Doubly Robust Policy Evaluation and Learning

Miroslav Dudik, John Langford and Lihong Li

Yahoo! Research
Discussed by Miao Liu

October 9, 2011

1. Introduction

2. Problem Definition and Approaches
   - Two Common Solutions
   - Doubly Robust (DR) Estimator

3. Experiments
   - Multiclass Classification with Bandit Feedback
   - Estimating Average User Visits

- This paper is about decision making in environment where feedback is received only for chosen action (medical studies, content recommendation etc).
- These problems are instance of *contextual bandits*.

Definition (Contextual bandit problem)

In a contextual bandits problem, there is a distribution D over pairs $(x, \vec{r})$ where $x \in \mathcal{X}$ is the context (feature vector), $a \in \mathcal{A} \triangleq \{1,...,k\}$ is one of the k arms to be pulled, and $\vec{r} \in [0,1]^{|\mathcal{A}|}$ is a vector of rewards. The problem is a repeated game: on each round, a sample $(x, \vec{r})$ is drawn from D, the context x is announced, and then for precisely one arm a chosen by the player, its reward r_a is revealed.

Definition (Contextual bandit algorithm)

A contextual bandits algorithm $\mathcal{B}$ determines an arm $a \in \mathcal{A} \triangleq \{1,...,k\}$ to pull at each time step t, based on previous observation sequence $(x_1, a_1, r_{a,1}), ..., (x_{t-1}, a_{t-1}, r_{a,t-1})$, and the current context x_t.

Definition

- The input data generation
  - A new example $(x, \vec{r}) \sim D$.
  - Policy $a \sim p(a|x, h)$, where h is the history of previous observations, i.e. the concatenation of all preceding contexts, actions and observed rewards.
  - Reward r_a is revealed, whereas other rewards $r_{a'}$ with $a' \neq a$ are not observed
- Note the distribution D and policy p are unknown.
- Given a data set $S = (x, h, a, r_a)$, the tasks of interest are
  1. Policy evaluation
  $$V^\pi = \mathbb{E}_{(x,\vec{r})\sim D}[r_{\pi(x)}]$$
  2. Policy optimization
  $$\pi^* = argmax_\pi V^\pi$$
- Better evaluation generally leads to better optimization.
- The key challenge here: only partial information about the reward is available which prohibits direct policy evaluation on data S.

Two common solutions

I. Direct method (DM)

- Forms an estimate $\hat{\rho}_a(x)$ of the expected reward conditioned on the context and action.
- Estimates the policy value by

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

Notes:

- If $\hat{\rho}_a(x)$ is a good approximation of the true expected reward, defined as $\rho_a(x) = \mathbb{E}_{(x,\vec{r})\sim D}[r_a|x]$, then (1) is a close to V^{π}.
- If $\hat{\rho}_a(x)$ is unbiased, $\hat{V}^{\pi}_{DM}$ is an unbiased estimate of V^{π}.
- A problem with DM: without knowing π, the estimate $\hat{\rho}$ is biased.

Two common solutions

II. Inverse propensity score (IPS)

- Forms an approximation $\hat{p}(a|x,h)$ of $p(a|x,h)$.
- Corrects for the shift in action proportions between the old, data collection policy and the new policy:

$$\hat{V}^{\pi}_{IPS} = \frac{1}{|S|} \sum_{(x,h,a,r_a) \in S} \frac{r_a \mathbb{I}(\pi(x) = a)}{\hat{p}(a|x,h)} \tag{2}$$

  where $\mathbb{I}(\cdot)$ is an indicator function.

Notes:

- If $\hat{p}(a \mid x,h) \approx p(a \mid x,h)$, then (2) will approximately be an unbiased estimate of V^{π} (from importance sampling theory).
- Since data collection policy $p(a \mid x,h)$ is typically well understood, it is often easier to obtain a good estimate $\hat{p}$.
- However, IPS typically has a much larger variance, due to the range of random variable increasing (Especially when $p(a|x,h)$ gets smaller).

