

Contextual bandits

David S. Rosenberg

NYU: CDS

March 12, 2021

Contents

1 The contextual bandit problem

2 Policies

The contextual bandit problem

The contextual bandit

A **contextual bandit problem** proceeds in rounds of the following steps:

- 1 Observe input/**context** x .
 - 2 Take action a .
 - 3 Receive reward $r \in \mathbb{R}$.
- Action a may depends on x and previously observed (x, a, r) triples.

Example: news article recommendation

- Consider a news website.
- Every day there are 10 top stories.
- We want to highlight one for each user.
 - Choice should be **personalized**.
- What is the action?
 - Selecting one of the top 10 stories
- What is the context?
- What is the reward?

The context

- What is the context?
 - Information about the user, if any (e.g. demographics)
 - Geographic location
 - User identifier (we can learn latent features, collaborative filtering style)
- Can we use things about an individual that may change over the rounds of a bandit?
 - possibly as a result of our actions?
 - e.g. recent reading history; e.g. articles (from previous rounds) shared with friends
- No – this would take us out of the contextual bandit framework.
- Reinforcement learning (RL) is a more general framework that allows for this.
 - i.i.d. contexts are replaced by “states” that may evolve over time.
- We'll give a brief introduction to RL in a later module.

The reward

- What can we use as a reward signal?
- Click (Y/N)
- Spent more than 30 seconds on article page (Y/N)
- More complicated function of time spent reading article
- Was article shared or favorited? (Y/N)
- Figuring out the right reward signal is nontrivial.
 - Requires domain understanding.
 - May need tweaking over time.

Context / reward examples

- User 325. $x = \{\text{Likes sports articles}\}$.
- Actions / rewards
 - Action 1: “Tom Brady retirement” **Reward: 10**
 - Action 2: “Player has meltdown after argument” **Reward: 2**
 - Action 3: “Government considers ban for actor using drugs” **Reward: 3**
- User 823. $x = \{\text{Likes human-interest stories}\}$
- Actions / rewards
 - Action 1: “Tom Brady retirement” **Reward: 1**
 - Action 2: “Player has meltdown after argument” **Reward: 5**
 - Action 3: “Government considers ban for actor using drugs” **Reward: 0**

└ The contextual bandit problem

└ Context / reward examples

- User 325. $x = \{ \text{Likes sports articles} \}$
- Actions / rewards
 - Action 1: "Tom Brady retirement" Reward: 10
 - Action 2: "Player has meltdown after argument" Reward: 2
 - Action 3: "Government considers ban for actor using drugs" Reward: 3
- User 823. $x = \{ \text{Likes human-interest stories} \}$
- Actions / rewards
 - Action 1: "Tom Brady retirement" Reward: 1
 - Action 2: "Player has meltdown after argument" Reward: 5
 - Action 3: "Government considers ban for actor using drugs" Reward: 0

- In the terminology of our discussion of causal inference, the reward for each action is a potential outcome.
- We only get to observe the reward corresponding to the action we took (or “treatment” given, in the causal inference terminology).
- Terminology note: Some authors refer to the outcomes we don’t observe as “counterfactual” (e.g. [MW15, Ch. 2]).
- Other authors use “counterfactual” to refer to all the potential outcomes that can happen (e.g. [HR20, p. 4]. And one of these counterfactuals, the observed outcome, is also “factual”.
- Some authors are careful to avoid the word “counterfactual” because of this ambiguity.
- Just be aware of the different usages – it doesn’t matter that much.

The rewards

Conditioned on a context $x \in \mathcal{X}$,

- a reward is generated for each possible action $a \in \mathcal{A} = \{1, \dots, k\}$.
- These k rewards are represented by **reward vector**

$$R = (R(1), \dots, R(k)) \in \mathbb{R}^k.$$

- We only observe one entry: $R(A)$, where A is the action we play.

Probabilistic model for contextual bandit

- Context and reward vector are related:
 - The same action will get different rewards in different contexts.

Stochastic contextual k -armed bandit model

- Context and reward vector $(X, R) \in \mathcal{X} \times \mathbb{R}^k$ **drawn jointly** from P .
- Context and reward pairs are i.i.d. over time:

$$(X, R), (X_1, R_1), \dots, (X_t, R_t) \text{ i.i.d. } \sim P.$$

Action selection

- Action at round t is A_t .
- At beginning of round t , the **history**, or previous **observation sequence** is

$$\mathcal{D}_t = \left((X_1, A_1, R_1(A_1)), \dots, (X_{t-1}, A_{t-1}, R_{t-1}(A_{t-1})) \right).$$

- At round t , action A_t may depend on context X_t and history \mathcal{D}_t .
- Note that we **cannot** say $A_t \perp\!\!\!\perp R_t$ – why?
- Because A_t depends on X_t , and R_t depends on X_t .
 - Information about R_t can propagate to A_t through X_t .

Action and reward are **conditionally** independent given context

We can say that $A_t \perp\!\!\!\perp R_t \mid X_t$ for each t .

└ The contextual bandit problem

└ Action selection

- Action at round t is A_t .
- At beginning of round t , the **history**, or previous **observation sequence** is

$$\mathcal{D}_t = \left((X_1, A_1, R_1(A_1)), \dots, (X_{t-1}, A_{t-1}, R_{t-1}(A_{t-1})) \right).$$

- At round t , action A_t may depend on context X_t and history \mathcal{D}_t .
- Note that we cannot say $A_t \perp\!\!\!\perp R_t$ – why?
- Because A_t depends on X_t , and R_t depends on X_t .
 - Information about R_t can propagate to A_t through X_t .

Action and reward are **conditionally independent** given context

We can say that $A_t \perp\!\!\!\perp R_t \mid X_t$ for each t .

