Ipsos MORI

UNLOCKING VALUE
WITH DATA SCIENCE

BAYES' APPROACH: MAKING DATA WORK HARDER

2016

DELIVERING VALUE WITH DATA SCIENCE
BAYES' APPROACH - MAKING DATA WORK HARDER

The Ipsos MORI Data Science team increasingly use Bayesian techniques to help our clients make the most of their data. This approach is particularly useful for:

- **making more accurate predictions**
- **improving the veracity of complex models**
- **compensating where data is sparse**

BUT WHO IS BAYES AND WHAT IS A BAYESIAN APPROACH?

Thomas Bayes
1701 – 1761

It may surprise you to learn that the side project of a Presbyterian minister in Tunbridge Wells in the mid-18th century is still influencing how we look at big data sets and make predictions today. Reverend Thomas Bayes gave his name to a very simple piece of mathematics which links together information from two stages in a process, or from two sources, to assess probability, or make a prediction – this is Bayes' Theorem or Rule.

One example involves the accuracy of medical tests and their use for screening. Suppose there is a disease for which the test is 99% accurate (ie 1 time out of 100 it gives the wrong result). Now suppose I take the test and it is positive. What is the chance that I have the disease? A Bayes approach says that it depends what my chances of having the disease were before I took the test. So if the disease only affects 1 person in 10,000, then my positive test result is much more likely to be a false positive. This is why a doctor will often only test for a disease if there are other reasons (such as "symptoms") which increase your chances of having it.

BAYES' THEOREM EXPLAINED

In its simplest form Bayes' Theorem involves working backwards from an observation to estimate the probability of how it happened. In the classic formula below, A and B are events, with A happening (or not) before B, P() means "the probability of" the event in the brackets and the bar "|" mean "given that":-

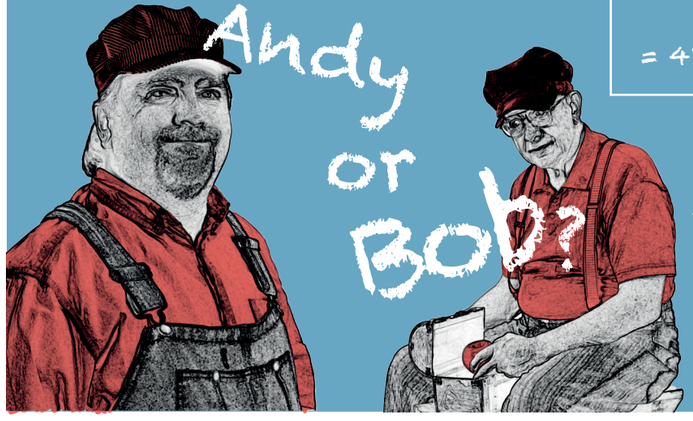
$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

So: "The Probability of A having happened, given that B has happened is equal to the probability of B happening given that A happened, multiplied by the probability of A happening, divided by the probability of B happening".

To put this into practice, an example from my morning commute might help. I catch a train to work from a small station, where just two different drivers are used during the week. Talking to the station staff I discover:

Andy works three days, Monday to Friday, while Bob works the other two, but Bob's work days are completely random. When Andy drives, the train is late 50% of the time, but when Bob drives it is late 80% of the time. This morning the train was late – so who is more likely to have been the driver?

Well of course either driver could have driven the train. One argument is that it is always more likely to have been Andy (because he works more days each week), while another is that Bob is more likely because he is more likely to be late. Bayes' Theorem combines those two elements together to give an exact answer.



The equation below shows the two ways that we can get to the outcome "train was late".

WE CAN WRITE THE PROBABILITY THAT ANDY WAS THE DRIVER OF THE LATE TRAIN AS:

Prob (Andy was the driver, GIVEN THAT the train was late).

And using Bayes' Theorem, this =

$$\frac{\{\text{Prob (Andy being the driver)} \times \text{Prob (train was late)}\}}{\{\text{Prob (train was late regardless of driver)}\}}$$

$$= \{0.6 \times 0.5\} / \{0.6 \times 0.5 + 0.4 \times 0.8\}$$

$$= 48\%$$

So it is slightly more likely (52% chance) that Bob drove the train this morning.

> "Andy works three days, Monday to Friday, while Bob works the other two, but Bob's work days are completely random. When Andy drives, the train is late 50% of the time, but when Bob drives it is late 80% of the time. This morning the train was late – so who is more likely to have been the driver?"

BAYES' APPROACH IN A NUTSHELL

We can use Bayes' Theorem to improve the accuracy of our predictions and modelling, thanks to Bayesian inference, where Bayes' Theorem is used to update the probability for a hypothesis as more evidence or information becomes available.

A Bayes approach starts with an assumption about the thing we are interested in and uses observed data to modify that assumption so that we get both an estimated value for this thing, as well as an indication of the reliability of that estimate.

The reliability will depend on the variation in the data, the sample size, but also now on the veracity of the initial assumption – so if the approach is to be beneficial then we need our initial assumption to be good!