Doubly Robust (DR) Estimator

- DR estimators take advantage of both the estimate of expected reward $\hat{\rho}_a(x)$ and estimate of action probabilities $\hat{p}(a|x,h)$.
- DR estimators was first suggested by Cassel et al. (1976) for regression, but not studied for policy learning:

$$\hat{V}^{\pi}_{DR} = \frac{1}{|S|} \sum_{(x,h,a,r_a)\in S} \left[\frac{(r_a - \hat{\rho}_a(x))\mathbb{I}(\pi(x)=a)}{\hat{p}(a|x,h)} + \hat{\rho}_{\pi(x)}(x) \right] \quad (3)$$

- Informally, (3) uses $\hat{\rho}$ as a baseline and if there is data available, a correction is applied.
- The estimator (3) is accurate if *at least* one the estimators, $\hat{\rho}$ and $\hat{p}$, is accurate.

A basic question

How does (3) perform as the estimates $\hat{\rho}$ and $\hat{p}$ deviate from the truth?

Bias Analysis

- Denote Δ the additive deviation of $\hat{\rho}$ from ρ and δ a multiplicative deviation of $\hat{p}$ from p:

$$\begin{aligned}\Delta(a,x) &= \hat{\rho}_a(x) - \rho_a(x),\\ \delta(a,x,h) &= 1 - p(a|x,h)/\hat{p}(a|x,h)\end{aligned} \tag{4}$$

- To remove clutter, use shorthands ρ_a for $\rho_a(x)$, $\mathbb{I}$ for $\mathbb{I}(\pi(x)=a)$, p for $p(\pi(x) \mid x,h)$, $\hat{p}$ for $\hat{p}(\pi(x) \mid x,h)$, and δ for $\delta(\pi(x),x,h)$.
- To evaluate $\mathbb{E}[\hat{V}^{\pi}_{DR}]$, it suffices to focus on a single term in Eq.(3), conditioning on h;

$$\begin{aligned}&\mathbb{E}_{(x,\vec{r})\sim D, a\sim p(\bullet|x,h)}\left[\frac{(r_a-\hat{\rho}_a)\mathbb{I}}{\hat{p}}+\hat{\rho}_{\pi(x)}\right]\\ &= \mathbb{E}_{x,\vec{r},a|h}\left[\frac{(r_a-\rho_a-\Delta)\mathbb{I}}{\hat{p}}+\rho_{\pi(x)}+\Delta\right]\\ &= \mathbb{E}_{x,a|h}\left[\frac{(\rho_a-\rho_a)\mathbb{I}}{\hat{p}}+\Delta(1-\mathbb{I}/\hat{p})\right]+\mathbb{E}_x[\rho_{\pi}(x)]\\ &= \mathbb{E}_{x|h}[\Delta(1-p/\hat{p})]+V^{\pi}=\mathbb{E}_{x|h}[\Delta\delta]+V^{\pi}.\end{aligned} \tag{5}$$

Theorem

Let Δ and δ be defined as above. Then, the bias of the doubly robust estimator is

$$|\mathbb{E}_S[\hat{V}^\pi_{DR}] - V^\pi| = \frac{1}{|S|}\left|\mathbb{E}_S\left[\sum_{(x,h)\in S} \Delta\delta\right]\right|$$

If the past policy and the past policy estimate are stationary (i.e., independent of h), the expression simplifies to

$$\mathbb{E}_S[\hat{V}^\pi_{DR}] - V^\pi = |\mathbb{E}_x[\Delta\delta]|.$$

- In contrast (assume stationarity):

$$\begin{aligned} |\mathbb{E}_S[\hat{V}^\pi_{DM}] - V^\pi| &= |\mathbb{E}_x[\Delta]| \\ |\mathbb{E}_S[\hat{V}^\pi_{IPS}] - V^\pi| &= |\mathbb{E}_x[\rho_{\pi(x)}\delta]|, \end{aligned} \tag{6}$$

  IPS is a special case of DR for $\hat{\rho}_a(x) \equiv 0$
- If either $\Delta \approx 0$, or $\delta \approx 0$, $\hat{V}^\pi_{DR} \approx V^\pi$, while DM requires $\Delta \approx 0$ and IPS requires $\delta \approx 0$.

