Assignment 1: DLS method

Subhojyoti Khastagir
SR. no.: 04-04-00-10-51-21-1-19446
M. Tech. - CSE
Department of Computer Science and Automation
subhojyotik@iisc.ac.in

1 Introduction

The Duckworth-Lewis-Stern method is used to come up with a fair target for the chasing team in an ODI match which has been interrupted in between due to which one or both of the two teams have lost some overs. It makes use of historical data and assumes a non-linear model to estimate the average runs achievable.

In this assignment, the DLS method was used to calculate the achievable runs for a given value of wickets in hand and overs from a given historical data of all ODI matches between 1999 to 2011.

2 Data

There are usually 50 overs per team in an ODI match which means 100 overs per match. But due to some matches which were interrupted, not all the matches had a total of 100 overs. Out of the 1423 matches provided in the csv data file, 432 matches had an incomplete first innings: which means that 432 matches had less than 50 overs with wickets still remaining in hand for the first innings, and 1177 incomplete second innings. From this data it is clearly evident that to be able to use data of uninterrupted matches, only the first innings' data can be used, as was instructed for the assignment.

The provided data file had lots of columns including "Innings", "Overs", "Run.Rate", "Day-night", "Home.Team", "Away.Team", etc., most of which were not required. Only the "Match", "Innings", "Over", "Runs", "Runs.Remaining" and "Wickets.in.Hand" columns were used for the calculation.

Luckily the data didn't have any spurious data like missing or non-sense entries. The "Date" column wasn't formatted consistently, but that was not an issue since the dates were not required for the calculation.

3 Methodology: Trial and error

The given model for the calculation of average runs was

$$Z(u, w) = Z_0(w)(1 - \exp(-Lu/Z_0(w)))$$

Here, the parameters are L, $Z_0(i); i \in [1, 10]$, which makes a total of 11 parameters. The data points for fitting the curve was calculated by taking the mean of "Runs.Remaining" for every over and wickets remaining. Using standard python library implementation of scipy.optimize.minimize, the parameters were optimized to reduce the loss function which was simple squared error loss.

With an initial guess of $L = 10, Z_0(i) = 100; i \in [1, 10]$, the squared error loss was 514.93 and the optimized values for the 11 parameters turned out to be

