ON ASSESSING REAL WORLD PREDICTION SKILL

BY DAVID ALDOUS

Professor in the Statistics Department at U.C. Berkeley

Introduction

Let me start with a puzzle. Is it possible to devise a quiz contest (on any topic, not necessarily mathematical) with the following properties?

Answers will be scored objectively – no subjective judgments (as would be needed for creative writing, for instance). Contestants who end with a better overall score will – beyond reasonable doubt – be better at the subject matter of the quiz.

The questions refer to substantive real-world matters, rather than fantasy (islands with liars and truth-tellers) or self-referential "how would most other contestants answer this question?" No person (or computer, etc) knows or will ever know the correct answer to any of the questions.

So this looks impossible at first sight – how can one grade objectively without knowing the answers? Now puzzles like this inevitably involve some kind of trick. But my trick is rather mild – an everyday quiz can be graded quickly, but for my quiz you have to wait a while to find your scores. If you can think of a less tricky such quiz, please let me know!

The Good Judgment Project

Here are 4 questions that people with an interest in world affairs might be pondering as I write (September 2017):

1. *Before 2018, will Russia officially announce that it is suspending its participation in or withdrawing from the Intermediate-Range Nuclear Forces Treaty?*
2. *Before 2018, will 5 or more countries experience 10 or more cases of poliovirus?*
3. *Before 2018, will there be a lethal attack on a US military vessel in the Red Sea, Gulf of Aden, Persian Gulf, or Gulf of Oman?*
4. *Before 2018, will China deploy a deep sea oil rig in another country's Exclusive Economic Zone without that country's permission?*

In the current *Good Judgment Project Classic Geopolitical Challenge* [1] participants are asked to assess the current probabilities of such future events. To reiterate, they are **not** asked to give a Yes/No prediction, but instead are asked to give a numerical probability, and to update as time passes and relevant news/analysis appears. Unlike school quizzes, you are free to use any sources you can – if you happen to be a personal friend of Vladimir Putin then you could ask him for a hint on the first question.

Do you think it is ridiculous to pose such questions to non-experts? If so, do you think that trial by jury is ridiculous? In both cases the point is to look at evidence and at expert opinion before giving an answer.

What makes this setting conceptually interesting is that no one will ever know the correct probabilities. Nevertheless one can judge participants' relative ability to assess such probabilities, after the outcomes are known. Explaining this paradox is the focus of this article.

Mean Squared Error

How can we assess someone's ability? We will use several very basic concepts from probability. A random variable X is, informally, a quantity with a range of possible numerical values, the actual value being determined by chance in some way. The *expectation* of X is a real number, written $E[X]$, analogous to the *average* of numerical data. If we seek to predict the value of a *random variable* X, our prediction has to be some constant x_0. The (random) *squared error* of our prediction is the random variable $(X - x_0)^2$, and the expectation of that random variable, in symbols $E[(X - x_0)^2]$, is called the *mean squared error (MSE)* of the prediction. And the "best" predictor in the sense of minimizing the MSE is just the constant $x_0 = E[X]$.

As a specific example, for a single throw X of a fair die, the MSE from predicting x is

$$\frac{1}{6}\sum_{i=1}^{6}(i-x)^2 = \frac{35}{12} + \left(x - \frac{7}{2}\right)^2$$

which is minimized at $x_0 = E[X] = 7/2$. The idea of using *squared* errors goes back to Gauss in the context of errors in astronomy observations, and is widely used in classical statistics because of its nice mathematical properties, which we will exploit in several ways.

An *event*, in the probability context, will either happen or not happen, and we can represent an event as a random variable, taking value 1 if the event happens and value 0 if not. This allows us to use "squared error" to score our predictions. If we predict 70% probability for an event, then our "squared error" is

$$\text{(if event happens) } (1.0 - 0.70)^2 = 0.09$$

$$\text{(if event doesn't happen) } (0.7 - 0)^2 = 0.49.$$

So suppose you participate in a *prediction tournament* like the Good Judgment Project. For simplicity let's suppose that participants just make a one-time *forecast*, a probability prediction, for each event. After the outcomes of all the events are known, your final score will be the average of these squared errors. As in golf, you are trying to get a **low** score.

In the next section I will argue that this is the right way to score. Just as in golf, your score really does indicate how good you are at the prediction game, give or take a small amount of luck.

A very little algebra

When you make a "probability p" forecast for a certain event, your squared error score will be

$$\begin{aligned}\text{score} &= (1-p)^2 \text{ if event occurs} \\ &= p^2 \text{ if not.} \qquad (1)\end{aligned}$$

Suppose you actually believe the probability is q. What p should you announce as your forecast? Under your belief, your mean score (by the rules of elementary mathematical probability) equals $q(1-p)^2 + (1-q)p^2$ and a line of algebra shows this can be rewritten as

$$(p-q)^2 + q(1-q). \qquad (2)$$

Because you seek to minimize the score, and because all you are able to choose is p, you should announce $p = q$, your honest belief – with this scoring rule you cannot "game the system" by being dishonest, that is by announcing a value of p which is not your true belief for the probability.

