Assignment-2: A Mathematical essay on Logistic Regression

Ahmed Shmels Muhe
M.Tech In Data Science
Indian Institute of Technology Madras, IIT Madras
Chennai, India
ge22m009@smail.iitm.ac.in

***Abstract*—In this revised (v. 2) mathematical essay on logistic regression, we will look at the mathematics underpinning it. We also show how logistic methods were used in one of history's most infamous shipwrecks, the sinking of the Titanic. The assignment's goal is to analyse data for a total of 2224 passengers, 1502 of whom died due to a lack of lifeboats, and hypothesise whether certain groups of people (divided by name, age, gender, socioeconomic class, and so on) were more likely to survive, as well as discuss any insights and observations. In this revised version, I include the mathematics behind logistic regression in Section 2. In Section 2.1, I explained how maximum likelihood estimation worked for logistic regression and the mathematics behind it. In Section 4, I evaluated the model with a confusion matrix and ROC curve and updated the writing and some of the outputs after some modifications in the code.**

***Index Terms*—logistic regression,under-fit, exploratory analysis**

I. Introduction

The classification problem attempts to establish a relationship between a categorical target variable and one or more explanatory variables [5]. Logistic regression is a discriminatory statistical modelling technique that itself models the probability of outputs in terms of inputs [1]. This, in general, is well suited for describing and testing hypotheses about relationships. A binary classifier is constructed by choosing a threshold and classifying inputs with greater probability as one class and those below the cutoff as another class.

In this essay, we are given data on the Titanic disaster, and we must analyse which groups of passengers were more likely to survive and which groups of passengers died in the disaster. We are provided numerous attributes on which to base our predictions, but we must determine whether any of them can substantially impact the model's performance. In this article, we will deduce mathematical solutions to logistic regression models and use existing Python libraries to fit the logistic model to the provided data. We'll also talk about the data insights we've gleaned and why we opted to change or remove specific features.

The dataset for this problem comprises the passengers' names, titles, ages, and whether they travelled along with siblings or spouses, parents, or children. It also contains the ticket class, fare, port of embarkation, and cabin number. We use logistic regression to learn and predict passengers' survivability, given these input features.

This work represents the concepts behind Logistic Regression and the evaluation metrics involved. Using passenger data, we then use this technique to establish the relationship to predict the survivability of passengers.

II. Logistic Regression

Logistic regression is an extension of linear regression, and by making two adjustments, we may expand linear regression to the (binary) classification scenario. First, we substitute the Gaussian distribution for y with a Bernoulli distribution, which is better suited for binary responses, $y \epsilon 0, 1$. That is, we employ [1].

Logistic regression is a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set. A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables [2].

Logistic regression can also play a role in data preparation activities by allowing data sets to be put into specifically predefined buckets during the extract, transform, load (ETL) process in order to stage the information for analysis.

$$p(y|x, w) = Ber(y|\mu(x)) \tag{1}$$

where $\mu(x) = E[y|x] = p(y = 1|x)$. Second, we compute a linear combination of the inputs, as before, but then we pass this through a function that ensures $0 \leq \mu(x) \leq 1$ by defining

$$\mu(x) = \sigma w^T(x) \tag{2}$$

where $\sigma(\eta)$ refers to the sigmoid function, also known as the logistic function. This is defined as

$$\sigma(\eta) = \frac{exp(\eta)}{exp(\eta) + 1} \tag{3}$$

Putting these two steps together we get

$$p(y|x, w) = Ber(y|sigm(w^T x)) \tag{4}$$

This is called logistic regression due to its similarity to linear regression (although it is a form of classification).

A simple example of logistic regression is shown above a figure, where we plot

$$p(y_i = 1|x_i, w) = \sigma(w_0 + w_1 x_i) \tag{5}$$

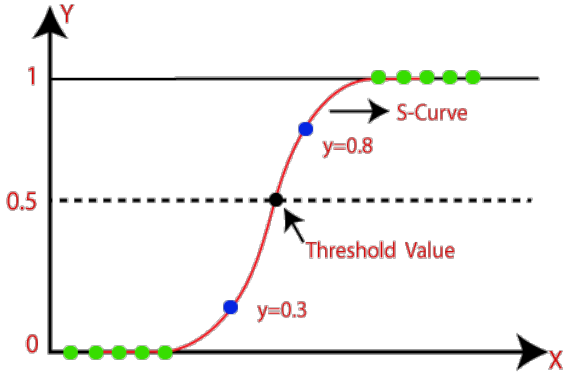

Fig. 1. example of logistic regression

If we threshold the output probability at 0.5, we can induce a decision rule of the form

$$y(x) = 1 \iff p(y = 1|x) > 0.5 \tag{6}$$

Logistic regression may be readily expanded to handle higher-dimensional inputs. If we set the threshold for these probabilities to 0.5, we have a linear decision boundary with the normal (perpendicular) provided by w. In the next part, we shall derive the loss function associated with logistic regression.

