Reinforcement Learning in Recommendation

Off-policy Policy Evaluation

Ondřej Podsztavek

Faculty of Information Technology
Czech Technical University in Prague

March 16, 2018

Outline

Motivation

- Recommendation algorithms are typically optimized for click through rate:

$$CTR \stackrel{\text{def}}{=} \frac{\text{total number or clicks}}{\text{total number of } \textbf{visits}} \times 100.$$

- But there is interest in optimizing customer lifetime value:

$$LTV \stackrel{\text{def}}{=} \frac{\text{total number or clicks}}{\text{total number of } \textbf{visitors}} \times 100.$$

- Reinforcement learning can possibly address this problem.

Goals

1. Design off-line performance metric.
2. Implement simulator to measure performance.
3. Measure performance of context-free bandit as simple baseline.
4. Move on to contextual bandits and full reinforcement learning algorithms.

Challenges

How to **compute good lifetime value strategy** and **evaluate** it off-line (**off-policy**) from logged data collected by different (**behavior**) policy.

Contextual Bandits

- Select articles to serve users based on contextual information about users and articles.
- Simultaneously adapting its selection strategy according to user-click feedback.
- Maximize CTR.

Feature-based Exploration and Exploitation Problem

- Large number users and content represented by features.
- Critical to **generalize** users and content.
- Balance user satisfaction in long run (*exploitation*) and gathering information about goodness of content (*exploration*).

Definition

Contextual bandit algorithm **A** proceeds in time steps $t = 1, 2, 3, \ldots$ and at each t:

1. **A** observes a user u_t and a set $\mathcal{A}_t$ of actions characterized by *context* vector $\mathbf{x}_{t,a}$ summarizing both the user u_t and the action a.
2. **A** chooses and action a_t and receives reward r_{t,a_t}.
3. **A** improves its selection strategy with the new observation $(\mathbf{x}_{t,a}, a_t, r_{t,a_t})$

Total T-trial Return

Total T-trial return is defined as:

$$G(T) \stackrel{\text{def}}{=} \sum_{t=1}^{T} r_{t,a_t}$$

and *optimal expected T-trial return*:

$$G^*(T) \stackrel{\text{def}}{=} \mathbf{E}\left[\sum_{t=1}^{T} r_{t,a_t^*}\right].$$

In context of article recommendation:

- An article is an action.
- If article is clicked on the reward is 1 else 0 then expected return is CTR.

LinUCB[1], Linear Upper Confidence Bound

Estimate mean reward of $\hat{\mu}_{t,a}$ and confidence interval $c_{t,a}$, such that with high probability:

$$|\hat{\mu}_{t,a} - \mu_a| < c_{t,a}, \quad a_t = \arg\max_a(\hat{\mu}_{t,a} + c_{t,a}).$$

LinUCB with *disjoint* linear models:

$$\mathbf{E}[r_{t,a}|\mathbf{x}_{t,a}] = \mathbf{x}_{t,a}^\top \theta_a^*.$$

LinUCB with *hybrid* linear models:

$$\mathbf{E}[r_{t,a}|\mathbf{x}_{t,a}] = \mathbf{z}_{t,a}^\top \beta^* + \mathbf{x}_{t,a}^\top \theta_a^*.$$

Both learned with *ridge regression*.

[1] Lihong Li et al. "A Contextual-bandit Approach to Personalized News Article Recommendation". In: *Proceedings of the 19th International Conference on World Wide Web*. 2010.

Approaches Off-line Evaluation

1. **On-line** evaluation *expensive* and *not reproducible*.
2. **Simulator** is challenging to implement moreover might be biased.
3. **Off-line** data could provide unbiased estimate but they are *partially-labeled* (only one action has reward feedback).