Variance Analysis

- Define $\epsilon = (r_a - \rho_a)\mathbb{I}/\hat{p}$ and note $\mathbb{E}[\epsilon|x,a] = 0$, we have

$$
\begin{aligned}
&\mathbb{E}_{x,\vec{r},a}\left[\left(\frac{(r_a - \hat{\rho}_a)\mathbb{I}}{p} + \hat{\rho}_{\pi(x)}\right)\right] \\
=&\mathbb{E}_{x,\vec{r},a}[\epsilon^2] + \mathbb{E}_x[\rho^2_{\pi(x)}] + 2\mathbb{E}_{x,a}[\rho_{\pi(x)}\Delta(1 - \mathbb{I}/\hat{p})] \\
&+ \mathbb{E}_{x,a}[\Delta^2(1 - \mathbb{I}/\hat{p})^2] \\
=&\mathbb{E}_{x,\vec{r},a}[\epsilon^2] + \mathbb{E}_x[\rho^2_{\pi(x)}] + 2\mathbb{E}_x[\rho_{\pi(x)}\Delta\delta] + \mathbb{E}_x[\Delta^2(1 - 2p/\hat{p} + p/\hat{p}^2)] \\
=&\mathbb{E}_{x,\vec{r},a}[\epsilon^2] + \mathbb{E}_x[\rho^2_{\pi(x)}] + 2\mathbb{E}_x[\rho_{\pi(x)}\Delta\delta] \\
&+ \mathbb{E}_x[\Delta^2(1 - 2p/\hat{p} + p^2/\hat{p}^2 + p(1-p)/\hat{p}^2)] \\
=&\mathbb{E}_{x,\vec{r},a}[\epsilon^2] + \mathbb{E}_x[(\rho_{\pi(x)+\Delta\delta})^2] + \mathbb{E}_x[\Delta^2 p(1-p)/\hat{p}^2] \\
=&\mathbb{E}_{x,\vec{r},a}[\epsilon^2] + \mathbb{E}_x[(\rho_{\pi(x)+\Delta\delta})^2] + \mathbb{E}_x[\frac{1-p}{p}\Delta^2(1-\delta)^2]
\end{aligned}
$$

- Summing across all terms in Eq.3 and combining with Theorem3, the variance is obtained

Theorem

Let Δ, δ and ϵ be defined as above. If the past policy and the policy estimate are stationary, then the variance of the doubly robust estimator is

$$\mathbf{Var}[\hat{V}^{\pi}_{DR}] = \frac{1}{|S|}\left(\mathbb{E}_{x,\vec{r},a}[\epsilon^2] + \mathbf{Var}_x[\rho_{\pi(x)}(x) + \Delta\delta] + \mathbb{E}_x\left[\frac{1-p}{p}\cdot\Delta^2(1-\delta)^2\right]\right).$$

- $\mathbb{E}_{x,\vec{r},a}[\epsilon^2]$ accounts for the randomness in rewards.
- $\mathbf{Var}_x[\rho_{\pi}(x) + \Delta\delta]$ is the variance of the estimator due to the randomness in x.
- The last term can be viewed as the importance weighting penalty.
- Similar expression can be derived for the IPS estimator

$$\mathbf{Var}[\hat{V}^{\pi}_{IPS}] = \frac{1}{|S|}\left(\mathbb{E}_{x,\vec{r},a}[\epsilon^2] + \mathbf{Var}_x[\rho_{\pi(x)}(x) - \rho_{\pi(x)}\delta] + \mathbb{E}_x\left[\frac{1-p}{p}\cdot\rho_{\pi}(x)^2(1-\delta)^2\right]\right).$$

- The variance of the estimator for the direct method

$$\mathbf{Var}[\hat{V}^{\pi}_{DM}] = \frac{1}{|S|}\mathbf{Var}_x[\rho_{\pi(x)}(x) + \Delta]$$

which does not have terms depending either on the past policy or the randomness in the rewards.

