A Mathematical Essay on Logistic Regression

Vignesh Kumar S
Department of Chemical Engineering
Indian Institute of Technology, Madras
Email address: ch18b118@smail.iitm.ac.in

***Abstract*—The purpose of this article is to understand Logistic Regression and the mathematics behind it. In this work, we also demonstrate the application of logistic methods illustrated through one of the most infamous shipwrecks in history, the sinking of the Titanic. Due to the complex nature and the abstractness involved in predicting the survival of a candidate, it is often difficult to obtain a good solution through manual approaches. Adopting a machine learning technique can help in obtaining insights and analyse different features in a large amount of data comprehensively.**

I. Introduction

In this paper, we will study Logistic Regression, a technique used to analyze and model the probability of a certain class, usually a dichotomous outcome. We will use this technique to analyze whether the survival of passengers was related to factors such as age, affordability, gender, travelling along with a group. This analysis would aid in obtaining insights among a large population and understand the trend behind survival and biases that could have existed at that time.

The classification problem attempts to establish a relationship between a categorical target variable with one or more explanatory variable(s). Logistic Regression is a discriminatory statistical modelling technique, which itself models the probability of outputs in terms of inputs. 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 below cutoff as another class.

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

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

II. Logistic Regression

The goal in classification problem is to take an input feature $\boldsymbol{x}$ and assign it to one of the two classes (K classes in general). The input space is thereby divided into decision regions whose boundaries are called decision surfaces. Logistic Regression is a linear model for classification, which means the decision surfaces are linear functions of the input vector $\boldsymbol{x}$.

The simplest case is to model $y(\boldsymbol{x})$ = $\mathrm{w}^T\boldsymbol{x} + w_0$ so that y is a real number. But for a classification problem, we seek to obtain probabilities that lie in (0,1) range. Hence, we apply a non-linear function $f()$ such that

$$y(\boldsymbol{x}) = f(w^T\boldsymbol{x} + w_0)$$

returns the posterior probabilities. Decision surfaces correspond to $y(\boldsymbol{x}) = constant$, which implies $w^T\boldsymbol{x} + w_0 = constant$ for one to one f.

The non-linear function $f()$ used in logistic regression is a sigmoid function that outputs a number between 0 and 1.

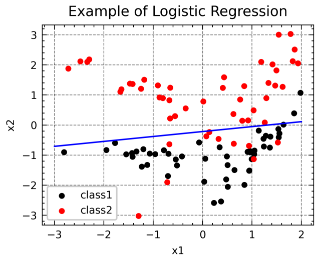
$$\sigma(u) = \frac{1}{1 + e^{-u}}$$

$$\hat{p} = \sigma(w^T\boldsymbol{x} + w_0)$$

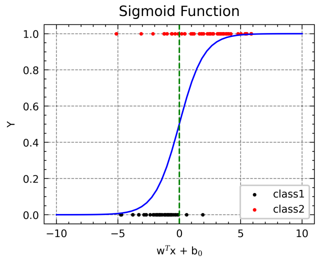
The estimated probabilities can easily be converted into a binary classifier by predicting

$$\hat{y} = \begin{cases} C_1 & \text{if } \hat{p} < Threshold \\ C_2 & \text{if } \hat{p} \geq Threshold \end{cases}$$

where C_1 and C_2 are the two classes. The threshold is usually set to 0.5.



(a) Logistic Regression decision surface separating two classes



(b) Predicting classes based on Threshold

A. Discriminative Model Setting

Let's try to model the above in a discriminative setting which models $P(Y|X)$, where X, y $\in$ $\mathbb{R}^d$ x $\{\pm1\}$. Note that we have used y as $\{\pm1\}$ for the ease of mathematical modelling whereas $\{0,1\}$ could also be used in general. We assume,

$$P(Y = 1|X = \boldsymbol{x}) = \sigma(w^T\boldsymbol{x}) \tag{1}$$

Since we are dealing with a binary classifier now,

$$P(Y=-1|X=\boldsymbol{x}) = 1-\sigma(w^T\boldsymbol{x}) = \sigma(-w^T\boldsymbol{x}) \quad (2)$$

The bias term w_0 could easily be incorporated in $\boldsymbol{w}$ as an extra feature and $\boldsymbol{x}$ as [$\boldsymbol{x}$,1] whenever required.

Equations (1) and (2) could be combined into a single expression as follows:

$$P(Y=y|X=\boldsymbol{x}) = \sigma(yw^T\boldsymbol{x}) \quad (3)$$

B. Parameter Learning

The parameters can be estimated by Maximum Likelihood estimation. Let the number of data points be *M* and *i* be an integer such that 1≤*i*≤*M* and be used to represent a particular instant in the datapoint.

$$\begin{aligned} Likelihood\,\mathcal{L}(\boldsymbol{w}) &= P(y_1, y_2, ...y_M|x_1, x_2, ...x_M) \\ &= \prod_{i=1}^{M} P(Y_i = y_i|X_i = xi, \boldsymbol{w}) \\ &= \prod_{i=1}^{M} \sigma(y_i w^T \boldsymbol{x_i}) \end{aligned}$$

$$\log\mathcal{L}(\boldsymbol{w}) = \sum_{i=1}^{M} \log\sigma(y_i w^T\boldsymbol{x_i})$$

We could maximise log-likelihood or minimise negative log-likelihood to estimate the parameters.

$$\hat{R}(\boldsymbol{w}) \triangleq -\log\mathcal{L}(\boldsymbol{w}) = \sum_{i=1}^{M} \log\left(1+exp(-y_i w^T\boldsymbol{x_i})\right)$$

where $\hat{R}(\boldsymbol{w})$ is called Empirical logistic loss function. To get the minimum cost, the gradient at optimum should tend to zero.

$$\nabla\hat{R}(\boldsymbol{w}) = \sum_{i=1}^{M} \sigma(-y_i w^T\boldsymbol{x_i})(-y_i\boldsymbol{x_i})$$

However, there are no closed-form solutions available for the above optimization problem. Hence, we resort to Gradient descent to reach the minima.

