A mathematical essay of logistic regression

Devansh Sanghvi
Department Of Biotechnology
Indian Institute of Technology, Madras
Chennai, India
be19b002@smail.iitm.ac.in

***Abstract*—This paper aims to develop an understanding of the mathematical properties of logistic regression and showcase its application in classifying whether a passenger aboard the Titanic survived or no.**

I. Introduction

Logistic Regression is a technique for modelling the likelihood of an event. It helps in understanding the relationship between the features and the target variables (survived or not in our case). The structure of logistic regression is very similar to the structure of linear regression. You have a set of explanatory variables (X1,X2....Xn) and our target binary variable (Y). The function behind it is more complicated than linear regression:

$$P(Y=1) = \frac{1}{1+e^{-(b+B*X)}}$$

P(Y=1) is the probability that the predicted class is 1, b is a constant that is not related to X and B represents the weights for the relationship between the X's and the Y. The logistic regression estimates the value of the constant, b and the weights using Maximum Likelihood Estimator. Upon calculating the weight and the the probability P(Y=1) for new data points and depending upon the threshold, decide if the predicted class should be 1 or 0 (survived/ not survived in our case). For example: if the threshold to classify the new data point is 0.5 and we get P(Y=1)= 0.75, we will classify that particular data point as 1.
To demonstrate the effectiveness of logistic regression, we analyze the ”Titanic” data-set provided to us. The data contains the passenger details of the 891 passengers that were aboard the Titanic on the eve of the infamous shipwreck. Passenger details included their name, sex, age, family details and their ticket details like their cabin, their passenger class, where they embarked from and their ticket prices. Another column in the data-set mentions if the passenger survived the Titanic crash or not. With the training data-set of the given passengers, we aim to predict if a given passenger would have survived the Titanic crash or not.
This paper will serve as a part of the requirements to be fulfilled for receiving credits in the course **EE4708: Data Analytics Laboratory**.

II. Logistic Regression

Logistic regression uses the maximum likelihood estimation to select the best fit line. It converts the probability into log(odds) using the logit function. To find the log(odds) value for each candidate, we project them onto the log odds value. Once we find the log (odds) value for each candidate, we'll convert the log(odds) to the probability using the sigmoid function.

$$log(\frac{p}{1-p}) = log(odds)$$

On finding the likelihood for the first line, we then rotate the log(odds) line by a bit and again calculate the likelihood. This is done using the Gradient Descent algorithm. For a particular slope, we calculate the cost of the classification prediction, and repeat the process until the best- fit log(odds) line is found. Certain assumptions are necessary for Logistic Regression:

A. Assumptions

- The dependent variable must be categorical.
- The model should have little or no multi-collinearity i.e. the independent variables should not be correlated with each other.
- The independent variables are linearly related to the log(odds).
- Logistic Regression requires quite large sample sizes.

B. Loss function in logistic regression

Mean squared error or mean squared errors cannot be used in logistic regression since the curve obtained from plotting the Mean squared error loss with respect to the logistic regression model weights is not a convex curve and therefore it is very difficult to find the global minimum. The non-convex nature of Mean Squared Error in logistic regression is because of the non collinearity introduced in the model because of the sigmoid function introduced in the model, making the relationship between the weight parameters and errors very complex. Hence the loss used by logistic regression (Binary Cross-Entropy Loss/ Log Loss):

$$\frac{-1}{N}\sum_{i=1}^{n} y_i \cdot log(p(y_i)) + (1-y_i) \cdot log(1-p(y_i))$$

C. Feature selection in logistic regression

Every raw data-set contains a lot of redundant features, that may impact the performance of the model. Feature selection is a component of feature engineering that involves the removal of irrelevant features and picks the best set of features for the model. In other words, it helps reduce the dimensionality of the data. There are various techniques used for feature selection, but one common method is the Recursive feature elimination:

It is used to select features by recursively considering smaller and smaller sets of features. We do this by eliminating the most useless features in our data in each iteration until we reach the desired amount of features.

D. *Review of model evaluation procedures*

To choose between different machine learning models, we need to evaluate the models using the right procedure. The simplest approach is to train and test on the same data, but this method results in overfitting the data. An alternative approach will be to split the data into train and test data. This is because testing accuracy is a better estimate of the model than the training accuracy. There are a couple of problems with train test split such as:

- It provides a high variance estimate
- Testing accuracy can change a lot depending on which data points are chosen in the test data-set.

While splitting the data into train/test data has its own shortcomings, they are still better than testing on the same data as training data and hence are used more often.

III. Preprocessing and exploratory analysis on the data

In this section, we will first preprocess the data to extract the important features in the data provided, then we will visualize the data to understand the effects of certain features and modify the data-set to optimize the results from the logistic regression model.

