Data Analytics: Assignment 1

BALAJI CHUNDI
SR NO:- 19728

August 22, 2022

1 Duckworth-Lewis-Stern Method

Frank Duckworth and Tony Lewis introduced a system based on "resources", which are the number of wickets and overs the batting team has left in its innings. The idea is that regardless of the total number of total overs in an innings, two teams that have the same number of these two resources are fairly matched, as they would be at the start of the game, before any runs have been scored.

2 Problem Statement:

Using the data of the first innings alone from the dataset *04_cricket_1999to2011.csv*
we need to find the best fit 'run production functions' in terms of w(wickets-in-hand) and u(overs-to-go). The model is assumed as

$$Z(u, w) = Z_0(w)[1 - exp{-Lu/Z_0(w)}]\,. \tag{1}$$

Loss function is "normalized sum of squared errors" across all wickets and overs.

3 Data Preprocessing

Selected only the data points corresponding to first innings of a match.

As oversRemaining is ranging from 0-49 only, I have added data points corresponding to 50 remaining overs.

Dropped data points which specify 0 wickets in hand.

Dropped data points corresponding to 0 oversRemaining.

Data rows having reported erroneous data in the column 'Error.In.Data' are dropped.

4 My Approach:

I have used two ways to initialize the parameters, first is 'random initialization' and second is considering the given data, I have calculated the mean runs corresponding to each wickets and used these as initial parameters.

I have extracted the important features from all the given columns, they are 'Innings.Total.Runs', 'Total.Runs', 'Total.Overs', 'Over' and 'Wickets.in.Hand'.

For first innings alone(across all the matches), for each data point(row) I have calculated 'runsRemaining' and 'oversRemaining' as follows:

$$runsRemaining = data[Innings.Total.Runs] - data[Total.Runs]$$
$$\text{oversRemaining = data[Total.Overs] - data[Over]} \tag{2}$$

Two extracted features('oversRemaining', 'runsRemaining' and a direct feature ('Wickets.in.Hand') are considered.
Next, I have defined the loss function, which is 'sum of squared errors' across all data points summed across overs and wickets.

$$\min_{Z_0(1),Z_0(2),\ldots,Z_0(10),L} \Sigma_{n=1}^{N}(y_n - Z(u_n, w_n, Z, L))^2 \tag{3}$$

Where N = total number of all the first innings data-points, $Z(u_n, w_n, Z, L)$ is the predicted run and y_n is the original runs.
I have used library *scipy.optimize* to minimize the function defined above:
scipy.optimize.minimize

I have tried different methods like: **L-BFGS-B, SLSQP, CG, BFGS**.
Out of all these methods, L-BFGS-B and SLSQP resulted in successful termination of the optimization. The values I got for the parameters by both these methods under random initialization and Mean initialization are listed below.

	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8	Z_9	Z_{10}	L
Initial	10	30	40	65	85	130	155	170	185	200	10
L-BFGS-B	13.589	27.354	51.444	79.226	104.428	138.006	168.659	207.183	239.402	284.395	10.853
SLSQP	13.582	27.340	51.45	79.221	104.417	138.007	168.629	207.150	239.378	284.373	10.856

Table 1: Optimized parameters by random initialization

	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8	Z_9	Z_{10}	L
Initial	8.848	18.547	33.619	50.820	70.415	95.602	123.329	156.273	188.387	226.717	10
L-BFGS-B	13.583	27.359	51.462	79.200	104.434	138.016	168.643	207.168	239.407	284.406	10.854
SLSQP	13.589	27.324	51.438	79.207	104.426	138.018	168.631	207.155	239.390	284.383	10.855

Table 2: Optimized parameters by Mean initialization Both the values are very close to each other, even the total normalized squared error loss for both the methods is around **1531.781** for both initialization.

Below are the plots of the above performed analysis.

