Policy Gradient for Contextual Bandits

David S. Rosenberg

NYU: CDS

March 26, 2021

Contents

Recap of the contextual bandit setting

[Online] Stochastic k-armed contextual bandit

Stochastic k-armed contextual bandit

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

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

   where $R_t = (R_t(1), \ldots, R_t(k)) \in \mathbb{R}^k$.
2. For $t = 1, \ldots, T$,
   1. Our algorithm **selects action** $A_t \in \mathcal{A} = \{1, \ldots, k\}$ based on X_t and history

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

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

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

Contextual bandit policies

- A contextual bandit policy at round t
  - gives a conditional distribution over the action A_t to be taken
  - conditioned on the history $\mathcal{D}_t$ and the **current context** X_t.
- In this module, we consider policies parameterized by θ: $\pi_\theta(a \mid x)$, for $\theta \in \mathbb{R}^d$.
- We denote the θ used at round t by θ_t, which will depend on $\mathcal{D}_t$.
- At round t, action $A_t \in \mathcal{A} = \{1, \ldots, k\}$ is chosen according to

$$\mathbb{P}(A_t = a \mid X_t = x, \mathcal{D}_t) = \pi_{\theta_t}(a \mid x).$$

Example: multinomial logistic regression policy

- Note: None of the discussion below depends on a specific policy class.
- However, it's helpful to have a policy class in mind.
- Let

$$\pi_\theta(a \mid x) = \frac{\exp\left(\theta^T \phi(x, a)\right)}{\sum_{a'=1}^{k} \exp\left(\theta^T \phi(x, a')\right)},$$

  where $\phi(x, a) : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$ is a joint feature vector.
- And $\theta^T \phi(x, a)$ can be replaced by a more general $g_\theta : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$,
  - e.g. a neural network.
- The whole conditional distribution $\pi_\theta(a \mid x)$ can also be represented as a neural network with a softmax output.
- The differentiability w.r.t. θ is key to a policy gradient method.

SGD for CPMs vs policy gradient

Conditional Probability Modeling (CPM)

- Input space $\mathcal{X}$
- Label space $\mathcal{Y}$
- Hypothesis space of functions $x \mapsto p_\theta(y \mid x)$
- Parameterized by $\theta \in \Theta$
- For any θ and x, $p_\theta(y \mid x)$ is a distribution on $\mathcal{Y}$.
- Mathematically, no different from a policy.

Conditional Probability Modeling (CPM)

- Given training set $\mathcal{D} = ((X_1, Y_1), \ldots, (X_n, Y_n)$ iid from $P_{\mathcal{X} \times \mathcal{Y}}$.
- Maximum likelihood estimation for dataset:

$$\begin{aligned} \theta &\in \arg\max_{\theta \in \Theta} \prod_{i=1}^{n} p_\theta(Y_i \mid X_i) \\ \iff \theta &\in \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \log\left[p_\theta(Y_i \mid X_i)\right] \end{aligned}$$

SGD for MLE of CPM

- Consider SGD to compute the MLE of a CPM.
- For observation (X_i, Y_i), we'll update θ by

$$\theta \leftarrow \theta + \eta \nabla_\theta \log p_\theta(Y_i \mid X_i)$$

  for some learning rate $\eta > 0$.
- This updates θ so there's more probability mass on **correct output** Y_i for input X_i.

The policy gradient update

- Below we'll derive the following policy gradient update to θ:

$$\theta \leftarrow \theta + \eta R_i(A_i) \nabla_\theta \log \pi_\theta(A_i \mid X_i)$$

- Compare this to the SGD update for CPM:

$$\theta \leftarrow \theta + \eta \nabla_\theta \log p_\theta(Y_i \mid X_i)$$

- Note that if $R_i(A_i) \equiv 1$, the two are equivalent.

Policy gradient vs conditional probability modeling

- In maximum likelihood with CPM, we're making the correct label Y_i more likely.
- With policy gradient, we're always increasing the probability of selected actions (with positive rewards), but the increase is larger with big positive rewards than with small positive rewards.

Policy gradient for contextual bandits

How to update the policy?

- Let A be an action chosen according to $\pi(a;\theta)$.
- Let $(X,R) \in \mathcal{X} \times \mathbb{R}^k \sim P$ be a generic context/reward vector pair.
- We want to find θ to maximize

$$\begin{aligned} J(\theta) &:= \mathbb{E}_\theta [R(A)] \\ &= \mathbb{E}_X \left[\mathbb{E}_{A|X\sim\theta} \left[\mathbb{E}_{R|X} [R(A) \mid A, X] \mid X \right] \right] \\ &= \mathbb{E}_X \left[\sum_{a=1}^{k} \pi_\theta (a \mid X) \, \mathbb{E}_{R|X} [R(A) \mid A = a, X] \right] \end{aligned}$$

- And now we differentiate w.r.t. θ.... but first...