A. *Preprocessing the data*

The initial training data-set has 891 samples and 12 features/ passenger details, namely: PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin and Embarked. Similarly, the testing data-set has 418 samples. Preprocessing the data-set includes (Note that the changes are made in both the training and testing data-sets):

1) Missing value assessment: Some features have a percentage of missing values and the first step is to decide what to do for these features. Age, Embarked both have a small percentage of NA values (19.87 and 0.22 respectively). We replace the missing age values by the median age of the known ages, 28.00 and impute the missing Embarked values by the port on which the most people boarded, S. Cabin has 77.10 percent of its values missing, and hence it is appropriate to ignore the feature altogether.
2) Additional Variables: Both Parch and SibSp are related to if the passenger is travelling with a family or alone, and the two features can be combined into one feature: if he was travelling alone or not. For the subjective features: gender, passenger class and embarked, we create categorical features to represent them numerically.

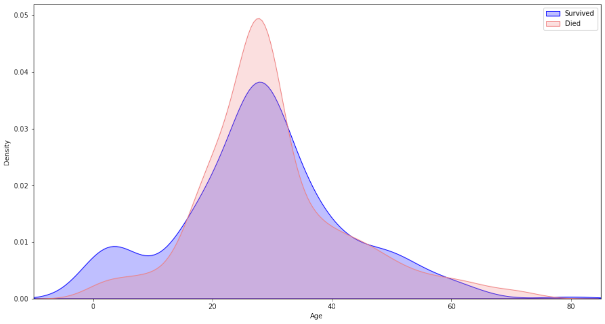

Fig. 1. Density Plot of Age for Surviving Population and Deceased Population

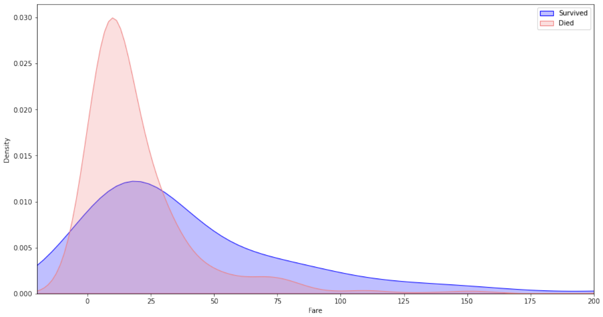

Fig. 2. Density Plot of Fare for Surviving Population and Deceased Population

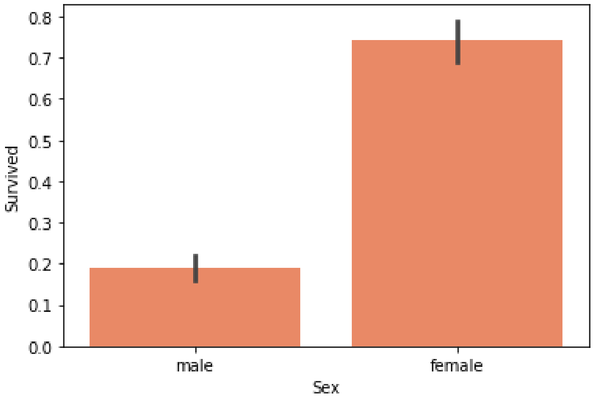

Fig. 3. Bar Plot of the fraction of the two gender populations that survived the crash

B. *Visualization*

After preprocessing the data, we visualize the data, to better understand the correlation between different features:

1) Exploration of Age: The two graphs show very similar distributions with respect to the age, but one noticeable difference is that a larger fraction of the survived population are children. It is therefore practical to add a feature to indicate if the passenger was a child or no.
2) Exploration of Fare: Passengers who paid less are less likely to survive and hence this could be an important indicator for our predictions.
3) Exploration of Gender: It can be clearly seen that

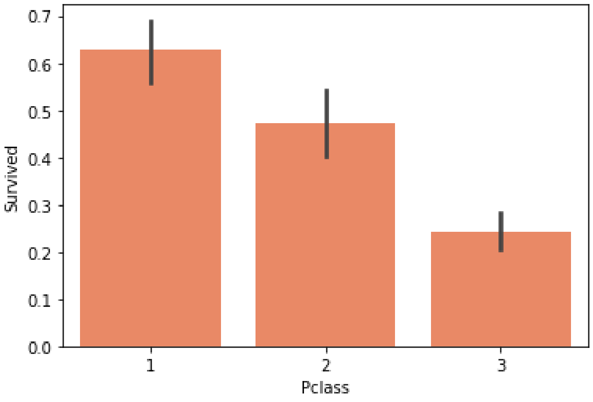

Fig. 4. Bar Plot of the fraction of passengers in each class that survived the crash