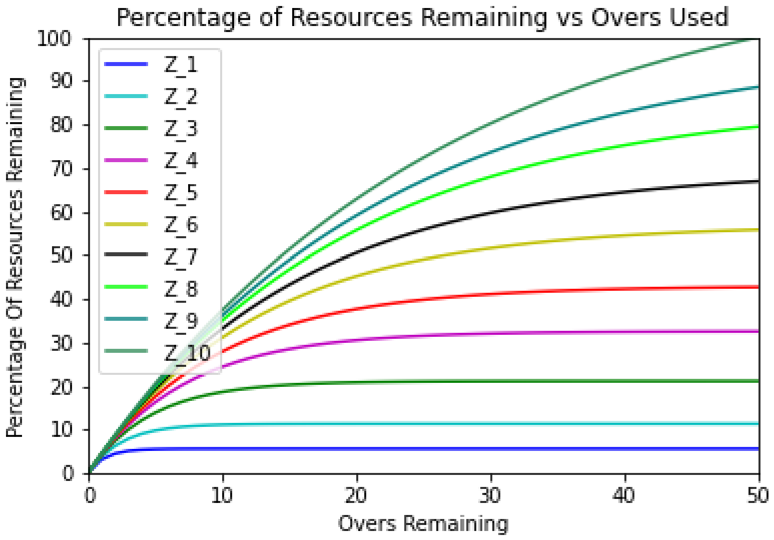

Figure 1: Plot of resources remaining vs Overs remaining, Method L-BFGS-B

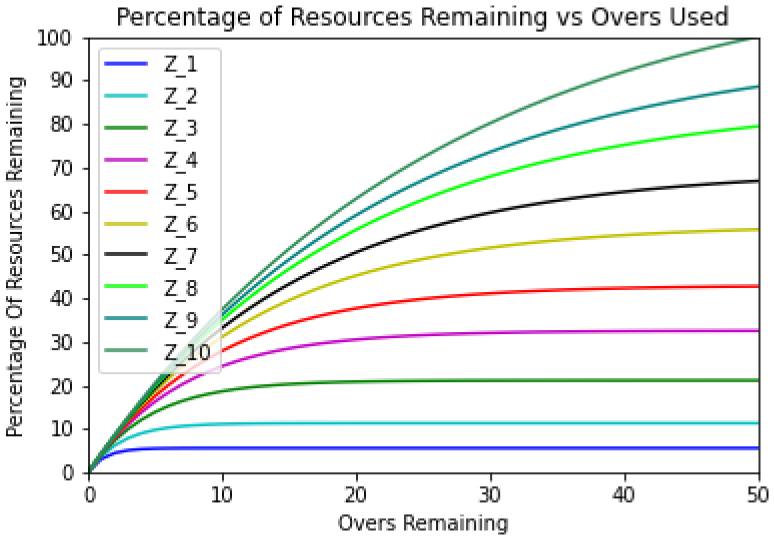

Figure 2: Plot of resources remaining vs Overs remaining, Method SLSQP

Additionally, I have tried to consider the matches which have been conducted for a total of 50 overs each side, that is, matches that got interrupted in between by some disturbance has been removed. The total normalized squared error loss obtained then was **1360.044** and the corresponding parameters from Z(1-10) and L are [14.32942637, 29.99915818, 57.92842286, 91.28158893, 117.23266168, 154.09697862, 184.59857237, 229.86324639, 261.19844699, 305.45643363, 10.39028225].

The resource remaining vs overs remaining plot for this instance under L-BFGS-B is shown below:

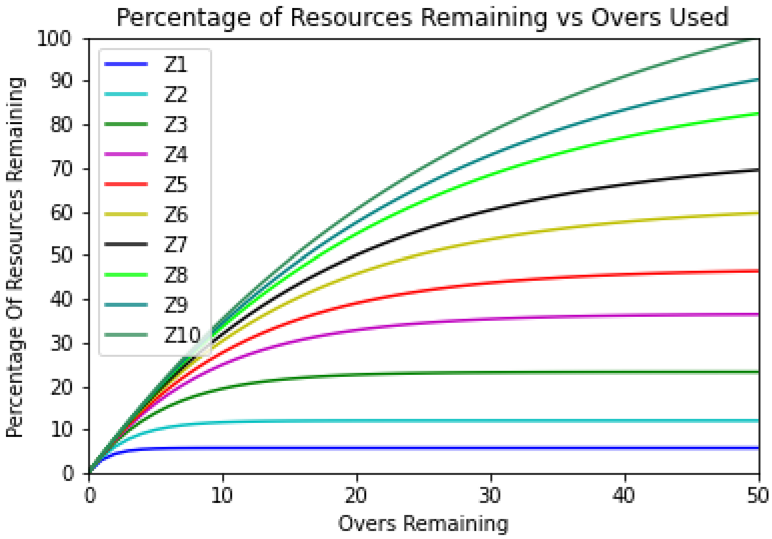

Figure 3: Plot of resources remaining vs Overs remaining, Method L-BFGS-B(for matches held for 50 overs per each side)

5 Note

1) Parameter initialization can be changed in the 'Minimize' function.
2) The part which considers matches held for 50 overs per each side is commented out in the final submission, in 'preprocessing' function.