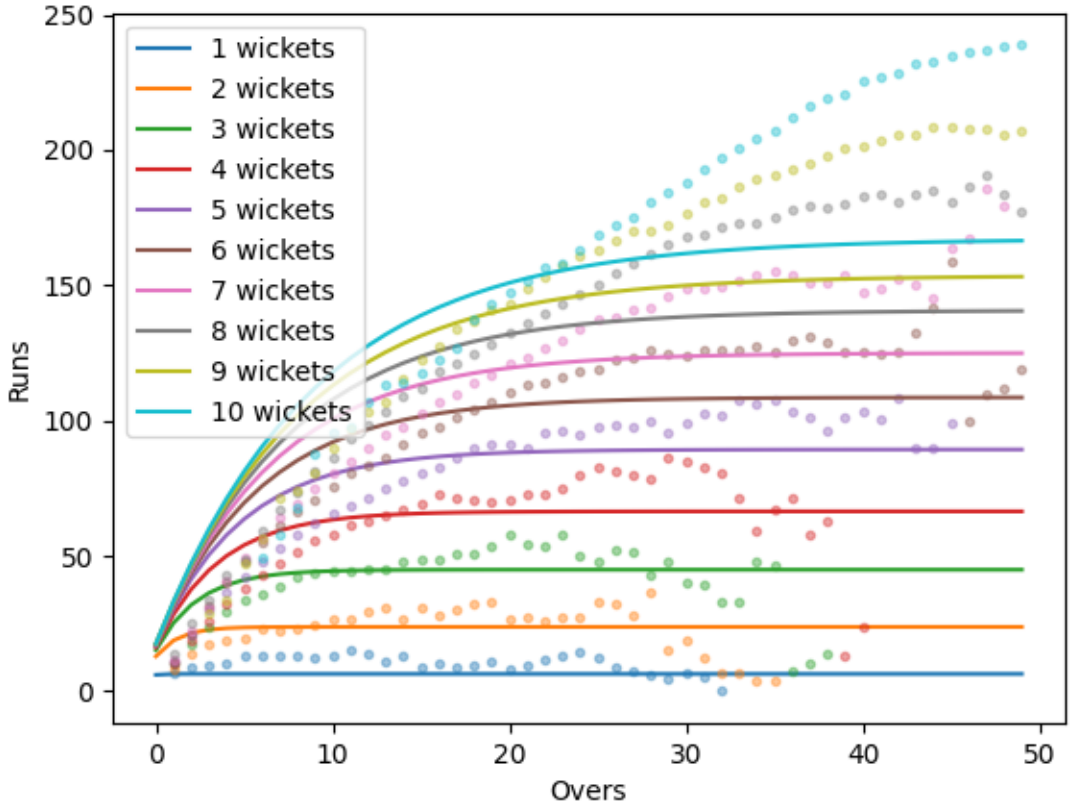

Figure 1: Initial attempt with a bad guess

L	$Z_0(1)$	$Z_0(2)$	$Z_0(3)$	$Z_0(4)$	$Z_0(5)$	$Z_0(6)$	$Z_0(7)$	$Z_0(8)$	$Z_0(9)$	$Z_0(10)$
18.56	6.38	23.75	44.9	66.46	89.32	108.58	125.0	140.73	153.56	167.14

The corresponding plot is shown in Figure 1:

The curves do not fit the data points very well. So this means that the initial guesses were too bad. After taking a look at the values that emerged after minimizing, the initial guess can be improved.

After revising the initial guesses to $L = 20, Z_0(i) = 20 \times i; i \in [0, 9]$, the squared error loss was 52.21 which was clearly an improvement, and the optimized parameters were

L	$Z_0(1)$	$Z_0(2)$	$Z_0(3)$	$Z_0(4)$	$Z_0(5)$	$Z_0(6)$	$Z_0(7)$	$Z_0(8)$	$Z_0(9)$	$Z_0(10)$
10.94	10.24	23.96	44.28	71.34	103.09	133.07	170.02	204.66	238.164	277.95

The corresponding plot is shown in Figure 2:

The plot now matches the data points pretty closely. But the process involves making guesses and improving upon the initial values of parameters. This cannot be done for larger datasets and the process needs to be automated. Here comes the next methodology.

4 Methodology: Heuristics

Going by the definition of the parameters $Z_0(w)$, we know that it denotes the runs that can be scored with w wickets in hand given unlimited overs. Other than guessing a random value or literally starting with a random value, this value can be obtained from the data itself, by calculating the cumulative runs scored for each value of wickets remaining, for uninterrupted matches, where all wickets were lost. The mean of this cumulative runs scored, taken over all such matches is used as a heuristic value

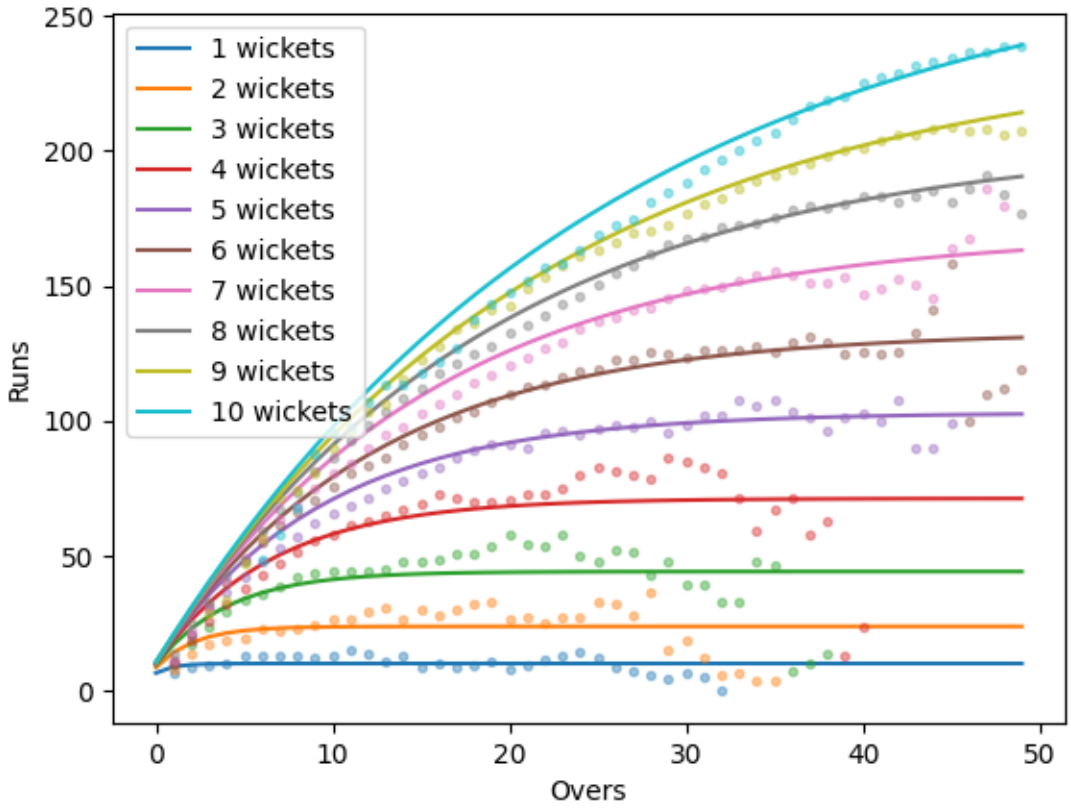

Figure 2: Next attempt with improved guess

for initial values of the parameters $Z_0(w)$. The value of L was kept at 10, as was obtained from the previous methodology.