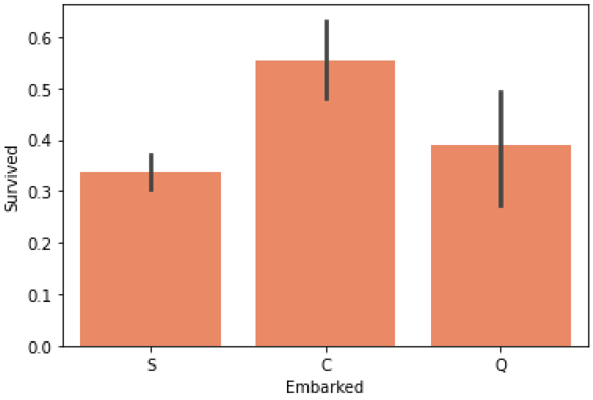

Fig. 5. Bar Plot comparing the fraction of passengers that survived with their respective ports they embarked from

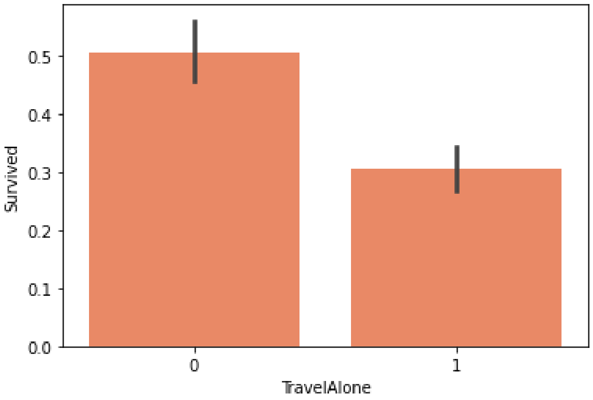

Fig. 6. Bar Plot of the fraction of passengers that survived while travelling alone or with family

females were preferred in the rescue procedure.

4) Exploration of Pclass: Fare and Pclass reiterate the same thing: the amount of money paid for a ticket is a strong indication of if the passenger survived. It is however ideal to ignore the Fare, as we will see in the feature selection process.
5) Exploration of Embarked: People who boarded the ship in Cherbourg, France, appear to have the highest survival rate, and people who boarded in Southampton were less likely to survive than the people who boarded in Queenstown.
6) Exploration of Travelling Alone: People without a family were more likely to die in the crash.

IV. The Logistic Regression Model

After preprocessing the data, we move to the heart of the problem: the prediction of whether an unseen passenger (not in the training data-set) survived the crash. Here are the steps for the same:

- Feature Selection: Before applying a logistic regression model to our data, we will eliminate the features that are least important using recursive feature elimination. The 8 remaining features are: Age,TravelAlone, Passenger classes, Embarked, Sex, IsMinor. Therefore, it is best to not include Fare in our model prediction. The correlation matrix for the same can be seen in image Fig.7.

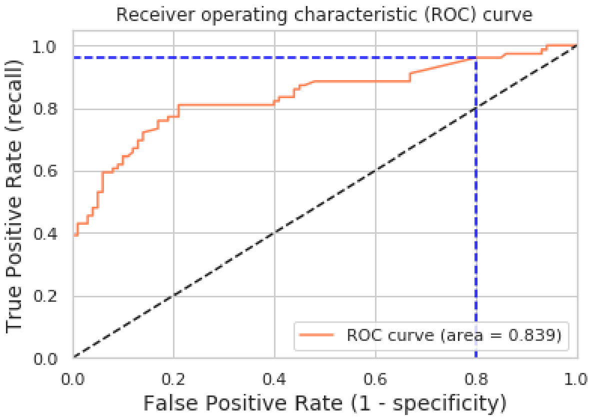

Fig. 7. Bar Plot of the fraction of the two gender populations that survived the crash

- Logistic Regression model and results After selecting the features, we run the logistic regression model using the simple test/ train model and evaluate our results using three parameters, which turn out to be the following:
  1) accuracy: 0.782
  2) logloss: 0504
  3) auc: 0.839

  Using the same model, and including Fare as a feature in our model we get the following results:
  1) accuracy: 0.796
  2) logloss: 0.455
  3) auc: 0.849

  We can see that the model is deteriorated and the feature Fare does not add valuable information to our model. The model can also be evaluated using K-fold cross validation and GridSearchCV which uses multiple scorers simultaneously.

V. Conclusion

The logistic regression model was succesfully demonstrated, bth visually and numerically, that there exists a correlation between gender and survival, and between the passenger class and survival. The addition of the L1 regularized term and the use of SAGA solver would result in a small increase in the accuracy of the model.

References

[1] ”Sinking of the Titanic”, wikipedia.org, 2022 [Online]. Available: **Link**.

[2] ”Logistic Regression”, scikit-learn.org, 2022 [Online]. Available: **Link**.

[3] ”What is logistic regression”, ibm.com, 2022 [Online]. Available: **Link**.