Minimax vs Bayes prediction

By
D. Blackwell
University of California, Berkeley

Let $x = (x_1, x_2, \ldots)$ be an infinite sequence of 0s and 1s, initially unknown to you. On day $n = 1,2,\ldots$ you observe $h_n = (x_1, \ldots, x_{n-1})$, the first $n-1$ terms of the sequence, and must predict x_n. What is a good prediction method, and how well can you do?

A *prediction method* p is just a function that associates with each finite sequence h of 0s and 1s a prediction $p(h) = 0$ or 1, your prediction of the next x when you have observed history h. Denote by $w_n(p,x)$ the proportion of correct predictions that method p makes against sequence x in the first n days.

Looking for a p that does well against every x seems hopeless, since for every p there is an x for which $w_n(p,x) = 0$ for all n: p is always wrong against x. You can construct x sequentially: choose x_1 so that p's first prediction is wrong:

$$x_1 = 1 - p(\text{empty sequence}).$$

Then choose x_2 so that p's second prediction, given x_1, is wrong:

$$x_2 = 1 - p(x_1)$$

etc.

But we can improve matters by allowing randomized predictions. For instance if we toss a fair coin everyday, predicting 1 on heads, 0 on tails, the strong law of large numbers guarantees that, for every x, the proportion of correct predictions will approach 50% as $n \to \infty$, with probability 1. A *random prediction method* is a function p that specifies the probability of predicting 1 next, given the past x-history $x_1, \ldots, x_{n-1}$ and the past predictions $y_1, \ldots, y_{n-1}$. Thus p associates with each x sequence a random sequence $y = (y_1, y_2, \ldots)$ of predictions such that

$$P(y_n = 1 \mid y_1, \ldots, y_{n-1}) = p(x_1, \ldots, x_{n-1}, y_1, \ldots, y_{n-1}).$$

For any x and p, define

$$\begin{aligned} &r_n = 1 \text{ if } y_n = x_n,\ r_n = 0 \text{ if } y_n \neq x_n, \\ (1)\qquad &u_n = (x_1 + \cdots + x_n)/n,\ v_n = (r_1 + \cdots + r_n)/n, \end{aligned}$$

so that u_n is the proportion of 1s among $x_1, \ldots, x_n$ and v_n is the proportion of correct

predictions in the first n days. It follows from a vector minimax theorem (Blackwell [1956]) that there is a random prediction method p_0 such that, for every x,

$$(*) \qquad \liminf_{n\to\infty} (v_n - \max(u_n, 1 - u_n)) \geq 0 \quad \text{a.s.}$$

Thus, using p_0, we know that, after many days, if the historical frequency of 1s is 70%, we shall have been right nearly 70% of the time, at least; if at a later time the historical frequency of 1s has dropped to 20%, at that point we shall have been right nearly 80% of the time, at least, etc.

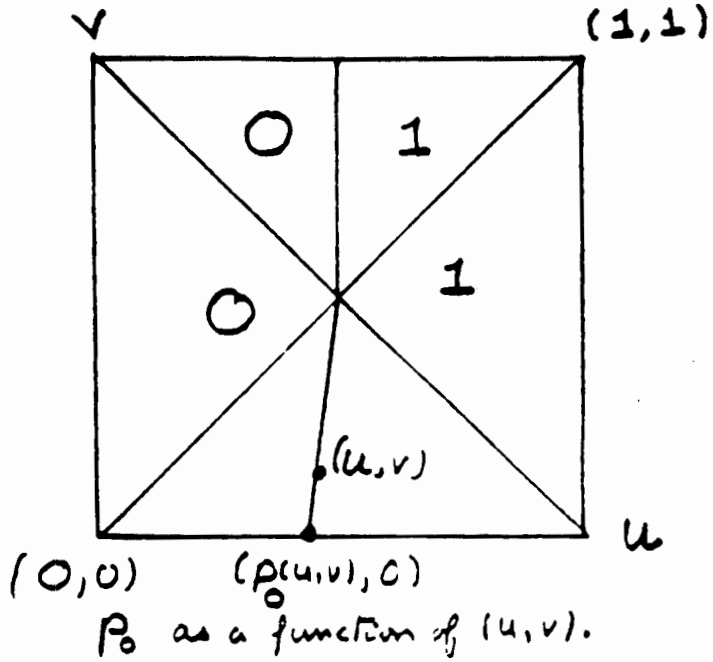
Here is one such p_0. Given $x_1, \ldots, x_{n-1}, y_1, \ldots, y_{n-1}$, p_0 will depend only on u_{n-1} and v_{n-1}, as follows. With $u = u_{n-1}$, $v = v_{n-1}$,

$p_0 = 0$ if $u \leq 1/2$ and $v \geq u$,

$p_0 = 1$ if $u > 1/2$ and $v \geq 1 - u$,

$p_0 = (u - v)/(1 - 2v)$ if $v < \min(u, 1 - u)$.

Here is a picture:



p_0 as a function of (u,v).