Data Setup

- Assume data are drawn IID from a fixed distribution: $(x, c) \sim D$, where $x \in \mathcal{X}$ is the feature vector and $c \in \{1, 2, \cdots, k\}$ is the class label.
- A typical goal is to find a classifer $\pi : \mathcal{X} \to \{1, 2, \cdots, k\}$ minimizing the classification error:

$$e(\pi) = \mathbb{E}_{(x,c)\sim D}[\mathbb{I}(\pi(x) \neq c)]$$

- A classifier π may be interpreted as an action-selection policy, i.e. the data point (x, c) can be turned into a cost-sensitive classification example $(x, l_1, \cdots, l_k)$, where $l_a = \mathbb{I}(a \neq c)$ is the loss for predicting a.
- A partially labeled data is constructed by randomly selecting a label $a \sim \text{UNIF}(1, \cdots, k)$ and revealing the component l_a.
- The final data is in the form of (x, a, l_a).
- $p(a \mid x) = 1/k$ is assumed to be known.

Table 1. Characteristics of benchmark datasets used in Section 5.1.

Dataset	ecoli	glass	letter	optdigits	page-blocks	pendigits	satimage	vehicle	yeast
Classes (k)	8	6	26	10	5	10	6	4	10
Dataset size	336	214	20000	5620	5473	10992	6435	846	1484

Policy Evaluation

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

- For each data set:
  1. Randomly split data into training and test sets of roughly the same size;
  2. Obtain a classifier π by running a direct loss minimization (DLM) algorithm of McAllester et al. (2011) on the training set with fully revealed losses;
  3. Compute the classification error on fully observed test data (ground truth);
  4. Obtain a partially labeled test set.
- Both DM and DR require estimating the expected loss denoted as $l(x, a)$ for given (x, a). Use a liner loss model: $\hat{l}(x, a) = w_a x$, parameterized by k weighted vectors $\{w_a\}_{a\in\{1,\cdots,k\}}$, and use least-squares ridge regression to fit w_a.
- Both IPS and DR are unbiased since $p(a|h) = 1/k$ is accurate.

Policy Optimization

$$\pi^* = argmax_\pi V^\pi$$

- Repeat the following steps 30 times:
  1. Randomly split data into tranining (70%) and test (30%) sets;
  2. Obtain a partially labeled training data set;
  3. Use IPS and DR estimators to impute unrevealed losses in the training data;
  4. Two cost-sensitive multiclass classification algorithms are used to learn a classifier from the losses completed by either IPS or DR: the first is DLM (McAllester et al., 2011), the other is the Filter Tree reduction of Beygelzimer et al. (2008).
  5. Evaluate the learned classifiers on the test data.

