EXperimental Learning

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Big Data and Social Analytics certificate course

© 2016 MIT / GetSmarter All Rights Reserved (not authorized for commercial use)

MODULE 5 UNIT 1
Video 1 Transcript

MIT BDA Module 5 Unit 1 Video 1 Transcript

Speaker key

AP: Alex Pentland

HY: Hapyak

AP: Okay, well. Welcome back. Hope the course is going well for you and that you're enjoying it.

Today we're going to talk about second-order analysis, which is looking at what we can do that isn't direct. Indirect inference. A lot of times people get stuck with the idea that they have to be able to measure the thing that they're interested in. But often you can measure stuff that's correlated with it, rather than the thing itself.

And so that's a way to get rid of data deserts, when you don't have any data, you find something that correlates with what you're interested in. But it also is something that you want to watch out for because correlation is not the same as causation. In other words, you can find something that goes up whenever this other thing goes up, but that doesn't mean this one causes that one. You can't change this one by changing that one. They might be caused by a third effect altogether, or the causation might go the opposite direction of what you think.

But even if you're not sure of which causes which, you can always use correlation to stratify; to say which one's bigger, which one's smaller, which one's in the middle. Because correlation means the two things go up and down together.

HY: What is the difference between correlation and causation, and why is it important to make this distinction?

Correlation refers to two or more factors behaving in the same manner, such as both increasing or decreasing simultaneously. However, this behavior is not necessarily due to the factors influencing each other. Causation, on the other hand, refers to a factor having a direct influence of the behavior of another factor. It is important to distinguish between the two because if you make a prediction based on the assumption of causation and the factors turn out to be only correlated then you are likely to arrive at incorrect insights.

00:01:32

AP: So let me give you a couple of examples of doing this type of thing. So, this is an example you might remember from earlier on. People moving around in San Francisco. The most, sort of, common places are the big dots; those are the bars and restaurants and so forth. And the data, of course, is just about movement and we don't know particular people, but we can use even anonymous data like this, data about motion only, to do lots of other things.

© 2016 MIT / GetSmarter All Rights Reserved (not authorized for commercial use)

So let me give you some examples. So for instance, imagine we were interested in predicting crime, and this is something that we did in London a couple of years ago. It turns out that when you see these flows of people change dramatically – so, instead of people from this town and this town spending some time in the city square, they stop going to that – that's the sort of thing that's a signal for crime.

Something changed there, probably something that was stressful for the society, and it turns out that predicts crime pretty well. Not crime by any particular person, but that that city square might have crime.

So we didn't know anything about a particular person. We couldn't measure crime directly, but by looking with the people as sensors, where are the people going? We can find something that correlates with crime. And that's a pretty good thing to be able to do.

00:02:54

Or here's another one. So this is in Riyadh, in Saudi Arabia. It turns out that unemployment correlates with different behavior than working. You can of course understand that if you are working, you go to work every day at the same time. If you're unemployed, you probably don't.

But here's something that they didn't know and you probably don't either. When some people have a lot hard time getting a job and other people get a job very quickly, what differentiates the two? It turns out that in Riyadh at least, it's that the people who have hard time getting the job, don't explore as widely as the people who get jobs quickly.

It's not that they're lazy, it's just that they may not have as good a transportation, so they look locally rather than really broadly. And that seems to predict whether or not they'll get a job quickly. Again, we didn't measure anything about skills, about money, about the type of person they are. We just looked at the behavior that they had and used that to infer things about speed of getting re-employed.

Here's a more general thing that's really surprising. So, if you look at these flows of people they predict the wealth of the neighborhood and of the city. So, neighborhoods that have really good mixing, either in terms of people physically or phone calls, tend to be very rich. Ones that, where people don't talk to each other very much and they're not connected to the rest of society tend to be very poor. And when they're poor, they have higher crime, worse health and so on and so on.

00:04:35

So, something that you wouldn't expect, which is the pattern of exploration, predicts all sorts of things, like infant health, longevity, GDP, crime level. Because they're correlated with each other. And they're correlated because people have this foraging behavior where the amount of exploration indicates the amount of stress they have, the amount of confidence they have. And poor societies have less of that than rich societies.

HY: MIT's research has found a direct causal relationship between human foraging behavior and the GDP levels of a country.

True

© 2016 MIT / GetSmarter All Rights Reserved (not authorized for commercial use)

Incorrect. They've found a correlation between foraging behavior and GDP levels, crime, and a number of other things, which has helped them predict those other things. However, other evidence from economics and small-scale experiments makes it likely that the connection is in fact causal: more mixing of ideas likely does cause greater GDP.

False

Correct, well done. They've found a correlation between foraging behavior and GDP levels, crime, and a number of other things, which has helped them predict those other things. However, other evidence from economics and small-scale experiments makes it likely that the connection is in fact causal: more mixing of ideas likely does cause greater GDP.

AP: So, a final thing is that we can begin to use some of these ideas in order to do planning. So, for instance, some colleagues of mine at Imperial, are beginning to use this relationship to plan where to put a new subway. Now they're using this relationship between people mixing and GDP to predict where to put the subway. But they're making a dangerous assumption and the dangerous assumption is that this is causal. Places with more mixing actually will have more GDP and not the other way round. It might be that places with more GDP have more mixing.

But in this case they're probably pretty safe because there's lots of other evidence that shows that bringing new ideas together is in fact what creates GDP. So, they've thought about the correlation versus causation problem and decided it probably really is causation, so we can use it as a design principle, not just for telling the good ones for the bad ones.

© 2016 MIT / GetSmarter All Rights Reserved (not authorized for commercial use)