भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Department of Computer Science and Engineering
Indian Institute of Technology Hyderabad

Report on Assignment 3

Dependent cost-sensitive regression

submitted by

Akshat Gupta – CS23MTECH11001
Ashish Emmenuel – CS23MTECH11004
Ashutosh Rajput – CS23MTECH11005
Simanta Das – CS23MTECH11018
Patnala Sai Kumar – CS23MTECH14020

CONTENTS

1. Abstract

Logistic regression one of a supervised learning method that is robust and reliable in binary classification. The Loss function it uses is Binary Crossentropy loss, which gives equal weightage for True Positive, True Negative, False positive and False Negative rates. But in real world problems it is not this case.

Here in this Project we are looking at two such approaches where we define a custom loss function in which there is different types of penalties for different misclassifications. The first one is **Bahnsen approach** and the second is **Gunnemann's approach**.

2. Introduction

2.1 Problem Statement

First let us understand what logistic regression is: It is used in classification of data in a binary way and it is a supervised machine learning algorithm. So basically I classify the data into two different sets only.

The problem with logistic regression is that it treats false positives and false negatives as equal as it uses a threshold to classify instances. So if the predicted probabilities are greater than the threshold then it is treated as a positive value else a negative value.

Then here comes Example Cost-sensitive regression : In this regression modeling the costs associated with misclassification varies between examples.It introduces the example-dependent costs into a logistic regression. In this misprediction costs of instances are different so it obtains the prediction result such that it leads to overall less cost. It is hard to implement as we have to explicitly define costs while training the model. It uses techniques such as data resampling, modification of algorithms and ensemble methods.

2.2 Description of Data set

We have been given a costsensitiveregression.csv file. It includes 13 columns

- Notcount
- Yescount
- APTM
- PFD
- PFG
- SFD
- SFG
- WP
- WS
- AH
- AN
- Status
- FNC

Here column 1 - 11(Notcount - AN) are independent variable

Column 12(Status) is the dependent variable

Column 13 (FNC) is the false negative cost and based on the risk parameter details

3. Methodology

3.1 Preprocessing the data

The first step of any machine learning problem is looking at the data, to get insights from it. In many real world problem it is not always the case that the data is in a good condition.

For such cases it is important to pre-process the data, to fill missing data or eliminate outliers that can negatively effect the model.

3.2 Preprocessing steps

1. Create the X_test, X_train , y_test, y_train from the given csv file as needed.
2. From the given test and train data, create PyTorch dataloader with 32 batch size.

3.3 Logistic Regression Model

1. Create a Logistic Regression class
2. 1 input layer with dimension (input dim, 32), and Sigmoid activation
3. 2 input layer with dimension (32, 1), and Sigmoid activation

3.4 Loss function

Train the model using both the loss functions.

- **Bahnsen approach**
- **Gunnemann's approach**

4. Discussion

Given below are the different plots for the two variant of loss.

