Assessing Human Error Against a Benchmark of Perfection

Date 05/11/17 No. 1

Anderson, Kleinberg, Mullainathan

Can we analyse a large dataset of decisions to discover the instances in which people are most likely to make errors? Instead of using ML to make optimal decisions, it will predict when humans are more prone to error.

Chess is an ideal system to analyse:
- It presents a human player with concrete decisions of different value
- The ground truth (correctness of a given decision) is feasibly computable
- Non-trivial, even for skilled players.

In order to use chess as a model system, there are three obvious approaches
i) Construct problems with a well-defined ground truth. But it is difficult to amass a large dataset.
ii) Use chess databases and chess engines, which can give a large dataset. However, even today it is difficult to find a mapping between the engine's choice and a determination of human error.
iii) Calculating the minimax value of a position: either white wins, black wins, or it is a draw (with best play). However, this is computationally intractable.

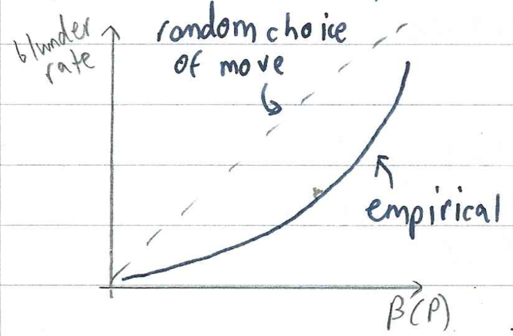
A better way is to assess errors using tablebases - chess has been solved for all positions with at most k pieces on the board, for small k. These are constructed by working backwards from terminal positions.
We can start with the large database of recorded games (ii), then restrict to the subset of $\leq k$-piece positions and compare the move played with the best move according to a table lookup. Limiting to this subset also lets us see how different people react to the same position.

Date No.

To predict errors, the features are:

- the skill of the decision-maker (Elo)
- time available to make the decision (time controls)
- inherent difficulty (proportion of legal moves which are blunders)

To be specific, in a position P there are $n(P)$ legal moves, of which $b(P)$ are blunders. We impose $1 \leq b(P) \leq n(P) - 1$. Empirically, it was found that the blunder rate is monotone in $b(P)$. This leads us to define the blunder potential $\beta(P) = b(P)/n(P)$ (as opposed to the empirical blunder rate.



A simple model that fits these results suggests that players are c times more likely not to blunder ($c > 1$). Then the empirical blunder rate $\gamma_c(P)$ is:

$$\gamma_c(P) = \frac{b(P)}{c(n(P) - b(P)) + b(P)} = \frac{\beta(P)}{c - (c-1)\beta(P)}$$

Fitting this gives $c \approx 15$ for amateurs and $c \approx 100$ for GMs.

Player skill

The empirical blunder rate is a smoothly declining function of the Elo. This skill gradient looks the same for different $\beta(P)$, but with vertical offsets - $\underline{\beta + 0.2 \Rightarrow 600 \text{ Elo}}$. Thus the difficulty of the position may be more important than skill, encouraging the fundamental attribution error.

The dataset even allows for blunder analysis on fixed positions. Not every position is skill-monotone (decreased blunders as Elo ↑).

Date No.

Time

- The dataset includes a large number of 3-minute games. Define $g(t)$ as the empirical blunder rate with t seconds left. $g(t)$ increases sharply as $t \to 0$, but flattens out for $t > 10$.
- Blunder potential still plays a very important role: $B + 0.2 \Rightarrow 50\text{sec}$.
- For higher Elo, extra time confers a relatively greater advantage.
- It turns out that more time spent correlates positively with empirical blunders.

Prediction

- Instead of just considering the current position, we can construct a game tree of depth d, with P at the root.
- If there are n moves, b of which are blunders, we denote the non-blunders as $m_1, m_2, \dots m_{n-b}$ and the blunders as $m_{n-b+1}, \dots m_n$, leading to positions $P_1, P_2, \dots, P_n$. Let T_0 be the indices of non-blunders, and T_1 for blunders.
- If $B(P_i)$ is high, it will be hard for the opponent to capitalise on your blunder, so you may be less likely to notice that it is a blunder.
- We can aggregate this to use as features, defining: $b(T_1) = \sum_{i \in T_1} b_i$, $n(T_1) = \sum_{i \in T_1} n_i$ and likewise for T_0
- For training, a balanced dataset of 600,000 instances was fit by a decision tree with features as discussed.

Results

- Skill and remaining time give 55% and 53% accuracy respectively, while difficulty gives 73%
- Most of the predictive power comes from depth-1 features.

- Humans have a slightly lower accuracy at predicting blunder positions.
- For positions with only one blunder, blunders become less likely with higher $b(T_1)$ ← people less likely to play moves which increase complexity.
- Once difficulty is controlled, skill and time give 63% accuracy.