Predicting the Treatment Status

Nikolay Doudchenko

1 Introduction

Many studies in social sciences deal with treatment effect models.[1] Usually there is a treatment variable which determines whether a particular individual (or more generally a unit of observation) was affected by some impact (the binary case) or the intensity of this impact. A researcher is interested in evaluating the effect of the treatment measured as some function of the joint distribution of $y_1^{(i)}$, the potential outcome for individual i conditional on being treated, and $y_0^{(i)}$, the potential outcome conditional on not being treated. There are many papers devoted to studying this setup due to a very fundamental problem that for each individual i either $y_1^{(i)}$ or $y_0^{(i)}$ is observed in the data, but never both.

In my project I consider a somewhat different problem. What if we don't know who has got the treatment and who has not? For example, suppose that we organize a lottery in which every participant draws a lottery ticket from a big jar. Some of the tickets are the winning ones (the individual is treated) while some aren't (the individual is not treated). We cannot observe whether a particular person has won, but we can observe some outcome potentially affected by the realization of the lottery. Can we infer anything about the treatment?[2] To be precise, I ask two questions: (i) Can we guess the treatment status? (ii) Is it possible to measure the treatment effect even though we don't observe the treatment dummy?

I approach this problem as an unsupervised learning problem. Although the treatment status is observed in the data that I use, I assume that in practice it won't be and there will be no way to train the model. The main tool in my analysis is what I call the "Hidden Variable Model" (HVM). Essentially, this is a hidden Markov model with the main difference being that there is no Markov structure and instead of observing a number of consecutive realizations for one individual we observe a number of "independent" individuals. The hidden state is exactly the treatment status which we don't observe. We may also know some additional details of the assignment process or how the treatment affects the outcome. For example, we may know that the treatment status was determined

[1] Another prominent example would be medical studies. It does not matter for the theoretical derivations. Neither does it for the simulations. The "real world data" that I use comes from economics so I stick to the social sciences context and terminology.

[2] Another interesting formulation is the following. What if there is an additional step required to complete the treatment? We know the individuals that were supposed to be treated, but we don't know who actually undertook that step. What can we do in this case?

based on a subset of features or that only some of the features affect the distribution of the outcome.

In this project I consider both simulated data and a "real" publicly available dataset that comes from a study by Angrist, Bettinger, Bloom, and King (2002).

The results suggest that as both the treatment and the outcome are stochastic the treatment status is hard to retrieve. Not surprisingly, the stronger is the treatment effect the easier it becomes to guess the treatment variable. Intuitively, if the treatment effect is positive and large, a large value of the outcome suggests that the individual was treated.

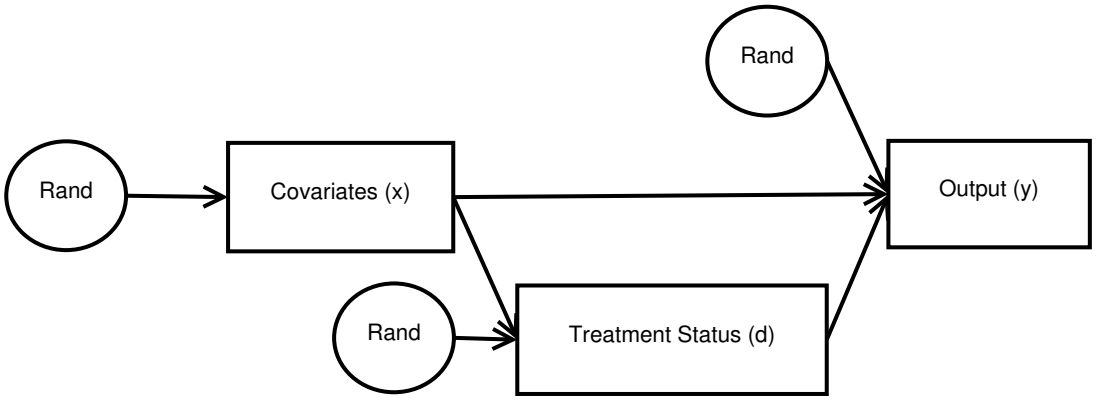
The estimates of the treatment effect are more precisely estimated. As obtaining these estimates is usually the main goal of the researcher, there is some hope that this project wasn't completely useless.

The rest of the text is organized as follows. Section 2 describes the statistical model referred to as the "Hidden Variable Model." Section 3 reports the results obtained using simulated data. Section 4 discusses the "real" dataset that I use in Section 5. Section 6 concludes.

