Raj Dasani
Data H195A

Final Prospectus

Project Overview: Assigning New Values to Chess Pieces via ML

Motivation of Project

As a massive fan of chess, and as one that plays a lot, I am always looking to get better in any ways I can. And for most casual players like myself, there are two ways we typically learn: (1) analyzing in a game by game basis and understanding the optimal computer move or (2) seeing all your games in aggregate via success rates based on how you decided to start a game (dubbed as “openings”, a common way to learn and get better in chess). But with how complex the game is, especially with the last two phases of the game (the middle and end-game), there is a lot about our individual playing strategy that we do not know from our consistency of our play throughout the entire game, to common, similar mistakes we make in the middle and end parts of the game.

The one way standard chess routine helps us understand the current status of a game is via the value of pieces. Typically, all the materials were valued as such: a pawn is worth 1 point, the knight and bishop are worth 3, the rook is worth 5, and the queen is worth 9. These values were made rather arbitrarily, first mentioned by the Modenese School in the 18th century, and have been debated for centuries ever since, with arguments about what the relative value of pieces or groups of pieces should be. But with every position / situation, pieces can have a different relative value, based on its positioning against the opponent’s pieces and its vision of the board. Many other chess experts have looked to create valuations of the pieces, such as Larry Kaufman suggesting values for each piece in the middlegame or with many experts understanding valuation of pieces via combinations of the other pieces on the board (such as having two bishops vs having a rook and a pawn).

Instead of simply trying to create my own valuation for each piece of the board, I want to cater my research towards improvement and creating an understanding of how different levels of players use different pieces. For example, can we find out if, on average, the players rated at an intermediate level, undervalue the bishop more than the levels above them? Through this we can create recommendations and cater learning strategies for players about how they can improve, and which pieces they should focus on utilizing more throughout the entirety of the game. This is where the beauty of ML can help us understand the ways a player can improve, by helping aggregate the high amount of variations of the game.

Dataset Description

The dataset I will use is a collection of Lichess games from a variety of ratings from the past year. Lichess is the second most popular chess-playing site in the world, after Chess.com. Lichess uses their API to collect the most recent games taken from their users, and collects information from the ratings of each player to the play by play script of the game, writing in standard chess notation. For the purpose of this report, it should be noted that Lichess's rating system is a bit different than chess.com's and the standard chess rating systems, as they have their own independent rating system, which many experts determine as "inflated." This still allows us to create the levels of players desired to understand patterns across different levels, but these may need to be inflated from the standard rankings in order to match Lichess's ratings.

The dataset I am using to start was downloaded from kaggle, and is used for the same reason as me, to analyze chess games. Lichess now releases a massive amount of games per month, all under their open database, within a Creative Commons CC0 license.

The data is bound to have its limitations, and may be a bit hard to parse through, especially the scripts of games, but once the first couple games have some reproducible code created, it will be straightforward to apply it to the rest of the games.

Methodology Description

When discussing how the data will be used, we will start from collecting games at every level, and pulling as many recent games as possible from the lichess website.

To process the data, I will start by extracting the text string of game moves (one of the columns) and parsing through the game to calculate for each piece: the amount of pieces it was taken for, the amount of squares each type of piece moved, the amount of checks it made, and if it was part of the checkmate (boolean). With these values done for every game, I will apply the same analysis for every game and then segment across levels and compare initial differences. Even before any ML is needed, look into if certain pieces were moved, used to take, or were in the end game more often than other pieces for each level, and if there were any significant differences within these categories.

Then for the analytics and modeling, I will start off with unsupervised learning in order to see if I can find a way to create an arbitrary value for a piece without using the old ones as an anchor. I will start this probably with some logistic regression to create log odds for each piece, as inspired by a blog report I read. After this, I will look into other unsupervised learning methods, such as looking into some with association: if groups of pieces or common patterns occurred that makes a piece or group of pieces more valuable at one level than another. This could help me a lot with understanding relative value above all in order to create dynamic values that could be helped for quantifying individual players or types of players and their play styles. How this will necessarily lead to plain numbers may yet to be seen, but I believe may be easily manipulated once a couple results are seen. In other words, once some of these methods are put in place and potentially combined, relative factors should emerge that I could use to create numerical values for pieces, potentially using the pawn = 1 as my base value. But additionally, using some supervised learning using the current valuations of pieces as anchors will be something I try to add to this report in some way, whether it is justifying/critiquing the current values, or understanding if those values apply to a certain level of chess player more than the other. Using some basic supervised learning, such as with regressions, I could see more directly what contributes to winning at different levels as my explanatory variable and use that to quantify some sort of "win value" for each piece.

Proposed Outcome

The ideal outcome will be creating a table and data frame that estimates the piece value for each general piece for each level of play. To be explicit that would be quantifying the value of the pawn, knight, bishop, rook, queen, and potentially king for the following standard grouping of players: Novice (<800), Beginner (800-1099), Intermediate (1100-1399), Intermediate 2 (1400-1699), Advanced (1700-1999), Expert (2000-2299). As noted before, these may need to be inflated because of the using of lichess data. Therefore, an example that may return is that a Beginner level player's Bishop is estimated at 2.5 while a player at a level of Intermediate 2 has a Bishop value of 3.7.

A side outcome that would be nice to create is an evaluation app/software/program that can do the same thing but for an individual player. Ideally, they would be able to input their own games and it would output an evaluation of their pieces. This would help students at different levels see what pieces they undervalue and how levels above them use them. Often many younger / less experienced players just watch grandmasters, where instead they could be watching and understanding the level above them and could be growing gradually.

Blind Spots, Potential Ethical Issues

Some blind spots from my analysis could start with generalizing ratings as a type of player. Obviously not every beginner will value pieces the same, and my desired outcome would create ratings for each piece based on this general rating.

The only key ethical concern to consider is the usage of chess games and data that the players may not 100% know they are part of. This goes through the agreement of the Lichess website and the CC0 license, which should hopefully clear up any ethical issues there.