A. Maximum likelihood estimation

The negative log-likelihood for logistic regression is given by

$$NLL(w) = -\sum_{i=0}^{N} \log[\mu_i^{(I(y_i=1)} \times (1-\mu_i)^{(y_i=0)}] = -\sum_{i=0}^{N} [y_i \log \mu_i + (1-y_i)\log(1-\mu_i)] \tag{7}$$

This is also called the cross-entropy error function. Unlike linear regression, we can no longer write down the Maximum Likelihood Estimation in closed form. Instead, we need to use an optimization algorithm to compute it. For this, we need to derive the gradient and Hessian. In the case of logistic regression, the gradient and Hessian are given by the following

$$g = \frac{d}{dw} f(w) = \sum_i (\mu_i - y_i) x_i = X^T(\mu, y) \tag{8}$$

$$h = \frac{d}{dw} g(w)^T = \sum_i (\Delta_w \mu_i) x_i^T = S^T S X \tag{9}$$

where $S = diag(\mu_i(1-\mu_i))$. Here the NLL is convex and has a unique global minimum.

III. THE PROBLEM FORMULATION

In this section, We will analyse the Titanic dataset and try to predict the survival of passengers based on the information available. The logistic regression technique is used as a classification model in further analysis. Let's start by addressing any potential missing values.

A. Missing Values

1) age: Since Age doesn't follow uniform distribution, using constant imputation might not always give the best results. Let's use linear regression-based imputer to fill the missing values.

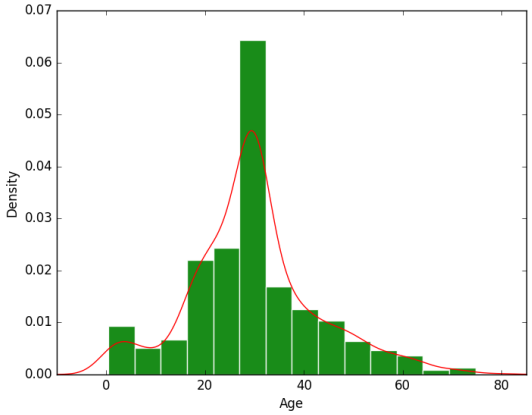

Fig. 2. Age missing values

2) Port of Embarkation: Whenever we encounter missing values for embarked, we impute the missing values with Southampton, the port most people boarded,

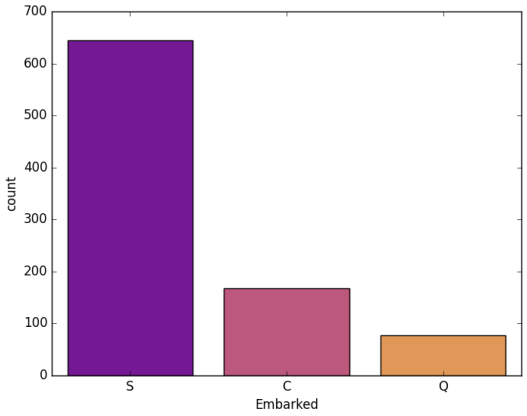

Fig. 3. Embarked-list

3) Fare: Again following a linear regression-based imputation would work well compared to a constant imputation method due to skewed distribution.

B. Exploratory Data Analysis

1) Age VS Survival: The distribution of the survived and dead is similar. One notable difference is that a large proportion of children survived. This shows evident attempts were taken to save children by giving them a place in life rafts.

2) Ticket Classes VS Survival: The first class was the safest and the third class was the unsafest travelling option. Since there is an inherent order in the nature of the class, let's stick with label encoding for the ticket class.

