COL 341

Report – Logistic Regression

While I analyzed the data I found that length of stay was linearly proportional to the total costs which it should be. So, I try to create features that are relevant to total costs. I created one hot encoding for them and then multiply that vector with total costs column so that I could get data which is linearly proportional to total costs. Here, since costs were numbers very less than 1, so I have taken log of total costs column and since, log is increasing function, so total costs while taking log would be linearly proportional to length of stay. Also, I clipped the total costs values between 10^{-15} and $1-10^{-15}$ so that there is no problem while taking log.

I have done normalization of birth weight and Patient Disposition columns as they contain large values.

I have removed gender and race columns as they were not related to length of stay.

I have removed Facility Id column as using this was not giving good result.

I have done one hot encoding of all the features and multiply them by updated total costs column as all features were related to total costs.

I have taken top 20 most frequent values from CCS Procedure Code, CCS Diagnosis Code, APR DRG Code as they have many distinct values.

I have taken top 10 most frequent values from Zip Code 3-digits, Hospital County and Facility Name as they have many distinct values.

For Operating Certificate Number, took top 150 most frequent features.

I merged Health Service Area and Emergency Department Indicator. I thought that emergency department indicator depends on health service area. For each merging, the length of stay depends on costs. So, did one hot encoding of merge columns of health service area and emergency department indicator with total costs.

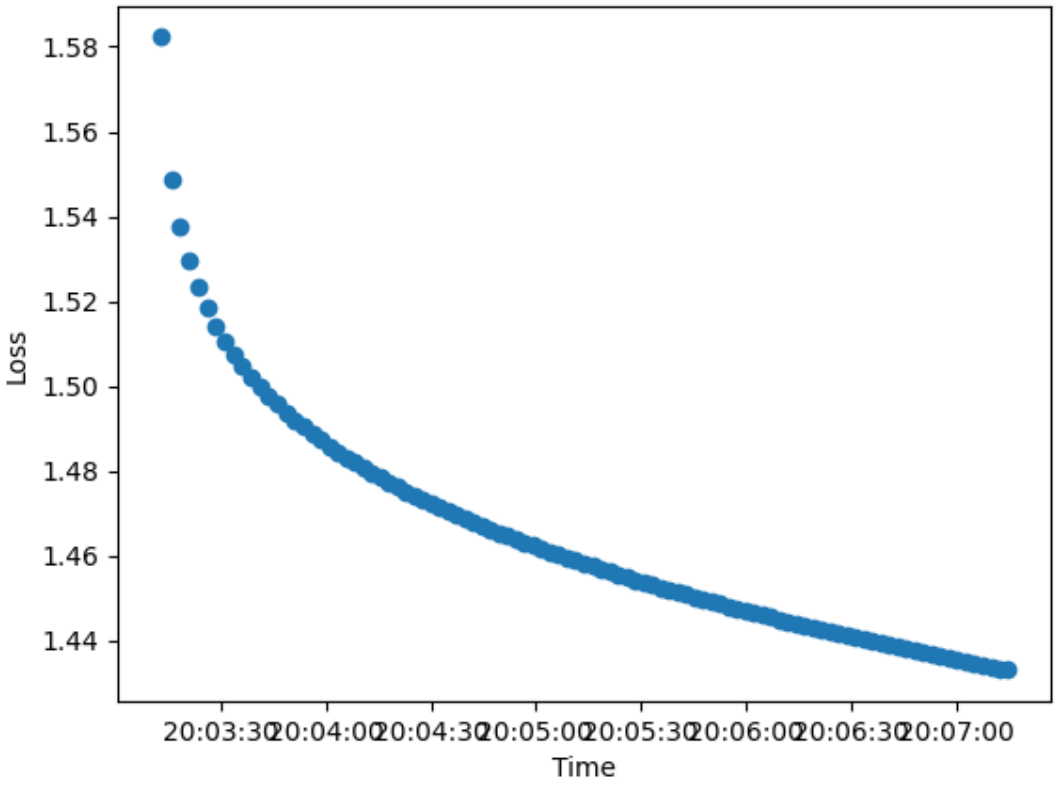
I merged Age Group and Emergency Department Indicator. I thought that older people will be taken to emergency more often as compared to young ones. For each merging, the length of stay depends on total costs. So, did one hot encoding of merge columns of age group and emergency department indicator with total costs.

I used alpha-beta backtracking using mini batch gradient descent.

For part c, I tried first batch gradient descent using fixed step size. I plotted graph of loss function verses time varying initial step size. Plotted graph for initial step size = [0.001,0.01,0.1,1, 2,2.5,3,5]. I found that taking initial step size = 2.5 was giving good results. Then I fixed initial step size = 2.5 and perform mini batch gradient descent varying batch sizes. I plotted graph of loss function for batch sizes = [100,300,500,750,1000] and found that batch size = 500 was giving good results. Then, I fixed batch size = 500 and initial learning rate = 2.5 and

perform alpha-beta mini batch gradient descent varying alpha and beta. I plotted graph for alpha = [0.3,0.35,0.4,0.45,0.5] and beta = [0.7,0.75,0.8,0.85,0.9,0.95] and found that taking alpha = 0.45 and beta = 0.75 was giving good results.

So, I took initial learning rate = 2.5, batch size = 500, alpha = 0.45 and beta = 0.75 for part c and d.



Graph of Loss v/s time taking my hyper parameters

Name : Raunak Jain

Entry Number : 2019MT10719