Sri Hari Malla
Indian Institute of Technology Hyderabad
Fundamentals of Machine Learning
CS19BTECH11039

PROGRAMMING REPORT

Sri Hari Malla - CS19BTECH11039

November 18, 2021

1 Logistic Regression:

1.1 Implementing Code:

- The code section can be found in the notebook submitted.
- It is Trained using Gradient Descent and Cross entropy as loss function.
- The code sections for 5 b were in subsequent cells.
- For better visualisation, some plots were drawn

1.2 Exploring Gradient Descent:

1.2.1 Cross Entropy Error Function:

Cost Function:

$$\text{j(w)}=\sum_i \left(\log\left(p\left(\bar{y}_i = 1 \mid \bar{x}_i\right)\right) + \left(1 - y_i\right)\log\left(p\left(\bar{y}_i = 0 \mid \bar{x}_i\right)\right)\right)$$

Cross Entropy function:

$$p(\bar{y} = 1|x_1x_2) = \frac{1}{1 + e^{-(-1+1.5x_1+0.5x_2i)}}$$

$$p(\bar{y} = 0|x_1x_2) = 1 - \frac{1}{1 + e^{-(-1+1.5x_1+0.5x_2i)}}$$

1.2.2 After One iteration:

- Can clearly see the output on the notebook submitted.
- Initial Weights : 1.5 0.5
- Initial Bias : -1

- Final Weights After One iteration : 1.50535086, 0.50196867
- Final Bias : -1.0031662597725644

  Cross Entropy function:

$$p(\bar{y} = 1|x_1x_2) = \frac{1}{1 + e^{-(-1.0031662597725644+1.50535086x_1+0.50196867x_2i)}}$$

$$p(\bar{y} = 0|x_1x_2) = 1 - \frac{1}{1 + e^{-(-1.0031662597725644+1.50535086x_1+0.50196867x_2i)}}$$

1.2.3 At Convergence:

- Used 500000 iterations.
- The plot function of cost in the notebook clearly suggests that it is converging at 5*1e5 iterations.
- Final Weights : 42.85263545, 9.55973708
- Final Bias : -28.346038607109172
- Accuracy Observed : 66.667
- Precision observed :
- Recall observed :

2 Taxi Fare - Predictions

- Code can be found in the notebook submitted.
- The Train dataset is too heavy.
- Need to pre process the data before training the model.

2.1 Pre Processing:

- The train data set is very very large with number of rows in millions.
- So, instead of picking the entire data, I picked the first 7 lakh rows of Train data.
- The data is considerably low with Nan values. Specially, no method was used to update the Nan values.
- Just dropped the rows having entire Nan values.

- Basic examination of the data suggests small changes in the train data.
- Dropped some rows with respect to the column attribute 'Passenger count'
- Dropped some points with respect to the dropoff and pickup latitude and longitude.
- Later found that, without preprocessing itself, it is giving more accuracy.
- Two special columns are entirely having string values. Used label encoding to encode the values.
- Seperated features and fares for using model on regression.

2.2 Support Method:

- Written a method to create a submission file with given predictions.
- Name argument is supported, if not sent, it creates a file with default name 'submission.csv'.

2.3 Model Selection:

2.3.1 Models used:

- Linear Regression Model
- XG Boost Regression Model
- Cat Boosting Regressor
- Light Boosting Regressor
- **Cat Boost Regressor and Light Boost Regressor outperformed XG boost and Linear regression models.**

2.3.2 Cat Boosting Regressor:

- Cat Boosting Regressor is found to be the most innovative algorithm for processing data having categorical features.
- For small datasets, GBC overfits easily whilst in CBC it is handled all by its own. Overfitting problems are relatively low in CBC when compared to Other algorithms.
- As the noise is low, it is expected to perform better.
- It generally performs better than other regression algorithms.
- The RMSE score measured was **3.998**

2.3.3 Light Boost Regressor:

- The Light Boost Regressor is same sort of CBR.
- Light GBM is a fast, distributed, high-performance gradient boosting framework based on decision tree algorithm, used for classification and regression.
- Light GBM use histogram based algorithm i.e it buckets continuous feature values into discrete bins which fasten the training procedure.
- It produces much more complex trees by following leaf wise split approach rather than a level-wise approach which is the main factor in achieving higher accuracy.
- The RMSE score measured was **3.79**

LaTeX generated document

THE END