Harvard Business Review

ANALYTICS

How to Make Better Predictions When You Don’t Have Enough Data

by Kira Radinsky and Yoni Acriche

DECEMBER 29, 2016

When Donald Trump first declared his candidacy for President of the United States, most analysts predicted that he has an incredibly small chance of becoming the Republican nominee. Probably the most prominent of these was Nate Silver from FiveThirtyEight. He estimated that Trump had a 2% chance of winning the nomination. This estimation was based on multiple significant historic data

points about past candidates, such as the background they came from, whether they were widely endorsed by the party, and their past successes and failures. This is a standard prediction approach based on the underlying assumption that what you are trying to predict (Trump) is comparable to its historical antecedents (past GOP candidates) and thus can be evaluated according to their performance. However, as it is clear to us now, in some unique cases like the Trump phenomenon, we could only learn little from recent direct history.

A similar problem crops up in polling. Political analysts use polls in order to estimate the likelihood of a candidate's success. However, polls are not perfect, and usually suffer from multiple types of biases — such as the effect of non-responders, the tradeoff of polling by calling landlines versus cellphones, and changes in voting turnout trends. To overcome these obstacles, political statisticians build models that try to correct polling errors by using data from previous elections. This method is based on the underlying assumption that current and historical polls suffer from the same type of errors. For example, analysts might assume that the population of non-responders is distributed similarly across time — an assumption that may or may not be true.

Compounding both problems, since presidential elections are a relatively rare event, our historical data is limited; in other words, the sample size is relatively small and outdated.

Predictive statisticians in the private sector face similar problems when trying to predict unexpected events, or when working from flawed or incomplete data. Simply turning the work over to machines won't help: most machine learning and statistical mining techniques also hold the assumption that historical data, which is used to train the machine-learning model, behaves similarly to the target data, to which the model is later applied. However, this assumption often does not hold as the data is obsolete, and it is often expensive or impractical to get the additional recent data that holds this assumption.

Thus in order to stay relevant, statisticians will have to get out of the purist position of fitting models that are based solely on direct historical data, and to enrich their models with recent data from similar domains that could better capture current trends.

This is known as Transfer Learning, a field that helps to solve these problems by offering a set of algorithms that identify the areas of knowledge which are "transferable" to the target domain. This broader set of data can then be used to help "train" the model. These algorithms identify the commonalities between the target task, recent tasks, previous tasks, and similar-but-not-the-same tasks. Thus, they help guide the algorithm to learn only from the relevant parts of the data.

In the example of the U.S. presidential elections, we might use this method to understand which international economic and social phenomena might predict the rise of an unexpected candidate like Trump. For instance, while the Trump phenomenon is new to the recent American political climate, on a global level political scholars have been observing this trend for quite a while. In "Trump, Brexit, and the rise of Populism: Economic have-nots and cultural backlash" Ronald Inglehart and Pippa Norris examine the recent rising support of populist parties in many western societies. In Britain, for example, while the UK Independence Party won only one seat in the May 2015 general election, its populist rhetoric fueled anti-European and anti-immigration sentiment, which later led them to win the EU Brexit referendum. Inglehart and Norris find many similarities between the populist rise in different countries; the same effects of economic insecurity in post-industrial economies and a backlash against diversifying societies have driven the same groups of voters to the ballots.

Transfer learning thinking suggests that using the 2016 Brexit voting data from the UK could have allowed statisticians to better understand current global turnout and voting trends. A model that considered data from beyond the U.S. thus might have predicted more support for Trump, especially in demographics that share the same anti-immigration views as was recently seen in the UK.

Politics offers only one case study that highlights the increasing need for novel statistical techniques that can adjust to frequently changing data. The problems that arise from using historical data are also prevalent in many other sectors. While businesses tend to make strategic investments using historical data, for example, we often ignore the possibility that the reality has already changed.

The small-sample-size problem crops up in other places, too. Consider a company with a successful operation in the U.S. that wants to expand to the German market. How can they translate the knowledge they've gained in the U.S. market and apply it to the German expansion? Is there any way to minimize the costs or the risks? Transfer learning methods can help the model to overweight the similarities between the U.S. and the German markets, such as population groups that share similar demographic and economic characteristics, and to underweight the dissimilarities. From a business perspective, this will enable decision-makers to simulate the performance of the company in an environment similar to that of the target market.

Instead of the common techniques of solely using the historical data of the same problem for making predictions, political statisticians and business predictors should also start to use data from similar problems occurring more recently — even if they might not be directly connected. To make the connection between the two problems, transfer learning algorithms help focus the learning process on the more relevant parts of the historical training data.

It's true that historical data is enormously valuable in making predictions. However, the ability to use more advanced techniques in data science will help leverage information from current comparable events, which is crucial for making more accurate predictions – especially when the historical data is limited, or the environment is uncertain. To avoid critical mistakes in prediction, data analysts need to adopt new methods that will enable them to translate knowledge from different time periods and domains.

Kira Radinsky, Ph.D. is the chief scientist and the director of data science of eBay, co-founded SalesPredict (acquired by eBay in 2016), and serves as a visiting professor at the Technion, Israel's leading science and technology institute.

Yoni Acriche is lead data scientist at eBay and previously the head of data science at Salespredict (acquired by eBay).

This article is about ANALYTICS

FOLLOW THIS TOPIC

Comments

Leave a Comment

POST

6 COMMENTS

Shanmuga Murugan 3 days ago

Transfer Learning can be combined with a level of Artificial Intelligence to predict more accurately. AI based self learning models would be able to tweak the algorithms on its own based on the variety of data. The test and training data based models are beginning to fail as the human behavior is becoming more and more unpredictable

REPLY 0 0

JOIN THE CONVERSATION

POSTING GUIDELINES