Clever Trick

- But first a clever trick:

$$\nabla_\theta \log \pi_\theta(a \mid x) \quad = \quad \frac{\nabla_\theta \pi_\theta(a \mid x)}{\pi_\theta(a \mid x)}$$

- Rearranging, we get

$$\nabla_\theta \pi_\theta(a \mid x) \quad = \quad \pi_\theta(a \mid x) \nabla_\theta \log \pi_\theta(a \mid x).$$

- This assumed that $\pi_\theta(a \mid x) > 0$.

Gradient of Objective Function

- For a given θ, we want to find direction to increase $J(\theta)$:

$$
\begin{aligned}
\nabla_\theta J(\theta) &= \nabla_\theta \mathbb{E}_X \left[\sum_{a=1}^{k} \pi_\theta(a \mid X)\, \mathbb{E}_{R|X}[R(A) \mid A = a, X] \right] \\
&= \mathbb{E}_X \left[\sum_{a=1}^{k} \nabla_\theta [\pi_\theta(a \mid X)]\, \mathbb{E}_{R|X}[R(A) \mid A = a, X] \right] \\
&= \mathbb{E}_X \left[\sum_{a=1}^{k} \pi_\theta(a \mid X)\, \nabla_\theta \log \pi_\theta(a \mid X) \mathbb{E}_{R|X}[R(A) \mid A = a, X] \right] \text{ (clever trick)} \\
&= \mathbb{E}_X \left[\mathbb{E}_{A|X\sim\theta} \left[\nabla_\theta \log \pi_\theta(A \mid X) \mathbb{E}_{R|X}[R(A) \mid A, X] \mid X \right] \right] \text{ (payoff of clever trick)} \\
&= \mathbb{E}_X \left[\mathbb{E}_{A|X\sim\theta} \left[\mathbb{E}_{R|X} [\nabla_\theta \log \pi_\theta(A \mid X) R(A) \mid A, X] \mid X \right] \right] \\
&= \mathbb{E}_\theta [R(A) \nabla_\theta \log \pi_\theta(A \mid X)]
\end{aligned}
$$

- In the setting of reinforcement learning, this result is often referred to as the Policy Gradient Theorem.

Unbiased estimate for the gradient

- Consider round t of SGD for optimizing $J(\theta)$.
- We play A_t from $\pi_{\theta_t}(a \mid X_t)$ and record $(X_t, A_t, R_t(A_t))$.
- To update θ_t, we need an unbiased estimate of

$$\nabla_\theta J(\theta_t) = \mathbb{E}_{\theta_t}\left[R_t(A_t)\nabla_\theta \log \pi_{\theta_t}(A_t \mid X_t)\right].$$

- Trivially,

$$R_t(A_t)\nabla_\theta \log \pi_{\theta_t}(A_t \mid X_t)$$

  is an unbiased estimate of $\nabla_\theta J(\theta_t)$.

- Suppose we ran multiple rounds with the same policy θ. We can also get a gradient estimate (a better one) by averaging all those results together. For convenience, we'll just index them by $1, \ldots, N$. So the gradient estimate would be

$$\theta \leftarrow \theta + \eta \left[\frac{1}{N} \sum_{i=1}^{N} R_i(A_i) \nabla_\theta \log \pi_\theta(A_i \mid X_i) \right].$$

- If each of those rounds had a different policy θ_i, then we could use importance sampling to get an unbiased estimate:

$$\theta \leftarrow \theta + \eta \left[\frac{1}{N} \sum_{i=1}^{N} \frac{\pi_{\theta_i}(A_i \mid X_i)}{\pi_\theta(A_i \mid X_i)} R_i(A_i) \nabla_\theta \log \pi_\theta(A_i \mid X_i) \right].$$

Basic policy gradient for contextual bandits

Policy gradient algorithm (step size $\eta > 0$):

1. Initialize $\theta_1 = 0 \in \mathbb{R}^k$.
2. For each round $t = 1, \ldots, T$:
   1. Observe context X_t.
   2. Choose action A_t from distribution $\mathbb{P}(A_t = a \mid X_t) = \pi_{\theta_t}(a \mid X_t)$.
   3. Receive reward $R_t(A_t)$.
   4. $\theta_{t+1} \leftarrow \theta_t + \eta R_t(A_t) \nabla_\theta \log \pi_{\theta_t}(A_t \mid X_t)$.