Performance Comparisons

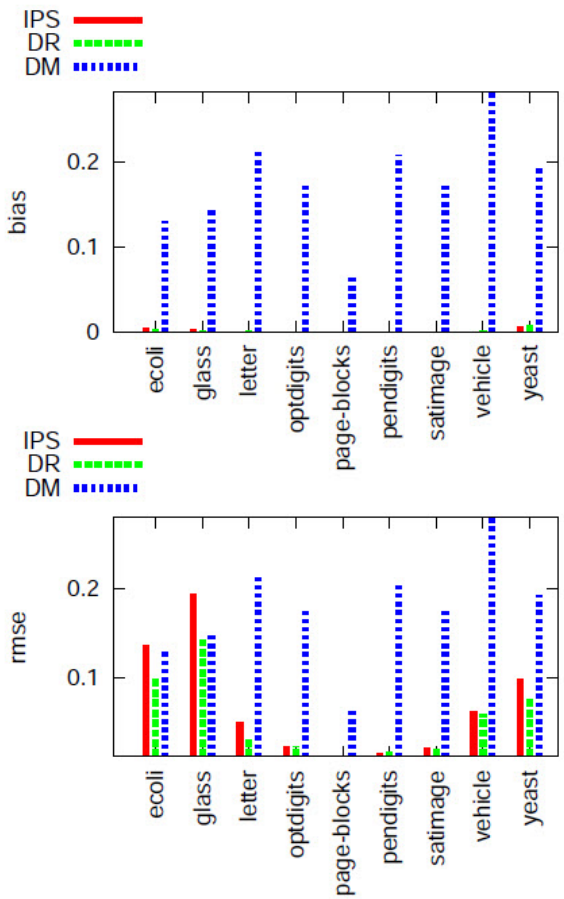

Figure 1. Bias (upper) and rmse (lower) of the three estimators for classification error.

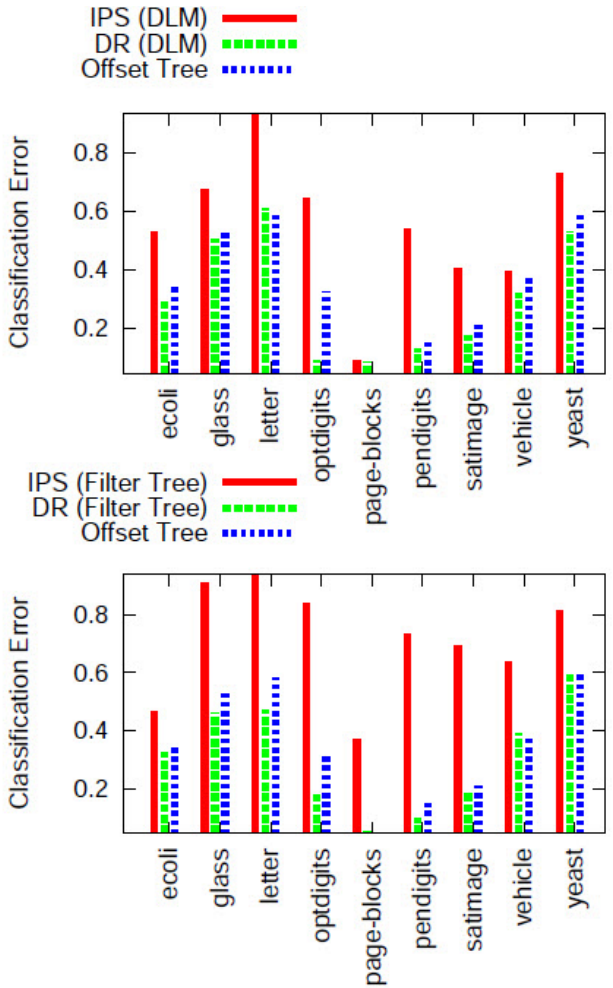

Figure 2. Classification error of direct loss minimization (upper) and filter tree (lower). Note that the representations used by DLM and the trees differ radically, conflating any comparison between the approaches. However, the Offset and Filter Tree approaches share a similar representation, so differences in performance are purely a matter of superior optimization.

