Data Analytics : Assignment 1

Duckworth-Lewis-Stern Method

Anuj Kumar
SR No : 21462
Department of Computer Science and Automation
email: *anujkumar1@iisc.ac.in*

August 21, 2023

1 Implementation Summary

1.1 Task Explanation

Using the first innings data alone in given data set and preprocessed the data, find the best fit 'run production functions' in terms of wickets-in-hand w and overs-to-go u. Given Model is

$$Z(u, w) = Z_0(w)\left[1 - \exp\left(-\frac{Lu}{Z_0(w)}\right)\right]$$

The error function asked to use is mean squared error function, summed across overs and wickets.And once we fit the model, we are asked to produce a plot of the ten functions, and report the (11) parameters associated with the (10) production functions which are (Z0[1],Z0[2],..,Z0[10],L) and the total error as normalized squared error.

1.2 Implementation of Task

1. First we got the data from the data file 04_cricket_1999to2011.csv. And In data processing step I fetched important columns from all the columns in the file. Here we kept the data for first innings only. and over are given as played so for remaining overs we subtracted over from 50 which gives remaining overs. finally processed data contains: Over, Runs.Remaining, Wickets.in.hand, Overs.Remaining.

2. Next I take all the data points of first innings and define the loss function as sum of squared errors loss function, summed across overs and wickets.Formally we can write loss function minimization as:

$$\underset{Z_0(1),Z_0(2),\ldots,Z_0(10),L}{\textbf{minimize}} \sum_{i=1}^{N}(y_i - Z(u_i, w_i, Z, L))^2$$

   Here N=All the data points of first innings, y_i a data point of runs remaining (actual runs), w_i a data point of wickets in hand and u_i a data point of overs remaining. And Z0 and L are the parameters.

3. $Z_0(w)$: shows expected runs can we made with w wickets in hand with no bound on overs. which initialized using averaging the runs remaining for given wicket = w from processed data. And L (slope) is initialized to $b(10) * Z_0(10)$, here $b(10) = 0.035$ and $Z_0(10)$ = average of runs remaining when 10 wickets are in hand.

4. Then I used scipy.optimize.minimize library function to minimize the objective function defined above in second step with method L-BFGS-B. which updates the parameters and loss.

2 Results

2.1 The plot with 10 curves

Here are plots in between Overs remaining and Average runs obtainable for every value of wickets. x-axis contains overs remaining (u) and y-axis value is denoted by Z(u,w) which is predicted runs.

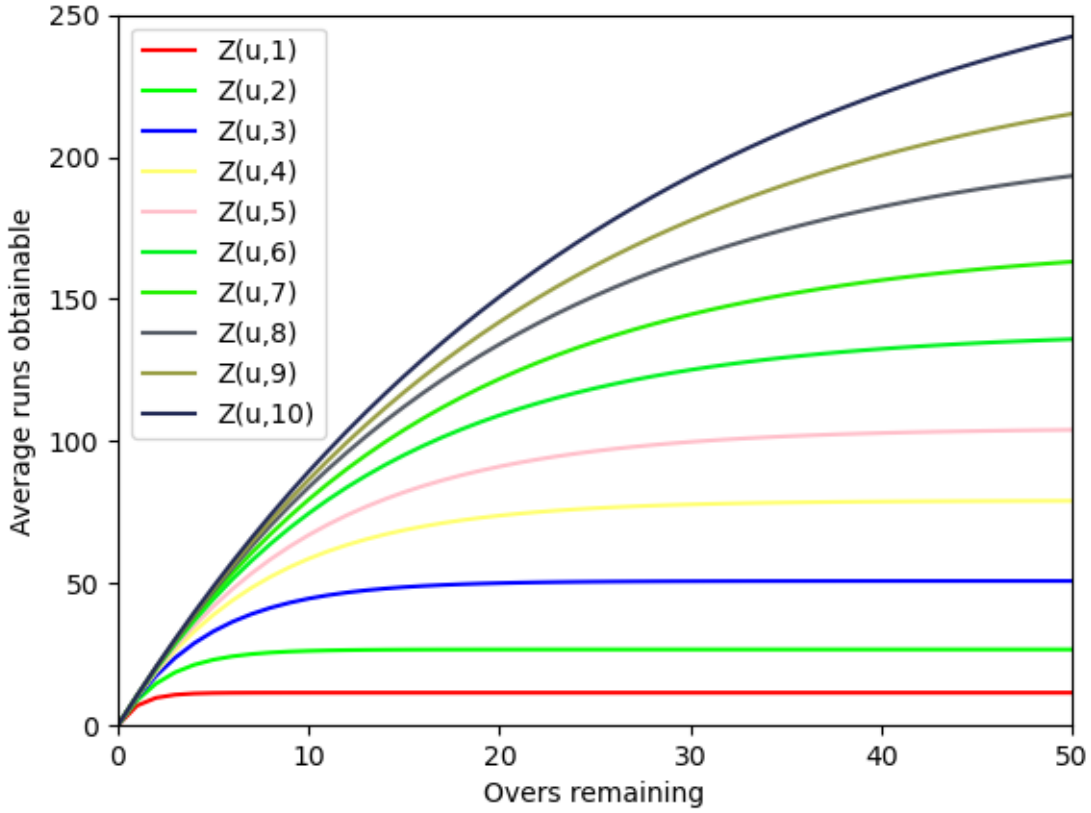

Figure 1: Average runs obtainable vs Overs Remaining

2.2 Average Loss

Average Loss is defined as Total squared error summed across overs, wickets, and data points for those overs and wickets, and normalised by the total number of data points across all overs and wickets. Normalized Squared Loss given by Model is **0.25573607320078784**

2.3 Value of model Parameters

Values of model parameters after training are given below :

1. Z0 Parameters (Z0(1),Z0(2),...,Z0(10)) values are : [11.52197614, 26.73159873, 50.82868327, 79.1134318, 104.61460773, 138.81851718, 170.50233284, 209.65479513, 241.50683739, 286.66483513]
2. L Parameter value is : 10.699681483308007

```
Data loaded.
Data preprocessed.
Z0 Parameters (Z0(1),Z0(2),...,Z0(10)) values are :
 [ 11.52197614  26.73159873  50.82868327  79.1134318  104.61460773
 138.81851718 170.50233284 209.65479513 241.50683739 286.66483513]
L Parameter value is :  10.699681483308007
Mean Squared Loss is :  1569.2562446439417
Normalized Squared Loss is  :  0.25573607320078784
```

Figure 2: Model Parameters and Loss given by model