- Note that $A \perp\!\!\!\perp R \mid X$ is the exact counterpart to the “ignorability” assumption in causal inference: $(Y(0), Y(1)) \perp\!\!\!\perp W \mid X$. The reward vector $R = (R(1), \dots, R(k)) \in \mathbb{R}^k$ corresponds to the potential outcome vector $(Y(0), Y(1)) \in \mathbb{R}^2$. The action $A \in \mathcal{A}$ corresponds to the treatment indicator $W \in \{0, 1\}$, and the covariate $X \in \mathcal{X}$ has the same interpretation in each setting.

Stochastic k -armed contextual bandit

Stochastic k -armed contextual bandit

- ① Environment samples **context** and **rewards vector** jointly, iid, for each round:

$$(X, R), (X_1, R_1), \dots, (X_T, R_T) \in \mathcal{X} \times \mathbb{R}^k \text{ i.i.d. from } P,$$

where $R_t = (R_t(1), \dots, R_t(k)) \in \mathbb{R}^k$.

- ② For $t = 1, \dots, T$,

- ① Our algorithm **selects action**/arm $A_t \in \{1, \dots, k\}$ based on X_t and history

$$\mathcal{D}_t = \left((X_1, A_1, R_1(A_1)), \dots, (X_{t-1}, A_{t-1}, R_{t-1}(A_{t-1})) \right).$$

- ② Our algorithm **receives reward** $R_t(A_t)$.

- We **never observe** $R_t(a)$ for $a \neq A_t$.

└ The contextual bandit problem

└ Stochastic k -armed contextual bandit

- It might look cleaner to say that at the beginning of every round, the environment generates $(X_t, R_t) \in \mathcal{X} \times \mathbb{R}^k$ from P . But we want to be very clear that $(X_1, R_1), \dots, (X_T, R_T)$ are
 - generated i.i.d. and are
 - generated before any of the actions A_1, \dots, A_T are generated.

Stochastic k -armed contextual bandit

Environment samples context and rewards vector jointly, i.i.d. for each round:

$$(X, R), (X_1, R_1), \dots, (X_T, R_T) \in \mathcal{X} \times \mathbb{R}^k \text{ i.i.d. from } P,$$

where $R_t = (R_t(1), \dots, R_t(k)) \in \mathbb{R}^k$.

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

Our algorithm selects action/arm $A_t \in \{1, \dots, k\}$ based on X_t and history

$$\mathcal{H}_t = ((X_1, A_1, R_1(A_1)), \dots, (X_{t-1}, A_{t-1}, R_{t-1}(A_{t-1}))).$$

Our algorithm receives reward $R_t(A_t)$.

We never observe $R_t(a)$ for $a \neq A_t$.

Policies

- Policies give some structure to action selection.
- A policy at round t
 - gives a conditional distribution over the action A_t to be taken
 - conditioned on the current context X_t and the history \mathcal{D}_t .
- We'll denote the policy at round t as $\pi_t(\cdot | X_t, \mathcal{D}_t)$.
- Choosing an action according to policy π_t means we choose A_t **randomly** s.t.

$$\mathbb{P}(A_t = a) = \pi_t(a | X_t, \mathcal{D}_t).$$

- Suppose we knew the function

$$r(x, a) = \mathbb{E}[R \mid A = a, X = x],$$

which gives the expected reward for any action a and context x .

- Then optimal policy would be

$$\pi_t(a \mid X_t, \mathcal{D}_t) = \mathbb{1} \left[a = \arg \max_a r(X_t, a) \right].$$

Example: “direct method”

- We don't know $r(x, a)$, but we can use \mathcal{D}_t as training data:

$$\left(\underbrace{(X_1, A_1)}_{\text{input}}, \underbrace{R_1(A_1)}_{\text{response}} \right), \dots, \left(\underbrace{(X_{t-1}, A_{t-1})}_{\text{input}}, \underbrace{R_{t-1}(A_{t-1})}_{\text{response}} \right).$$

- Approximating $r(x, a)$ is a regression problem!
- Let $\hat{r}_t(x, a) = \text{TrainingAlgorithm}(\mathcal{D}_t)$.
- The policy for the **direct method** is defined as

$$\pi_t(a \mid X_t, \mathcal{D}_t) := \mathbb{1} \left[a = \arg \max_a \hat{r}_t(x, a) \right].$$

- This is a **pure exploitation** method.

Some other approaches

- ϵ -greedy is an obvious extension of the direct method.
- Thompson sampling: prior is over models $\hat{r}_t(x, a)$
 - equivalently, prior is over model parameters
- Policy gradient: directly optimizing over the policy to improve expected reward
 - we'll return to this in a few weeks as a warm-up for REINFORCE.

References

- The term contextual bandit was introduced in [LZ07], but the idea has been around much longer.
- A nice history of contextual bandits is given in [TM17], which cites a 1979 paper as the first appearance of contextual bandits.

References I

- [HR20] Miguel A. Hernán and James M. Robins, *Causal inference: What if*, Boca Raton: Chapman & Hall/CRC, 2020,
<https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/>.
- [LZ07] John Langford and Tong Zhang, *The epoch-greedy algorithm for contextual multi-armed bandits*, Proceedings of the 20th International Conference on Neural Information Processing Systems (Red Hook, NY, USA), NIPS'07, Curran Associates Inc., 2007, pp. 817–824.
- [MW15] Stephen L. Morgan and Christopher Winship, *Counterfactuals and causal inference*, 2 ed., Cambridge University Press, 2015.
- [TM17] Ambuj Tewari and Susan A. Murphy, *From ads to interventions: Contextual bandits in mobile health*, Mobile Health, pp. 495–517, Springer International Publishing, 2017.