2 Model

Suppose that for each observation $i = 1, \ldots, m$ there are n feature (or covariates) $x^{(i)} = (x_1^{(i)}, \ldots, x_n^{(i)})^T$ which determine the probability, $p(d = 1|x^{(i)})$, of being assigned to the treatment group. The treatment status, $d^{(i)}$, and the covariates jointly determine the distribution of the outcome variable, $p(y|d^{(i)}, x^{(i)})$. The graphical representation of the model is given in Figure 1.

Figure 1: Graphical Representation of the Model



Here the word "Rand" means that the covariates, the treatment status and the outcome are stochastic.

I assume that only $x^{(i)}$ and $y^{(i)}$ are observed and $d^{(i)}$ is hidden. The likelihood of $(y^{(i)}, x^{(i)})$ is given by

$$p(y^{(i)}, x^{(i)}) = \sum_{d=0,1} p(y^{(i)}, d|x^{(i)})p(x^{(i)}).$$

The log-likelihood is

$$\begin{aligned}\log p(y^{(i)}, x^{(i)}) &= \log\left(\sum_{d=0,1} Q_i(d)\frac{p(y^{(i)}, d|x^{(i)})}{Q_i(d)}\right) + \log p(x^{(i)}) \\ &\geq \sum_{d=0,1} Q_i(d)\log\frac{p(y^{(i)}, d|x^{(i)})}{Q_i(d)} + \log p(x^{(i)}) \\ &= \sum_{d=0,1} Q_i(d)\log\frac{p(y^{(i)}|d, x^{(i)})p(d|x^{(i)})}{Q_i(d)} + \log p(x^{(i)}),\end{aligned}$$

where $Q_i(d)$ is a $p.m.f.$ of some distribution over $d \in \{0, 1\}$. The inequality above follows from Jensen's inequality.

If we assume that $p(d|x^{(i)})$ and $p(y|d, x^{(i)})$ have particular parametric forms, these parameters can be estimated using the expectation-maximization (EM) algorithm.

There is one fundamental issue though. We can try to split the dataset into two parts, but it is impossible to say which of these two parts is the treatment group and which is the control group. In practice there is usually some additional knowledge that allows us to make an assumption about the sign of the treatment effect. Here I assume that the treatment effect is positive which allows us to predict that the group with the higher average $y^{(i)}$ is the treatment group.

2.1 Expectation-maximization

I assume that $p(d|x^{(i)}) = p_d$ and is known. In other words, the assignment to the treatment and control groups is completely random (does not depend on the covariates) and the researcher knows the fraction of people assigned to the treatment group.[3]

I also assume that y is binary (the case that corresponds to the real data that I use) and that

$$p(y|d, x^{(i)}) = \left(g(\theta_0 d + \theta^T x^{(i)})\right)^y \left(1 - g(\theta_0 d + \theta^T x^{(i)})\right)^{1-y},$$

where $g(s) = 1/(1 + \exp(-s))$ is the sigmoid function and θ_0, $\theta = (\theta_1, \ldots, \theta_n)^T$ are the parameters that we need to estimate.

During the E-step of the algorithm I compute $Q_i(d) \propto p(y^{(i)}, d|x^{(i)})$. During the M-step I use Newton's method to maximize the lower bound for the log-likelihood.[4]

3 Simulations

The goal of this section is to assess the performance of the model when we know exactly how the data is generated. Throughout this section $n = 10$, $x_1^{(i)} = 1$ for all $i = 1, \ldots, m$,

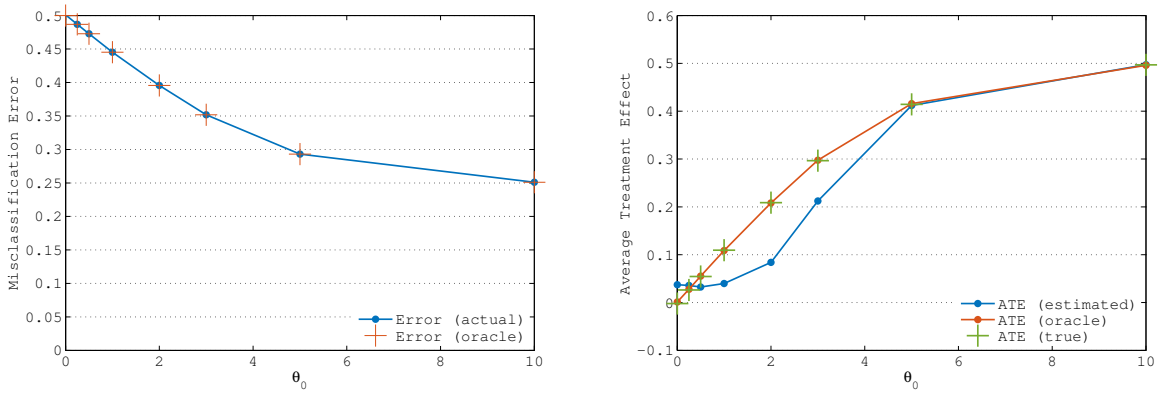
[3] It is easy to modify the M-step for the case when p_d is unknown.

