STA571: Advanced Machine Learning

Application of Doubly Robust Method to Marketing Data (March 2017)

Author: Susan Cherry

1 Motivation

We are often interested in policy evaluation and optimization, that is estimating the total reward of a given policy and maximizing the total expected rewards. These sorts of problems can be thought of as contextual bandits, in which a decision maker observes some "contextual" information from which he or she can choose an action. A "reward" is given based on the chosen action. The goal of the decision maker is to maximize total expected rewards. Examples of situations that can be modeled by contextual bandits include the effectiveness of internet advertising (we only observe whether a user clicks on a certain advertisement, not whether that ad was presented to the user or not) and health policies (we only observe the outcome for a patient given a certain treatment, not alternative treatments).

In this project, I will deal only with offline data, which means that I assume access to historical data but do not have the ability acquire new data. Using offline data is necessary, as it is often more feasible to use existing data to evaluate policies than to run the policies online. Previous approaches to dealing with offline data include the Direct Method and Inverse Propensity Scores, but previous work shows that these approaches suffer from large bias and variance. This project will evaluate the use of the Doubly Robust technique, which has been shown to yield accurate estimates when either a good model of rewards or good model of past policy is present. While we have not discussed the Doubly Robust technique in class, this work is somewhat related to the Linear-Time Estimators for Propensity Scores paper that we will read later in the semester. It is more broadly related to the use of Machine Learning techniques in the estimation of causal treatment effects.

2 Problem definition

The goal of this project is to evaluate the effectiveness of the Doubly Robust estimator compared to the Direct Method and Inverse Propensity Scores on the effect of a direct marketing campaign conducted by a Portuguese banking institution (the data comes from the UCI Bank Marketing dataset). The methods used in this project are similar to the methods proposed in the "Doubly Robust Policy Evaluation and Learning Paper" published by Dudik, Langford, and Li. In their paper, however, their experiments are based off of real world data but the exploration and partial feedback are simulated. Using the Bank Marketing dataset, the exploration and partial feedback will also come from real world data.

The dataset contains 45,000 observations and has data on individuals' background characteristics,x, whether the individual was contacted by the marketing campaign, a (this is the policy I will evaluate), and whether the individual has subscribed a term deposit, r (outcome of the policy/reword). The goal of this project to to evaluate how well the Doubly Robust method performs on the Bank Marketing data. Previous research implies that the Direct Method will have a large bias and that the Inverse Propensity Scoring has large variance. My project will evaluate whether this is true and determine if the Doubly Robust method gives

better and more accurate results. if it does, then this provides evidence that Doubly Robust methods should be used more often when evaluating and optimizing policy.

3 Models and methods

I will compare Doubly Robust (DR) method, Direct Method (DM), and Inverse Propensity Scoring (IPS). Estimating policy value in contextual bandits is difficult because only the reward for the chosen action (whether the individual was contacted by the marketing campaign or not) is observed and not any of the other actions. DM and IPS are the two most common methods for these situations. DM forms an estimate of the expected reward, $\hat{r}(x,a)$ conditional on the context and action. The policy value is estimated by: $\hat{V}_{DM} = \frac{1}{n}\Sigma_{k=1}\Sigma_{a\in A}v)a|x_k)\hat{r}(x_k a)$. DM is particularly prone to bias. IPS is less prone to bias. It forms an approximation of $\hat{\mu}_k(a|x)$ and uses this estimate to correct the shift in action proportions between the exploration policy and the new policy: $\hat{V}_{IPS} = \frac{1}{n}\sum_{k=1}\frac{v(a|x_k)}{\hat{\mu}_k(a|x)}r_k$. IPS typically has a large variance.

DR (the new approach that my project will test) attempts to correct the problems of the old approaches. It takes advantage of both the estimate of the reward and the estimate ($\hat{r}$) of the action probabilities ($\hat{\mu}$). We have: $\hat{V}_{DR} = \frac{1}{n}\Sigma_k[)\hat{r}(x_k,v) + \frac{v(a|x_k)}{\hat{\mu}_k(a|x)}(r_k - \hat{r}(x_k,a_k))]$, where $\hat{r}(x_k,v) = \Sigma_{a\in A}v(a|x)\hat{r}(x,a)$. This estimator is unbiased if at least one of the estimators, $\hat{r}$ or $\hat{\mu}_k$ is accurate, which is why it is called "doubly robust".

I observe (x_k, a_k, r_k) from the dataset and estimate p_k, which is the recorded probability that the exploration policy chose to contact the individual or not. I estimate p_k by using $\hat{\mu_k}(a_k|x_k)$. Following Dudik, Erhan, Langford, and Li, I will estimate the value of two policies: the exploration policy and the argmax policy, which is based on a linear estimator $r'(x,a) = w_a x$. Using the exploration policy is useful, because it provides ground truth and a sanity check. I will fit $r'(x,a)$ on training data by importance weighted linear regression with importance weights $\frac{1}{p_k}$.

4 Results and validation

To evaluate my results, I will compare DR, IPS, and DM for both the exploration and argmax policies. I will report both the estimated reward and the variance. Theory suggests that the variance will be the largest for IPS and smallest for DR. Theory also suggests that the bias will be largest for DM and smallest for DR. The exploration policy provides ground truth, so I can evaluate the bias to see if the theory holds.

The results of this project will allow me to determine if the doubly robust policy gives more accurate estimates than the direct method and inverse propensity scoring. If so, my project will provide further evidence that the doubly robust method should become standard practice with contextual bandits and with policy evaluation and optimization.