Using this method, the loss turned out to be equal to what the previous attempt yielded, that is 52.21, and the optimal values for the parameters were

L	$Z_0(1)$	$Z_0(2)$	$Z_0(3)$	$Z_0(4)$	$Z_0(5)$	$Z_0(6)$	$Z_0(7)$	$Z_0(8)$	$Z_0(9)$	$Z_0(10)$
10.94	10.24	23.96	44.28	71.34	103.09	133.07	170.02	204.66	238.164	277.95

which is once again the same as the previous attempt.

The plot, which also looks similar, is shown in Figure 3

5 Conclusion

The provided data was clean enough for the process to be an easy and smooth one. Not much preprocessing was required to extract and perform calculation on the data and the end result was very satisfactory, as can be confirmed from the close fitting curves in the plots. The final values for the 11 parameters are

L	$Z_0(1)$	$Z_0(2)$	$Z_0(3)$	$Z_0(4)$	$Z_0(5)$	$Z_0(6)$	$Z_0(7)$	$Z_0(8)$	$Z_0(9)$	$Z_0(10)$
10.94	10.24	23.96	44.28	71.34	103.09	133.07	170.02	204.66	238.164	277.95

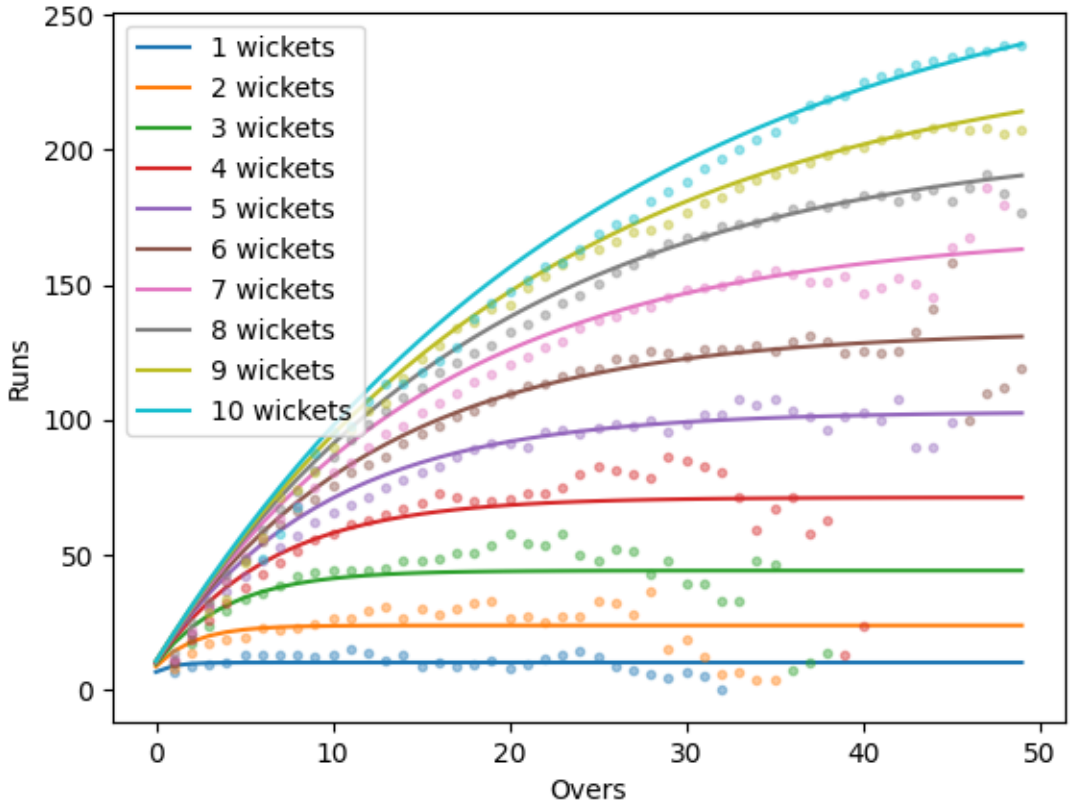

Figure 3: Using heuristic instead of guess