personalized, treatment allocation policies.

However, this is not at all simple: naive estimates of (conditional) treatment effects based on observational data can be totally and utterly wrong.

Hope to have shared an alternative view on both RCTs as well as registry data; for more detail, please ask questions!

Contact

m.c.kaptein@uvt.nl