Assignment-1

Y Rithvik

August 21, 2023

1 Implementation Summary

1.1 Data Preprocessing

Implemented the following in *preprocess_data* function

- **Date Correction:** Corrected the wrong date format by converting the date in DD/MM/YYYY and MM DD-DD, YYYY format to DD-MM-YYYY.
- **Removed useless columns:**Removed useless columns keeping columns which are used for training the model and calculating loss.
- **Removed all entries of 2nd Innings:** Removed all entries for which Innings is 2.
- **Add new entry for each match:** Added an entry for each 1st innings which contains runs remaining for 50 overs left and 10 wickets remaining.

1.2 Train Model

Training procedure: During training, non-linear regression is performed using **scipy.optimize.minimize** which computes the optimized parameters(Z_0 values and L value) by minimizing the normalized squared error over all data points. This is mathematically described in eq(1).
Initialization of parameters: Parameters $Z_0(w)$ is initialized by taking average runs scored with wickets in hand = w across all data points present and parameter L is initialized by the average runs scored in the 50th over. These initialized parameters are passed as arguments to scipy.optimize.minimize function.

$$Z_0^*(1), \ldots, Z_0^*(10), L^* = \arg \min_{Z_0(1),\ldots,Z_0(10),L} \frac{1}{\text{tl}} \sum_{w=1}^{10} \sum_{u=0}^{50} L_{squared}(u,w) \tag{1}$$

$$where$$

$$L_{squared}(u,w) = \sum_{\substack{y_{\text{true}} \in \{\text{df['Runs.Remaining']} \mid \\ \text{df[Overs.Remaining]}=u \,\&\, \text{df['Wicket.In.Hand']}=w\}}} (y_{\text{true}} - Z(u,w))^2$$

$$Z(u,w) = Z_0(w)\left[1 - \exp\left(\frac{-Lu}{Z_0(w)}\right)\right] \tag{2}$$

$$\begin{aligned} u &= \text{Overs Remaining} \\ w &= \text{Wickets In Hand} \\ df &= \text{preprocessed data} \\ tl &= \text{total data points} \end{aligned}$$

Model parameters are finally set to the obtained optimized values.

1.3 Plots and Loss

Plots: Plotted 10 graphs with the number of overs remaining(u) on x-axis ranging from $0 - 50$ and $Z(u, w = i)$ for i in $\{1, 2, \ldots, 10\}$ on y-axis, each graph corresponds to different i value.
Loss: Normalized squared error loss is computed which is nothing but the optimization function in eq(1)

2 Results

2.1 The plot with 10 curves

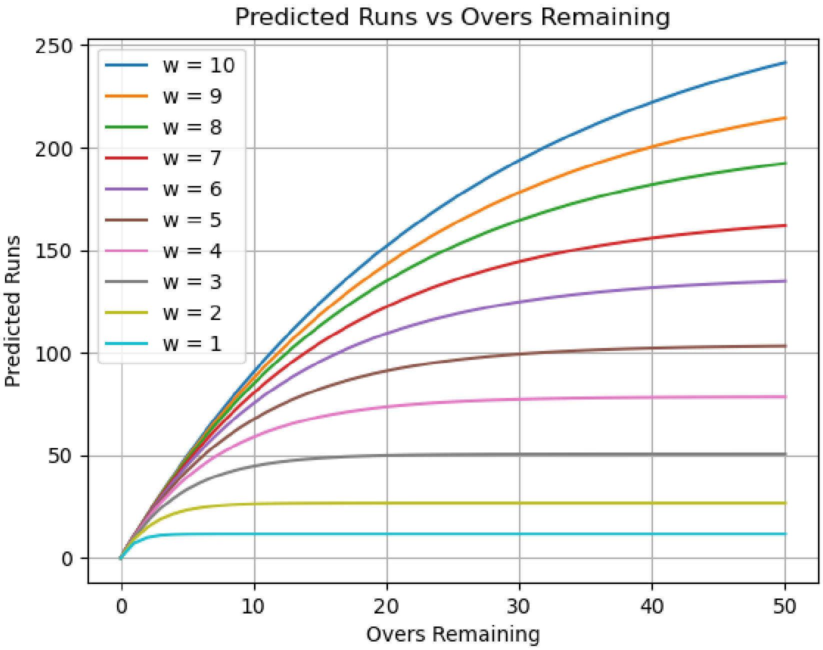

Figure 1: Plot with 10 curves

2.2 Average Loss

Normalized Squared Error Loss over all data points is: **1609.5452968525506**

2.3 Values of Model Parameters

Parameter	**Value**
$\mathbf{Z_0}$**(1)**	**11.663168681996252**
$\mathbf{Z_0}$**(2)**	**26.79481207621622**
$\mathbf{Z_0}$**(3)**	**50.58490378153681**
$\mathbf{Z_0}$**(4)**	**78.50011449659158**
$\mathbf{Z_0}$**(5)**	**103.82277082459568**
$\mathbf{Z_0}$**(6)**	**137.45181574885703**
$\mathbf{Z_0}$**(7)**	**168.57036617190334**
$\mathbf{Z_0}$**(8)**	**207.2123051296936**
$\mathbf{Z_0}$**(9)**	**238.72706883783005**
$\mathbf{Z_0}$**(10)**	**282.26586466973885**
L	**10.91456538467796**