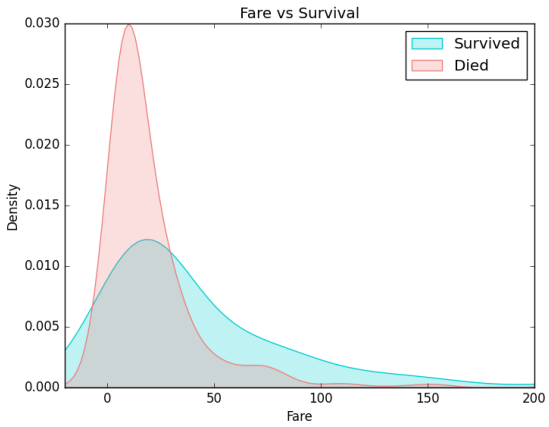

Fig. 4. Fare vs survival

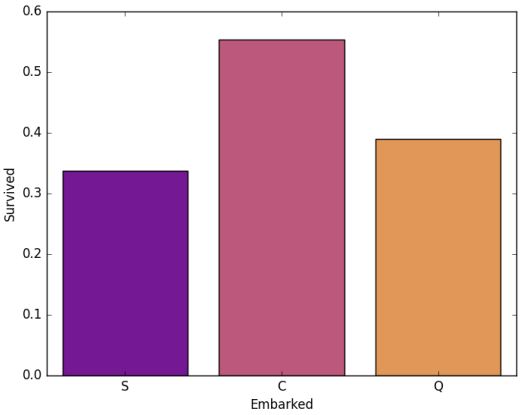

Fig. 7. Embarked Survived

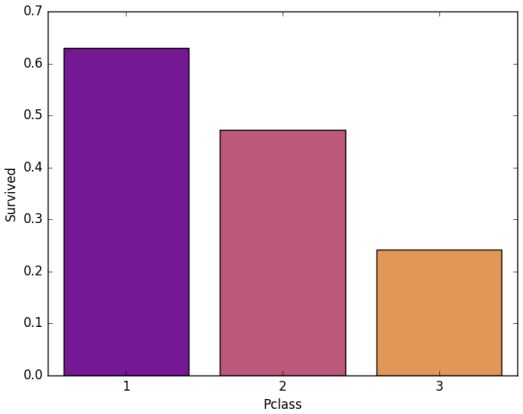

Fig. 5. Passenger class

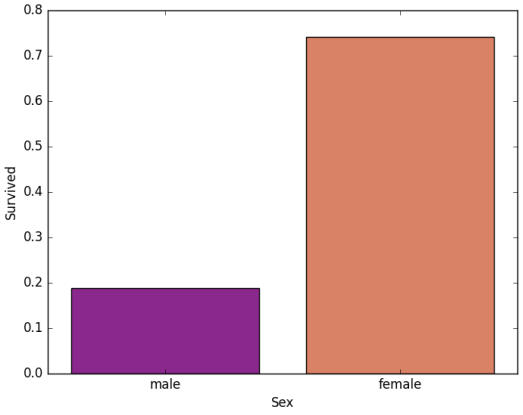

Fig. 6. gender ratio

3) Gender vs Survival: Females had more chances of survival. This could be because more importance has been given to them during evacuation (in accordance with the movie).

4) Fare vs Survival: Passengers who were able to afford more had a higher survival rate compared to those who couldn't. This could also be related to the location of rooms offered and the ease of access to lifeboats during the tragedy.

5) Embarkation vs Survival: People who boarded from Cherbourg, France had the highest survival rate, and those who boarded from Southampton were less likely to survive. While this doesn't make much sense intuitively, the average fare spent by a Cherbourg passenger was 59.95 dollars, the highest among all the three locations. This could mean that people who boarded from Cherbourg, on average, were significantly richer and influential and had better chances of survival.