Estimating Average User Visits

- The dataset: $D = \{(b_i, x_i, v_i)\}_{i=1,\dots,N}$, where $N = 3854689$, b_i is the ith(unique) bcookie[1] , $x_i \in \{0,1\}^{5000}$ is the corresponding (sparse) binary feature vector, and v_i is the number of visits.
- A sample mean might be a biased estimator if the uniform sampling scheme is not ensured. This is known as "covariate shift" [2], a special case of contextual bandit problem with $k = 2$ arms.
- The partially labeled data consists of tuples (x_i, a_i, r_i), where $a_i \in \{0,1\}$ indicates whether bcookie b_i is sampled, $r_i = p_i v_i$ is the observed number of visits, and p_i is the probability that $a_i = 1$.
- DR estimator requires building a reward model $\hat{\rho}(x) = \omega x$
- The sampling proabilities p_i of the b_i is defined as (Gretton et al. 208)$p_i = \min\{\mathcal{N}(\langle x_i, \bar{x}\rangle), 1\}$, where $\bar{x}$ is the first principle component of all features $\{x_i\}$.

[1] A bcookie is unique string that identifies a user.

[2] `http://sifaka.cs.uiuc.edu/jiang4/domain_adaptation/survey/node8.html`.

Performance Comparisons

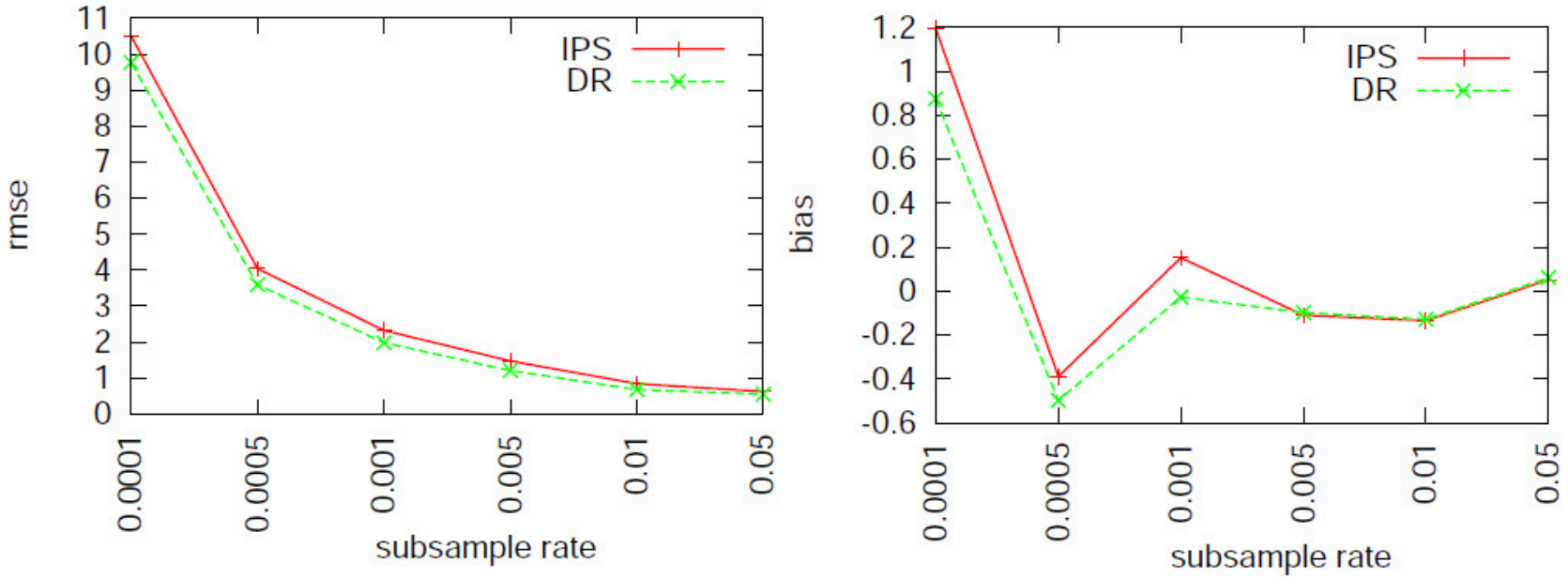

Figure: 3. Comparison of IPS and DR: rmse(left), bias(right). The ground truth value is 23.8.