For $v < \min(u, 1 - u)$, p_0 is the u-coordinate of the intersection of the line through (1/2, 1/2) and (u,v) with the u-axis. The proof that p_0 satisfies (*) for every x is a bit messy. We shall refer to p_0 as the minimax predictor.

If, instead of insisting on (*) for every x, we put a probability distribution λ on the set

X of sequences x, and look for a prediction method p that satisfies (*) except for a set of xs having λ probability zero, things become much easier. Then the most obvious nonrandom prediction — predict the more probable result — works, and the proof is simple.

Theorem 1. *Let* $x = (x_1, x_2, \ldots)$ *be any 0-1 stochastic process, and define*

$$y_n = 1 \text{ if } P(x_n = 1 \mid x_1, \ldots, x_{n-1}) \geq 1/2,$$

$$= 0 \text{ otherwise.}$$

Then, with r_n, u_n, v_n *as in (1), (*) holds.*

Proof. Write $E_0(\cdot)$ for $E(\cdot \mid x_1, \ldots, x_{n-1})$ and put $z_n = E_0(x_n)$, $t_n = E_0(r_n - x_n)$, so that $t_n = \max(z_n, 1 - z_n) - z_n \geq 0$. With $w_n = r_n - x_n - t_n$, we have $E_0(w_n) = 0$, so that $E(w_n \mid w_1, \ldots, w_{n-1}) = 0$. Thus from a well-known strong law of large numbers (see Stout [1974], Theorem 3.3.1)

$$\bar{w}_n = (w_1 + \cdots + w_n)/n \to 0 \text{ a.s.}$$

But $\bar{w}_n = v_n - u_n - \bar{t}_n$, where $\bar{t}_n \geq 0$, so that

$$\liminf(v_n - u_n) \geq 0 \text{ a.s.}$$

Similarly, with $T_n = E_0(r_n - (1 - x_n))$, we have $T_n = \max(z_n, 1 - z_n) - (1 - z_n) \geq 0$ and, with $W_n = r_n - (1 - x_n) - T_n$, we get $\bar{W}_n \to 0$, and

$$\liminf(v_n - (1 - u_n)) \geq 0 \text{ a.s.}$$

Thus

$$\liminf(v_n - \max(u_n, 1 - u_n)) \geq 0 \text{ a.s.}$$

We shall refer to the predictor y in the Theorem as the *Bayes predictor*.

The contrast between the minimax predictor p_0 and the Bayes predictor y is strong. The minimax predictor is not obvious, it is randomized, it satisfies (*) for every x, but the proof is not easy. The Bayes predictor is extremely obvious, it is not randomized, it satisfies (*) only for almost all x, and the proof is simple.

The random predictor p_0, as defined, does not take advantage of apparent patterns in the data. For instance against the sequence x = (010101....) the long run frequency of correct predictions would be only 50%. We could however split our prediction problem into two separate problems: predicting the sequence x_0 of xs after a 0 and the sequence x_1 of xs after a 1. For the above x we have $x_0 = (1111...)$ and $x_1 = (0000...)$

and p_0, applied separately to x_0 and x_1, predicts perfectly.

More generally, denote by S_n the set of 2^n sequences of 0s and 1s of length n. For each x and each $s \in S_n$, put $x_s(k) =$ the x-coordinate immediately after the kth occurrence of s in x (if there is a kth occurrence). By applying p_0 separately to each x_s, we can take advantage of patterns in the x_s: there is a random prediction method p_n such that for every x, and every s that occurs infinitely often in x, we have

$$(**) \qquad \liminf (v_n(s) - \max(u_n(s), 1 - u_n(s)) \geq 0 \quad \text{a.s.}$$

What is the corresponding non-randomized predictor that satisfies (**) for almost all x? The same as before: predict the more likely result — not separately for each s, but overall, given the entire past (information never hurts). We formuate the result as

Theorem 2. *Let* $x = (x_1, x_2, \ldots)$, *be any 0-1 stochastic process and let* $a = (a_1, a_2, \ldots)$ *be any process. Define*

$$y_n = \begin{cases} 1 \text{ if } P(x_n = 1 \mid a_1, x_1, \ldots, x_{n-1} a_n) \geq 1/2, \\ 0 \text{ otherwise.} \end{cases}$$

Then, with r_n, u_n, v_n *as in (1), (*) holds.*

To get the (**) result for the Bayes predictor apply Theorem 2 separately for each s, with $x = x(s)$ and the a process being the blocks of coordinates between the coordinates of the $x(s)$ process.

The proof of Theorem 2 is like that of Theorem 1, with $E_0(\cdot)$ defined as $E(\cdot \mid a_1, x_1, \ldots, x_{n-1} a_n)$.

References

Blackwell, D. [1956]. An analog of the minimax theorem for vector payoffs. *Pacific Jour. of Math.* [6], pp. 1-8.

Stout, W.F. [1974]. Almost Sure Convergence. *Academic Press.*