Off-line Evaluation of Contextual Bandits[2]

$h_0 \leftarrow \emptyset,\ \hat{G}_{\mathbf{A}} \leftarrow 0,\ T \leftarrow 0$
for $t = 1, 2, 3, \ldots, L$ **do**
  get the t-th event $(\mathbf{x}, a, r_a)$ from S {stream S of length L}
  if $\mathbf{A}(h_{t-1}, \mathbf{x}) = a$ **then**
    $h_t \leftarrow \text{concatenate}(h_{t-1}, (\mathbf{x}, a, r_a))$
    $\hat{G}_{\mathbf{A}} \leftarrow \hat{G}_{\mathbf{A}} + r_a,\ T \leftarrow T + 1$
  else
    $h_t \leftarrow h_{t-1}$
  end if
end for
return $\hat{G}_{\mathbf{A}} / T$

Assumptions

Stable arms set. Logging policy chooses arms uniformly at random. Data, events are IID.

[2] L. Li et al. "Unbiased Offline Evaluation of Contextual-bandit-based News Article Recommendation Algorithms". In: *ArXiv e-prints* (Mar. 2010). arXiv: 1003.5956 [cs.LG].

Direct Method

Estimate the value of policy π (policy evaluation):

$$V^{\pi} = \mathbf{E}_{(x,\mathbf{r})\sim D}[r_{\pi(x)}|x].$$

Policy optimization is to find an optimal policy with maximum value:

$$\pi^* = \arg\max_{\pi} V^{\pi}.$$

Form an estimate of $\hat{\rho}_a(x) = \mathbf{E}_{(x,\mathbf{r})\sim D}[r_a|x]$ of expected reward considering a context and an action:

$$\hat{V}^{\pi}_{\mathrm{DM}} = \frac{1}{|S|}\sum_{x\in S}\hat{\rho}_{\pi(x)}(x).$$

Estimate $\hat{\rho}$ might be biased (is based on different policy).

Inverse Propensity Score

Approximate behavior policy $\hat{p}(a|x)$ of $p(a|x)$ and correct the proportion between target and behavior policy:

$$\hat{V}^{\pi}_{\mathrm{IPS}} = \frac{1}{|S|} \sum_{(x,a,r_a)\in S} \frac{r_a \mathbf{I}(\pi(x) = a)}{\hat{p}(a|x)}$$

In practice no problem with bias but high variance.

Doubly Robust Estimator[3]

Take advantage of both *direct model* and *inverse propensity score*:

$$\hat{V}^{\pi}_{\mathrm{DR}} = \frac{1}{|S|} \sum_{(x,a,r_a)\in S} \left[\frac{(r_a - \hat{\rho}_a(x))\mathbf{I}(\pi(x) = a)}{\hat{p}(a|x)} + \hat{\rho}_{\pi(x)}(x) \right].$$

Intuition is to use $\hat{\rho}$ as a baseline and if data available apply correction.

[3] Miroslav Dudík, John Langford, and Lihong Li. "Doubly Robust Policy Evaluation and Learning". In: *CoRR* abs/1103.4601 (2011). arXiv: 1103.4601. URL: http://arxiv.org/abs/1103.4601.

Full Reinforcement Learning

Reinforcement learning algorithms distinguish between a visit and a visitor. Moreover, they can *learn from delayed reward*.

Motivation

"I expected to find something in recommendation systems, but I believe those are still dominated by collaborative filtering and contextual bandits. (...) Every Internet company ever has probably thought about adding RL to their ad-serving model, but if anyones done it, they've kept quiet about it."[4]

Advantages over contextual bandits

Sufficient if users establish long-term relationships by returning back (do not expect i.i.d. visits).

---

[4] Alex Irpan. *Deep Reinforcement Learning Doesn't Work Yet*. 2018. URL: https://www.alexirpan.com/2018/02/14/rl-hard.html.

Markov Decision Process

Definition

MDP is a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$, where $\mathcal{S}$ is a set of possible states, $\mathcal{A}$ is a finite set of actions, $\mathcal{P}(s, a, s')$ is a probability of transition to s' when action a is taken in state s, $\mathcal{R}(s, a) \in \mathbb{R}$ is a reward received when action a is taken in state s and $\gamma \in [0, 1]$ is a discount factor.