However, if we assume a "flat" initial assumption, whereby we have no prior assumptions (statisticians refer to this as a "uniform prior") then the Bayesian result and the traditional result are numerically the same. So a Bayesian approach should be no worse and in some situations can give us a better result. The key to being Bayesian is describing the world in terms of probabilities and then using data to adjust those probabilities – to get a 'tighter' range of likely outcomes.

WHY IS BAYES IMPORTANT?

Although Bayes' Theorem is described in terms of discrete variables, the rule can be extended to continuous variables, and that is often how we apply it in market research. The key principle is that we describe a situation in terms of probabilities and use data to adjust our description.

HOW IPSOS MORI APPLIES BAYESIAN TECHNIQUES TO ADD VALUE

1. Getting more accurate estimates – using Bayesian Estimation we can combine results from a survey with previous knowledge or data. For example, let us suppose that based on our previous surveys on drinking behaviour, we felt that the proportion of the population who regularly drink is between 30% and 40%, with 35% the "most likely" value. We then survey 2,000 people and find that 38% regularly drink. Now we can combine the new observation (the survey) with the a priori distribution (a probability distribution built from our previous expectations) to create a posterior distribution, which is our new probability distribution for the underlying proportion of the population.

2. Improving key driver analysis – using Bayesian Modelling and Bayesian Networks we can better understand the relationship between two variables. This might be, for example, a satisfaction score (A) alongside a performance rating for a key attribute (B). Typical modelling work attempts to find some linear relationship (eg satisfaction score = 4.1 + 0.3 * delivery performance + random error) whereas Bayesian Networks allow for the relationship between A and B to be driven by statistical distributions , so that A has a set of probabilities of taking different values according to the value of B. This can produce more flexible models (non-linear), but requires more input from the analyst in terms of making and testing assumptions.

3. Optimising combinations - Hierarchical Bayesian models are used in choice exercises (conjoint, max diff) where we give each respondent a limited selection of all possible choices. The models are described as Bayesian, because we take a distribution of parameters, apply the observed data and use it to refine our estimated distribution. We repeat this process a number of times so the estimation process is an iterative method which starts from an initial assumed distribution and converges towards a stable solution. This technique, often used for product, price and feature optimisation allows segmentation at a more individual level based on underlying needs.

4. Greater granularity – Small Area Estimation is where we again apply Hierarchical Bayesian approaches to build a series of separate models for different areas covered within a larger survey in order to achieve greater granularity, for example for implementation of local or regional results.

SO WHAT ARE THE BENEFITS FOR AN END USER?

The key benefit is the efficient use of all the data that we have – both survey and other sources – to obtain a more robust measurement.

However there are 3 key additional benefits:

1. Where we have strong confidence in our prior information then we can quote a narrower interval of reliability.

2. We can now run multilevel modelling even with very low numbers of groups.

3. Respondent-level estimates (including conjoint utilities) mean that we can provide more sophisticated simulation tools to clients. These give better estimates of, for example, the source of volume from product changes, which we use to measure the risk to overall profitability.

NEW BAYESIAN TECHNIQUES

We are constantly exploring applications of new Bayesian methods such as Approximate Bayesian Computation, Variational Bayesian Methods and the intriguingly titled MAD Bayes to investigate what benefits this may bring to our analysis of big data for our clients. For example, a recent paper by Bradley Efrom (RSS Journal 2015) has shown a way to give the frequentist accuracy of a Bayesian estimate which should enable us to provide confidence intervals for a wider variety of modelling work.

Finally - the health warning. Like all analytical work, we should end with a warning about limitations and reliability. There are situations where a Bayesian approach does not make sense – when a population is changing for example and/or we are measuring that change. Also a warning to beware of two claims we have seen recently:

1. "We don't need as many data points if we use a Bayesian approach"

2. "We don't need as robust a sample if we use a Bayesian approach".

Although there may be a small element of truth behind these statements, we would in general disagree with their widespread application as a panacea to poorly designed surveys. Remember - if you are using less data or less accurate data, then you must be relying on something else. Ensure first that your assumption about the prior is robust.

Ipsos MORI

ABOUT IPSOS MORI

Ipsos MORI, part of the Ipsos group, is one of the UK's largest and most innovative research agencies, working for a wide range of global businesses, the FTSE100 and many government departments and public bodies.

We specialise in solving a range of challenges for our clients, whether related to business, consumers, brands or society. In the field of data science, we have a large and diverse team of experts including mathematicians, statisticians, data scientists and behavioural economists. We are constantly seeking to break new ground in the understanding and application of large and complex data sets.

We are passionately curious about people, markets, brands and society. We deliver information, and analysis that makes our complex world easier and faster to navigate and inspires our clients to make smarter decisions.

CONTACT

Clive Frostick
Head of Analytics,
Ipsos MORI
T: +44 (0)20 8861 8755
E: clive.frostick@ipsos.com
www.ipsos-mori.com
@ipsosmori