A Simple Guide to Doubly Robust Estimation

Amit Sharma

Apr 18, 2016 · 5 min read

> Two roads diverged in a wood, and I—

> I took the one less traveled by,

> But if I could go back, I will try

> To take both: why one, then another.

When I first heard about doubly robust estimation, it sounded like magic, almost too good to be true. When working with messy data, we are used to making tradeoffs. From statistics, there is the bias-variance tradeoff: you can't improve one without impacting the other. From computer science, there are time and space complexity tradeoffs: an algorithm can take up less time or less space, but devising one that reduces both is always hard.

Thus, in a world with no free lunches, the doubly robust promise stood out. Given two possibly faulty ways of estimating a quantity, a doubly robust estimator guarantees an **unbiased estimate**, whenever one of them is correct. This can be a boon when working with data with unknown distributions and sampling strategies.

How does it work? The math is non-intuitive, so it is best to start with an example.

Finding the average height of students

Let us suppose we want to find the average height of people in a class. We could ask all the students to measure their height and then compute the mean of those observations.

$$\hat{H}_{mean} = \frac{\sum_{i=1}^{n} h_i}{n}$$

If all the students in the class were in attendance, the above formula will give us the correct answer. If not, and if some students were absent because of independent reasons, the mean estimate will still be an unbiased estimate of the average height in the class.

But what if some students were missing systematically? Say there was a basketball game and so many of the taller students were missing that day. Then the computed mean will be lower than the true average height of the class. One way to solve this problem is to know what kind of students are expected to be missing. Suppose we know that there are 5 students taller than 6” in the class, out of which 3 are missing. This means that any student taller than 6” has a 40% chance of being present in the class. To make matters simple, let us assume that all other students are present and therefore have a 100% chance of being in the class.

To account for absence of taller students, we could give more importance to observations from taller students.

$$\hat{H}_{IPS} = \frac{\sum_{i=1}^{n} h_i/p_i}{n}$$

where p_i is the probability that a student of height h_i is present in the class. In our example, p_i is 1 for everyone except tall students.

$$p_i = \begin{cases} 1 & \text{if } h_i < 6 \\ 0.4 & \text{if } h_i >= 6 \end{cases}$$

Another way of writing the same formula is by summing over all students enrolled in teh class, and using an indicator variable o_i to denote whether the student was present.

$$\hat{H}_{IPS} = \frac{1}{n}\sum_{i=1}^{N} \frac{o_i h_i}{p_i}$$

Using this formula, we can expect to account for the missing tall students and achieve an unbiased estimate for the average height.

Using a doubly robust estimator

The problem though is that in many cases, we may not know which one of the above estimators to use. It could be that we do not have enough information about the missing students, and thus cannot accurately write down the probability of omission, p_i accurately. Or that the entire assumption about the missing basketball team is incorrect, and students across the height spectrum were absent at random.

Fortunately, the doubly robust estimator allows us to estimate the average height even when we are not sure about which assumption is true. It is given by:

$$\hat{H}_{DR} = \frac{1}{n}\sum_{i=1}^{N} [\frac{o_i h_i}{p_i} - \frac{o_i - p_i}{p_i}\hat{h}_{i,mean}]$$

Resolving the mystery

Let's try to see why it works. First, let's rearrange some terms.

$$\hat{H}_{DR} = \frac{1}{n}\sum_{i=1}^{N} h_i + \frac{1}{n}\sum_{i=1}^{N} \frac{o_i - p_i}{p_i}(h_i - \hat{h}_{i,mean})$$

The DR estimator will be unbiased whenever the right term is zero.

- Let us suppose that the students are missing at random. Then, $\hat{H}_{mean} = \sum_{i=1}^{N} h_i$, so the right term is zero.
- Similarly, when we suspect that taller students are absent more than other students and we can estimate p_i accurately, then $\sum_{i=1}^{N} o_i - p_i = 0$ and again, we find that right term evaluates to zero.

Voila! In both cases, we obtain the correct average height using the DR estimator.

Generalizing to regression and propensity scores

The same principle can be generalized to more complex scenarios. $\hat{H}_{mean}$ can be generalized as a regression, based on some observed multi-dimensional data about students (X).

$$\hat{h}_{i,REG} = \alpha_0 + \sum_{j=1}^{M} \alpha_j x_j$$

Similarly, the calculation of propensity scores can be generalized, using observed data X.

$$p_i = P(o_i = 1|x_i) = Logit(\alpha_0 + \sum_{j=1}^{M} \alpha_j x_j)$$

Most generally, we require *some* way for estimating $\hat{h}_{i,REG}$ and p_i. Any functional form, such as a learned decision tree or a machine learning model, can be substituted in place of the regression or logit models.

Okay, where is the catch?

The catch, as you might imagine, is that in most practical cases, establishing even one of these assumptions is non-trivial. With messy data from the real-world, it is anybody's guess whether the data is missing at random, or what the correct probabilities of omission are.

Still, using a doubly robust estimator provides a useful check against modeling assumptions, as long as we do not err badly on both counts.

causal inference