Now write q for the true probability of the event occurring (recall we are dealing with future real-world events for which the true value q is unknown), and write p for your forecast probability. Then your (true) mean score, by exactly the same calculation, is also given by (2). The term $(p-q)^2$ is the "squared error" in your forecast probability.

Now consider two participants, A and B, making forecasts p_A and p_B for the same event which has (unknown) probability q. Then (2) implies that

$$E[\text{score (A)}] - E[\text{score (B)}] = (p_A - q)^2 - (p_B - q)^2. \qquad (3)$$

In a prediction tournament there will be a large number n of events, with unknown probabilities $(q_i,\ 1 \le i \le n)$ and with forecasts $(p_{A,i}, p_{B,i},\ 1 \le i \le n)$ chosen by the participants. We **would like to** measure how good a participant is by the average squared-error of their forecast probabilities

$$\text{MSE}(A) = \frac{1}{n}\sum_i (p_{A,i} - q_i)^2 \qquad (4)$$

But this is impossible to know, because we don't know the q's. However, (3) implies that for the final scores (the average of the scores on each event)

$$\begin{aligned}&E[\text{final score (A)}] - E[\text{final score (B)}] \\ &\qquad = \text{MSE}(A) - \text{MSE}(B). \qquad (4)\end{aligned}$$

Now your actual final score is random, but by a "law of large numbers" argument, for a large number of events it will be close to its mean. Informally,

$$\begin{aligned}\text{final score (A)} &= E[\text{final score (A)}] \\ &\quad \pm \text{ small random effect.} \qquad (5)\end{aligned}$$

Putting all this together,

$$\begin{aligned}\text{MSE}(A) - \text{MSE}(B) &= \text{final score (A)} - \text{final score (B)} \\ &\quad \pm \text{ small random effect.}\end{aligned}$$

Now we are done: the MSEs are our desired measure of skill, and from the observed final scores we can tell the relative skills of the different participants, up to a small amount of luck.

The mathematical bottom line

Rephrasing the argument above, an individual's score is conceptually the sum of three terms. Write q_i for the (unknown) true probability that the i'th event happens.

- A term $\frac{1}{n}\sum_i q_i(1-q_i)$ from irreducible randomness. This is the same for everyone, but we don't know the value.
- Your individual MSE (4), where "error" is (your forecast probability - true probability)
- Your individual luck, from randomness of outcomes.

The analogy with golf continues to be helpful. A golf course has a "par", the score that an expert should attain. Your score on a round of golf can also be regarded as the sum of three terms.

- The par score.
- The typical amount you score over par (your *handicap*, in golf language).
- Your luck on that round.

So a prediction tournament is like a golf tournament where no-one knows "par". That is, you can assess people's relative abilities, but we do not have any external standard to assess absolute abilities.

And the real world?

We've seen the mathematics, but what is the bigger picture? After all, one could just say it's obvious that some people will be better than others at geopolitical forecasts, just as some people are better than others at golf.

To me it is self-evident that one should make predictions about uncertain future events in terms of probabilities rather than Yes/No predictions. So it is curious that, outside of gambling-like contexts, this is rarely done. Indeed the only common context where one sees numerical probabilities expressed is the chance of rain tomorrow.

A major inspiration for current interest in this topic has been the work of Philip Tetlock. His 2006 book [2] looks at extensive data on how good geopolitical forecasts from political experts have been in the past (short answer: not very good). That book contains more mathematics along the "how to assess prediction skill" theme of this article.

What makes some people are better than others at forecasting, and can we learn from them? That is the topic of Tetlock's 2015 book [3], which reports in particular on an IARPA [5] sponsored study of a prediction tournament similar to the current one [1], though where participants were assigned to teams and encouraged to discuss with teammates. Their conclusions relate success to both cognitive style of individuals and to team dynamics.

Finally, readers of this magazine may be interested in a recent paper [4] claiming that Canadian strategic forecasters are better than their U.S. counterparts!

References

[1] Good Judgment Open. www.gjopen.com

[2] Philip E. Tetlock, Expert Political Judgment: How Good Is It? How Can We Know?, Princeton University Press (2006).

[3] Philip E. Tetlock and Dan Gardner, Superforecasting: The Art and Science of Prediction, Crown (2015).

[4] David R. Mandela and Alan Barnes. Accuracy of forecasts in strategic intelligence, Proceedings of the National Academy of Science 111 10984 – 10989 (2014).

[5] Intelligence Advanced Research Projects Activity. https://www.iarpa.gov

About the author

David Aldous has been Professor in the Statistics Dept at U.C. Berkeley, since 1979. A central theme of his research in mathematical probability is the study of large finite random structures, obtaining asymptotic behavior as the size tends to infinity via consideration of a suitable infinite random structure. He has recently become interested in articulating critically what mathematical probability says about the real world.