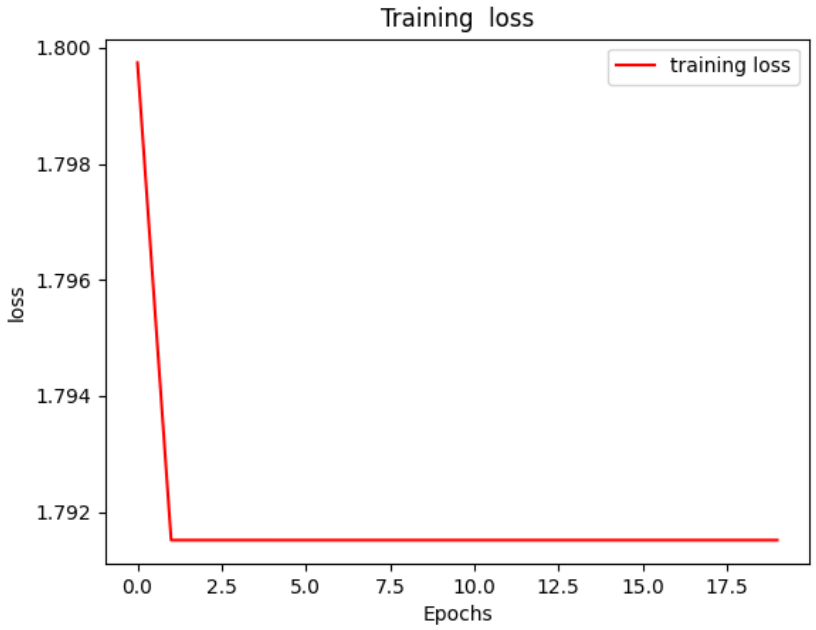

Figure 1: Training Loss for Bahnsen approach

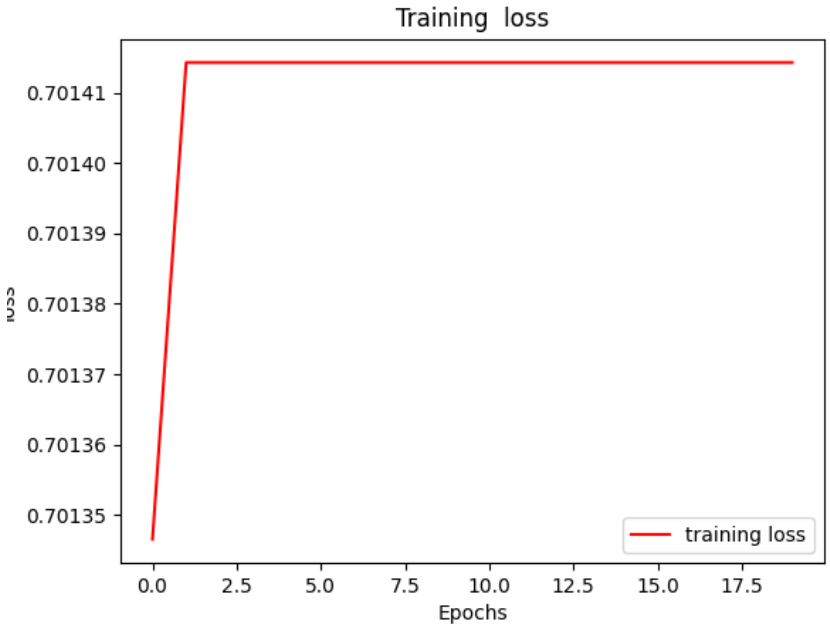

Figure 2: Training Loss for Gunnemann's approach

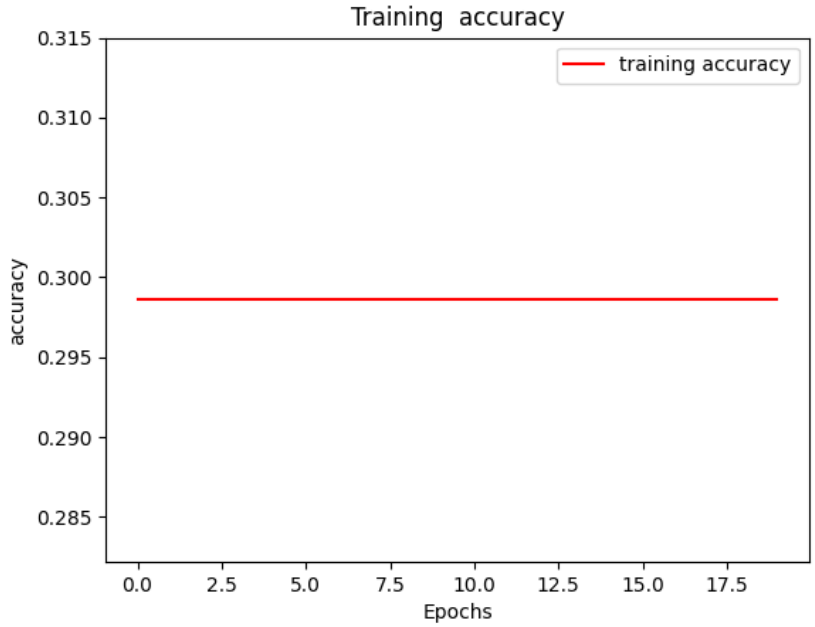

Figure 3: Traing accuracy for Bahnsen approach

From the graphs it is clear that the model is not very robust and doesn't perform well. It is because of the heavy penalty for True Positive rate.

5. Conclusion

After studying the two different approach we come to the conclusion that both model are not very robust. It is because of the heavy penalty for the True Positive rate. Because of the the true positive predictions are also getting penalised and causing the model to under-fit.

Each problem has its unique requirements and so thus the solution. To solve such kind of problems, it is important to study the data, and develope domain knowledge about the problem area.

6. References

1. Günnemann, Nikou, and Jürgen Pfeffer. "Cost matters: a new example-dependent cost-sensitive logistic regression model." *Advances in Knowledge Discovery and Data Mining: 21st Pacific-Asia Conference, PAKDD 2017, Jeju, South Korea, May 23-26, 2017, Proceedings, Part I 21.* Springer International Publishing, 2017.

2. Bahnsen, Alejandro Correa, Djamia Aouada, and Björn Ottersten. "Example-dependent cost-sensitive logistic regression for credit scoring." *2014 13th International conference on machine learning and applications. IEEE, 2014.*