IV. Model Evaluation

A. confution matrix

Logistic Classifier summary	
Metric	Score
Accuracy	0.8529
Precision	0.7906
Recall	0.8391
F1 score	0.8141

After Data cleaning, Preprocessing and Feature Engineering, Using the concepts learnt through SVM, a kernel-based Logistic regression model with Gaussian kernel was trained and tuned over hyper-parameters $\Lambda and C$. This gave a significant performance boost compared to the previous model due to the fact that this enables the model to capture nonlinearities better. This could also count due to the improved imputation of the method used in the new analysis. The confusion matrix and the evaluation metrics for Gaussian Kernel Logistic Regression are reported below:

In a classification setting, accuracy may not be the best score to consider especially in a skewed class situation [4]. The training data in our case is quite balanced and not a concern. We also evaluate the F1 score and the scores of precision and recall. Precision is the ratio of correct positive predictions to the total positive predictions of our model. The recall is the ratio of positive instances that are correctly detected by

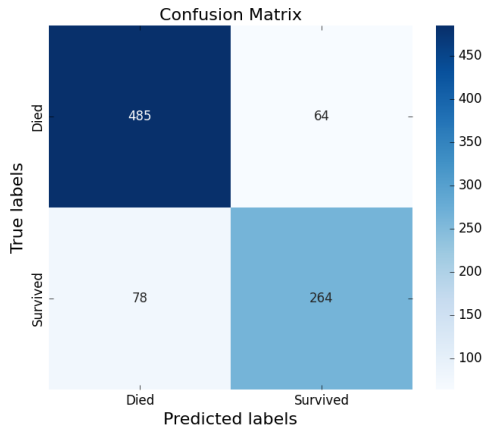

Fig. 8. Confusion Matrix

our classifier. The F1 score is simply a harmonic average of these two. Our classifier seems to have performed well in all the metrics discussed. Both the accuracy and F1 score have improved by approximately two per cent which suggests an improvement over the linear logistic regression model

B. ROC curve

ROC curve is another common tool used with binary classifiers. The ROC curve plots the true positive rate (another name for recall) against the false positive rate. The FPR is the ratio of negative instances that are incorrectly classified as positive. It is equal to one minus the true negative rate, which is the ratio of negative instances that are correctly classified as negative. The TNR is also called specificity. Hence the ROC curve plots sensitivity (recall) versus 1 – specificity.

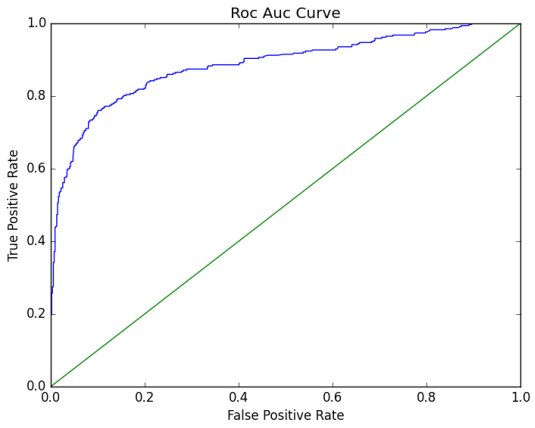

Fig. 9. false positive rate vs true positive rate

The dotted line represents the ROC curve of a purely random classifier; a good classifier stays as far away from that line as possible (toward the top-left corner). One way to compare classifiers is to measure the area under the curve (AUC). A perfect classifier will have a ROC AUC equal to 1, whereas a purely random classifier will have a ROC AUC equal to 0.5. The area under the ROC curve of our logistic regression classifier is 0.8833.

V. CONCLUSIONS

While there was some amount of luck involved, Socio-Economic factors such as income, influences, class tickets and factors such as gender, and travelling with children seem to have impacted the survival rate. This could also be used to explain the inherent biases and favouritism we hold as a society towards the rich section of the population. Extensions to traditional Logistic Regression like the use of Gaussian Kernels and the kernel trick usually applied to SVMs and better imputation give improvements in the prediction of logistic regression classifiers.

REFERENCES

[1] Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. " O'Reilly Media, Inc.".

[2] Murugan, S., Kumar, B. M., & Amudha, S. (2017, September). Classification and prediction of breast cancer using linear regression, decision tree and random forest. In 2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC) (pp. 763-766). IEEE.

[3] Steyerberg, E. W., Harrell Jr, F. E., Borsboom, G. J., Eijkemans, M. J. C., Vergouwe, Y., & Habbema, J. D. F. (2001). Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. Journal of clinical epidemiology, 54(8), 774-781.

[4] Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. The journal of educational research, 96(1), 3-14.

[5] Cabrera, A. F. (1994). Logistic regression analysis in higher education: An applied perspective. Higher education: Handbook of theory and research, 10, 225-256.

[6] Wright, R. E. (1995). Logistic regression.

[7] LaValley, M. P. (2008). Logistic regression. Circulation, 117(18), 2395-2399