In recommendation context:

- $\mathcal{S}$ is set of feature vectors describing a user,
- $\mathcal{A}$ is set of articles to recommend,
- $\mathcal{P}$ described (**unknown**) dynamics of users and
- $\mathcal{R}(s, a)$ is 1 if a user click on the article a else 0.

Reinforcement Learning Objective

Goal

Find a decision rule called *policy* π that maximizes the expected performance $\mathbf{E}[R(\tau)|\pi]$.

- Policy $\pi(a|s)$ denotes the probability of taking action a in state s.
- Episode produces a *trajectory* $\tau = s_1, a_1, r_1, \ldots, s_T, a_T, r_T$.
- T is a time horizon.
- $R(\tau) = \sum_{t=1}^{T} \gamma^{t-1} r_t$ is the *return* of trajectory τ.

Off-line Evaluation in Full Reinforcement Learning

1. **Simulator-based**: Fit a MDP model from the data and evaluate the against model.
2. **Simulator-free**: Evaluate based on *importance sampling* which correct the mismatch between target and behavior policy.

Importance Sampling[5]

Estimate the expected value of a random variable x with distribution d from sample drawn from distribution d':

$$\mathbf{E}_d[x] = \int x d(x) dx = \int x \frac{d(x)}{d'(x)} d'(x) dx = \mathbf{E}_{d'}\left[x \frac{d(x)}{d'(x)}\right].$$

Unbiased and *consistent* estimate:

$$\frac{1}{n}\sum_{i=1}^{n} x_i \frac{d(x_i)}{d'(x_i)}.$$

[5] Doina Precup, Richard S. Sutton, and Satinder P. Singh. "Eligibility Traces for Off-Policy Policy Evaluation". In: *Proceedings of the Seventeenth International Conference on Machine Learning*. ICML '00. 2000.

Importance Sampling Estimator

Provide unbiased estimate of π_b value.
Define the per-step importance ratio:

$$\rho_t \stackrel{\text{def}}{=} \frac{\pi_t(a_t|s_t)}{\pi_b(a_t|s_t)}$$

and cumulative importance ratio:

$$\rho_{1:t} \stackrel{\text{def}}{=} \prod_{t'=1}^{t} \rho_{t'}.$$

Trajectory-wise *importance sampling* estimate:

$$V_{\text{IS}} \stackrel{\text{def}}{=} \sum_{t=1}^{H} \gamma^{t-1} \rho_{1:t} r_t.$$

High Confidence Off-policy Evaluation[7]

Model-free approach to off-policy evaluation. Compute lower bound on true performance $\mathbf{E}[R(\tau)|\pi]$ using *importance sampling*. Three approaches:

- concentration inequality[6],
- Student's *t*-test,
- bias corrected and accelerated bootstrap.

Suitable for safe policy improvement.

---

[6] Philip S. Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. "High Confidence Off-Policy Evaluation". In: *Proceedings of the Twenty-Ninth Conference on Artificial Intelligence*. 2015.

[7] Georgios Theocharous, Philip S. Thomas, and Mohammad Ghavamzadeh. "Ad Recommendation Systems for Life-Time Value Optimization". In: *Proceedings of the 24th International Conference on World Wide Web*. 2015.

Doubly Robust Off-policy Value Evaluation[9]

Doubly robust estimator for sequential decision-making. Unbiased and much lower variance than *importance sampling*.

MAGIC[8]

Better extension of doubly robust estimator.

---

[8] Philip S. Thomas and Emma Brunskill. "Data-Efficient Off-Policy Policy Evaluation for Reinforcement Learning". In: *CoRR* abs/1604.00923 (2016). arXiv: 1604.00923. URL: http://arxiv.org/abs/1604.00923.

[9] Nan Jiang and Lihong Li. "Doubly Robust Off-policy Evaluation for Reinforcement Learning". In: *CoRR* abs/1511.03722 (2015). arXiv: 1511.03722. URL: http://arxiv.org/abs/1511.03722.

Futuristic Vision

Model-based reinforcement learning which while recommending models users and based on its model plans what to recommend.