Modeling Social Data - Lecture 11, April 14th

bk2628

April 20, 2017

1 Prediction vs Causation

1.1 Prediction

Making a forecast without changing anything. For example: seeing your neighbor with an umbrella might predict rain

1.2 Causation

Make a change in the current scenario, i.e., the current state of the world and anticipate what will happen.

1.2.1 Reverse causal inference

Finding the cause of something that has already happened. For example: what caused my kid to get sick? Reverse causal inference is generally quite hard.

1.2.2 Forward causal inference

Forward causal inference is more "what happens if we do a certain thing?".

1.2.3 Example : Hospitalization on health

What's the effect of going to the hospital today on your health tomorrow? If we try to estimate the above model based on just observational data we may be missing out an unobserved common cause like the health of the person on the day of the hospital visit.

Observational estimates

Let's say all sick people in our dataset went to the hospital today, and healthy people stayed home
Observed difference in health tomorrow = (Sick and went to hospital) – (Healthy and stayed home)
Observed difference in health tomorrow = [(Sick and went to hospital) – (Sick if stayed home)] + [(Sick if stayed home) - (Healthy and stayed home)]

Causal effect = (Sick and went to hospital) – (Sick if stayed home)

Selection bias = (Sick if stayed home) - (Healthy and stayed home)

Observed difference in health tomorrow = Casual effect - Selection bias

There is a selection bias in this example because the people visiting the hospital are not random as today's health of the person affects the decision of visiting the hospital today.

2 Predictive Systems

Aim : predict the future activity for a user.

2.1 Search engine example

Say you wanted to buy a toy from amazon and so, you search for it on Google. Google would show you an ad for Amazon.

But the search results also have Amazon. This raises the following question:

The counterfactual question is "would have i still landed on Amazon.com even without the ad?"

Maybe or maybe not.

There can be hidden causes for the observed results, and ignoring such causes can lead to completely different conclusions.

In the above example, people may have anyway used Amazon.com to order the toy even without the ad. But our observations may suggest that the activity was because of the ad.

2.2 Simpson's Paradox

Selection bias can be so large that observational and causal estimates give opposite effects.

(example: going to hospitals makes you less healthy)

Other examples : Comparing an old algorithm with a new one.

Old Algorithm	New Algorithm
5%(50/1000)	5.4%(54/1000)

After dividing the group into low and high activity users

Low activity users

Old Algorithm	New Algorithm
2.5%(10/400)	2%(4/200)

High activity users

Old Algorithm	New Algorithm
6.6%(40/600)	6.2%(50/800)

New algorithm has a better success rate overall but when we divide the data, old algorithm seems to have a higher success rate.

Conclusion: we can add as many variables as we want to the model and different algorithms would perform different under different variables.

Going back to the hospital example:

How can we check if going to the hospital makes you less healthy?

One way is to "clone" the world which means just change the one thing that we want to measure while keeping everything else the same.

Guy goes to the hospital Vs Guy doesn't go to the hospital everything else remains the same.

This works well in theory but how do we actually clone the world?

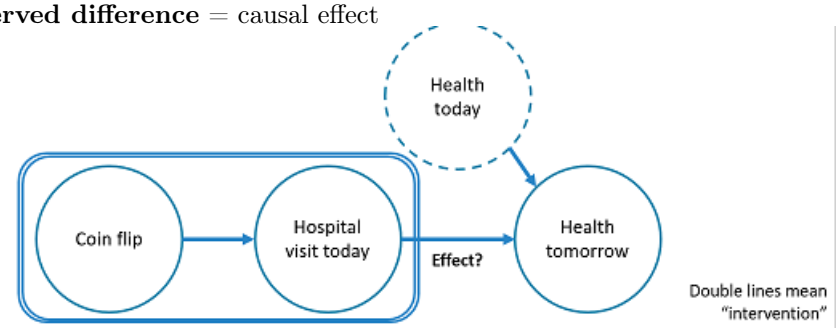
Take the data set(sick or healthy) and use a coin flip to decide who goes into which version of the world. Since the decision is based on a coin flip, the expected result would be to get similar samples in both the worlds.

Basic identity of causal inference:

observed difference = causal effect - selection bias

Selection bias is zero since there in no difference, on average, between the two worlds.

observed difference = causal effect



Random sampling makes sure that health today has no effect on the health tomorrow. The only thing we are checking is "does going to the hospital today affects you health tomorrow?".

2.2.1 Experiments : Caveats/Limitations

- Randomizations is not always feasible or ethical.
- Experiments cost a lot of money and time.
- Anyone can flip a coin, but convincing parallel worlds are hard to simulate.
- It's inevitable that some people would deviate from their random assignments.

2.2.2 Experiment : Experimental drug trial

Two goals for this experiment:

1) Internal Validity: Could anything other than the treatment have produced results?

One cause could be that doctors gave the medicine to some special patients, thus breaking the randomization.

2) External Validity: Do the results of the experiment hold in a setting that we care about?

Would the medicine be as successful in the real world as it was in the clinical trial?

One example is birth control pills which may not be as effective in the real world as they are during the test on account of people forgetting to take them on same days.

3 How we conduct behavioral experiments?

3.1 Lab Experiment

3.1.1 Better internal validity

- Greatest procedural control: we can define the setting the experiment takes place.
- Can carefully curate situations

3.1.2 Less external validity

- Artificial context, simple tasks :

  The user may have behaved differently in a "natural" environment.
- Demand effects
- Homogeneous (WEIRD) subject pool

  WEIRD stands for Western, educated, and from industrialized, rich, and democratic.
- Time/scale limitations

3.2 Field Experiment

3.2.1 Better generalization

- Experiment findings apply to at least one real world setting

3.2.2 But:

- Less Control, more potential confounds
- Demand of experiment conflicts with goals of real organizations

- More effort to conduct and manage
- More room for error

4 Examples of Causal Inference

4.1 Natural Experiments:

Sometimes nature runs experiment for us, e.g.:

4.1.1 As-if random:

People are randomly exposed to water sources (Snow, 1854)

4.1.2 Instrumental variables:

An instrument independently shifts the distribution of a treatment. Example: A lottery influences military service (Angrist, 1990)

What can we try to infer? 1) Effect of lottery in making people join the military. 2) What would be somebody's future earnings if your name was called in the lottery?

Natural experiments: Caveats

- Good natural experiments are hard to find.
- They may have many (untested) assumptions.
- The treated population may not be the one of interest.