[4] I tried to use gradient ascent too, but it leads to less accurate estimates of the average treatment effect. Moreover, the estimates become very sensitive to the initial values of θ_0, θ.

$x_j^{(i)} \sim \mathcal{N}(0,1)$ and independent across $i = 1, \ldots, m$ and $j = 2 \ldots, n$. I set $p_0 = p_1 = 1/2$, $\theta_1 = 0$, $\theta_j = 1$ for $j = 2, \ldots, 5$ and $\theta_j = -1$ for $j = 6, \ldots, 10$. I assume that $d^{(i)}$ and $y^{(i)}$ are generated according to the model described in Section 2.

The next two plots show how the misclassification error[5] and the estimate of the average treatment effect[6] behave when θ_0 changes.[7] These plots show that the algorithm fails

Figure 2: Misclassification Error and Average Treatment Effect vs. θ_0



to accurately predict the treatment status. Although it becomes easier as the treatment effect increases, even when $\theta_0 = 10$ on average 25% of the observations are misclassified. However, as both the treatment status and the outcome are stochastic, even the "oracle" model predicts poorly and produces essentially the same misclassification errors as the main model.

The estimates of the average treatment effect look much more promising. There are considerable discrepancies when θ_0 is small, but as it increases the estimate of the average treatment effect converge to that of an "oracle" model.

To economize the space I don't report the behavior of the misclassification error and the estimate of the average treatment effect as a function of the sample size.[8] The sample sizes above 1,000 seem to have essentially no effect on the performance.

4 Data

The dataset that I use comes from Angrist, Bettinger, Bloom, and King (2002).[9] I use the data from a randomized trial in Colombia conducted in the 90s. The idea of the

[5]This is an unsupervised learning problem so the term "misclassification error" might be confusing. However, the $d^{(i)}$ are actually observed so it is possible to calculate how many times the model has correctly guessed $d^{(i)}$.

[6]As in the statistical model that I consider the treatment status is independent of the potential outcome (the randomization assumption), the "true" average treatment effect is $E[y_1^{(i)} - y_0^{(i)}] = E[y^{(i)}|d^{(i)} = 1] - E[y^{(i)}|d^{(i)} = 0]$ and can be easily estimated.

[7]In the plots the word "oracle" refers to the case when we know the true parameters of the model, but do not observe the $d^{(i)}$'s. The word "true" refers to the case when $d^{(i)}$'s are observed.

[8]I presented these results at the poster session.

[9]This section is based heavily on the description of the experiment in the paper.

experiment was to randomly assign vouchers to the low-income families with children attending public primary schools. These vouchers could be used to pay the tuitions in private secondary schools.

The data consist of 1,315 individual level observations. The features include student's gender and age, the city where he or she lives, the year of the lottery (1993, 1995, or 1997), a dummy for whether the student's household has a phone, age and years of schooling of student's parents. There is also a treatment status variable which I want to predict, but I assume that it is not observed by the researcher. The outcome variable which the treatment status is supposed to affect is a dummy for whether the student had finished the 8th grade by the time of the follow up survey.

5 Results

The table below reports the performance of the model on the "real" data. I compare the HVM with the k-means model on the full set of features with $k = 2$. The "true" average treatment effect (computed using the actual values of $d^{(i)}$'s) is 0.068. Therefore, the numbers below demonstrate that both the HVM and the k-means perform poorly. As noted in the previous section, when the treatment effect is low the HVM tends to produce inaccurate estimates (which substantially depend on the initial values).

Model	Misclassification Error	Average Treatment Effect
k-means	0.499	0.353
Hidden variable	0.465	0.002

6 Conclusion

The results suggest that the Hidden Variable Model allows us to estimate the average treatment effect and tends to perform somewhat better than a simple k-means model. However, it fails to accurately predict the treatment status and the estimate of the average treatment effect depends on the initial values.

To improve the performance of the model I would like to add more (derived) features and consider alternative specifications of $p(y|d, x^{(i)})$.

References

Angrist, J., E. Bettinger, E. Bloom, and E. King (2002). Vouchers for private schooling in colombia: Evidence from a randomized natural experiment. *The American Economic Review*.

Kuroki, M. and J. Pearl (2014). Measurement bias and effect restoration in causal inference. *Biometrika 101*(2), 423–437.