EXperimental Learning

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Big Data and Social Analytics certificate course

MODULE 3 UNIT 1
Video 2 Transcript

MIT BDA Module 3 Unit 1 Video 2 Transcript

Speaker key

AP: Arek Stopczynski

HY: Hapyak

AS: Let's now talk about noise versus bias in the data and by noise we mean a random error that happens in the data versus bias that is a systemic error in the data. Since noise is random the values are distributed around the true value which means that adding more data usually allows you to get rid of the noise. An example of noise in the data is randomly missing values in the data and as you collect more of them you will be arriving closer and closer to the truth.

00:00:52

The bias, however, is when the data that you are seeing doesn't have the mean in the true value or the mean is shifted and then adding more and more data doesn't actually help. You will be more sure and more sure that what you are seeing is correct but actually it will always be shifted from the true value that you want to measure. And there are ways to deal with those two problems so let's take a very simple, almost toy, example and let's see how we can go for spotting the problem and maybe even fixing it.

The example we want to look at is let's imagine we are studying a population of students on a campus and we want to estimate how much time people are spending indoors versus outdoors. And it's relatively simple problem and let's imagine that we thought, oh, let's attack it by simply looking at the location data. So every few minutes the phone will check the GPS coordinates and we are able to translate these coordinates into indoor versus outdoor, based on a high resolution map of campus.

So we go about to do that and, simply, every time we have a measurement – GPS measurement – and we are seeing that it was inside, we would count it as inside. Every time it's outside we would count it as outside. Can you think about a problem that may happen when we measure things in that way? And if yes, do you think that doing the collection of the data for two months instead of one month, effectively doubling the amount of data, would actually help?

HY: The difference between bias and noise is that bias can’t be reduced through an increase in the volume of data.

True

Correct, well done. Bias will remain no matter how much the volume of the data is increased.

False

Incorrect. Bias will remain no matter how much the volume of the data is increased.

00:02:30

AS: What we can expect would happen is GPS is more likely to fail when you are inside buildings. So the natural bias in the data would be that we would be over counting the data points that would be generated outside of the buildings, simply because every time you are inside there is a higher probability that GPS will not be able to produce the data. And no matter whether you collect a month of data or two months of data or ten months of data, this bias, you can expect, will be relatively consistent in the data. So just collecting more and more data will not bring you closer to the true value.

And this is a very important learning. Even in big data analysis, having more data is not always the solution. Sometimes you actually need to be smart about how you approach the analysis. So could we actually fix it? Could we actually take this toy example of trying to estimate the fraction of time people are spending indoors versus outdoors if we know there is slightly a bias in the data?

00:03:38

Yes, if we know that, we can, for example, try to estimate what is the probability of failure of GPS inside a building and outside of the building and correct our values by this fraction or we can be more aggressive in smoothing of the GPS data when we are inside and outside. So, for example, if we are seeing that someone entered the building and then exited the building, we expect the behavior, the human behavior, to be smooth. So it's not someone jumping in and out of the building every few minutes.

When they entered and they were exited, there's most likely that they have been inside, on this day that we are observing, for a period of time. So we can do more smoothing that would get rid of a lot of this bias in the data. So there are ways to do it but first of all we need to think about is there a reason why, even with infinite or almost infinite amount of data, I would be seeing something that is actually not true? And if you can identify that the question is, of course, how you can correct it?

In the previous models we talked about the data quality and how do you measure that. And having a robust measure for data quality, and actually inspecting it, is a way for you to get the feeling whether you can expect certain biases in the data, related to your particular question, or not. So in our little example with sensing indoor versus outdoor, if you would see that you are missing data in blocks, and because human behavior is actually smooth, you could be now worrying that, oh, I'm actually missing data around certain conditions and maybe this is actually related to the question I'm trying to answer. So there is a risk that there is a bias.

00:05:24

So look at the data quality and think about, am I seeing errors that are not just randomly distributed with respect to the question I'm trying to answer? Am I seeing certain blocks where these errors seem to accumulate because those can be indication of the biases? And, very importantly, remember more data is not always the answer. There are problems that require deeper dive and understanding the data, not just adding more of it.