C. Understanding the Loss function

For the ith datapoint, w$^T\boldsymbol{x_i}$ determines the nature of prediction. If the sign (+ or -) of w$^T\boldsymbol{x_i}$ matches with y$_i$,the predicted value is correct and the empirical loss function is less. If the predicted value does not match with the true prediction, the cost incurred becomes high. So the loss function maps a high cost to a misclassification point and hence minimising this, we obtain a solution with lower misclassification error.

D. Gradient Descent

The Gradient is the direction along which a function increases the maximum. Going along the opposite direction of Gradient gives the maximum decrease. On each iteration, the cost is reduced by moving along this direction to a certain step size, which is influenced by the hyperparameter learning rate (η). This is continued till a minima is reached or a stopping criterion is met.

1) Parameter Initialization: Parameters can be assigned random values, to begin with. In logistic regression, often, the parameters are initialized to zero.

2) Stopping Criteria: The convergence to local minima is checked by evaluating if a stopping criterion is met. Some of the choices for stopping criteria include:

- Change in the cost function is less than a threshold
- Change in gradient is less than a threshold
- Change in parameters is less than a threshold
- Number of iterations has reached its threshold

```
ALGORITHM Initialize $\boldsymbol{w_1}$
Repeat till convergence
{
    $\boldsymbol{w_{t+1}} := \boldsymbol{w_t} - \eta\nabla\hat{R}(\boldsymbol{w_t})$
    $\quad = \boldsymbol{w_t} + \eta\sum_{i=1}^{M}\sigma(-y_i w^T\boldsymbol{x_i})(-y_i\boldsymbol{x_i})$
}
```

where η is the learning rate and t is the iteration number.

E. Regularised Logistic Regression

In order for the model to generalise well and to prevent overfitting, a penalty function could be added to the loss function. The choices of penalty function could be $l1$ norm (Lasso), $l2$ norm (Ridge) or a combination of both (Elastic Net). For a $l2$ norm penalty function, Empirical loss function is:

$$\hat{R}(\boldsymbol{w}) = \sum_{i=1}^{M} \log\left(1+exp(-y_i w^T\boldsymbol{x_i})\right) + \frac{\lambda}{2}\|\mathbf{w}\|_{\mathbf{2}}^{\mathbf{2}}$$

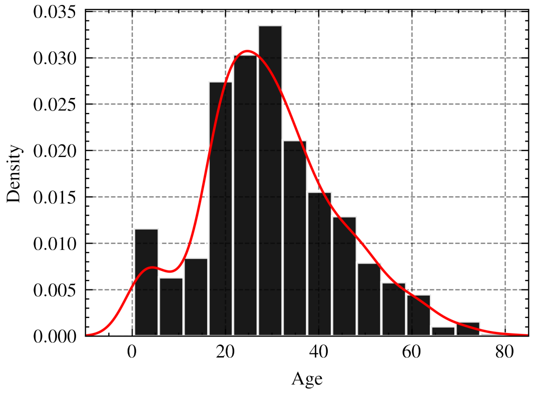
The parameters are obtained through the gradient descent method illustrated earlier.

III. The Problem

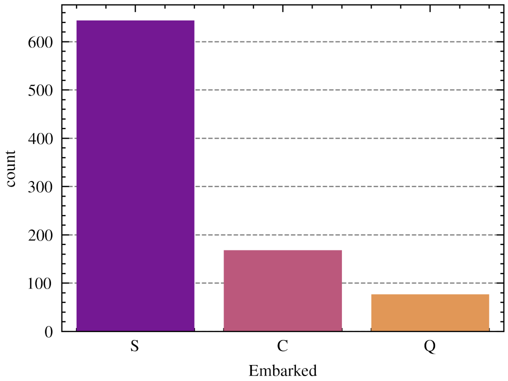
In this section, We will analyse the Titanic dataset and try to predict the survival of passengers based on the information available. The logistic regression technique is used as a classification model in further analysis. Let's start by addressing any potential missing values.

A. Missing Values

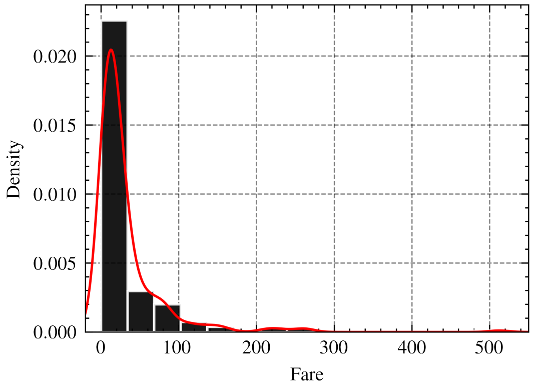
1) Age: Since Age doesn't follow uniform distribution, using constant imputation might not always gives the best results. Let's use linear regression based imputer to fill the missing values.



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

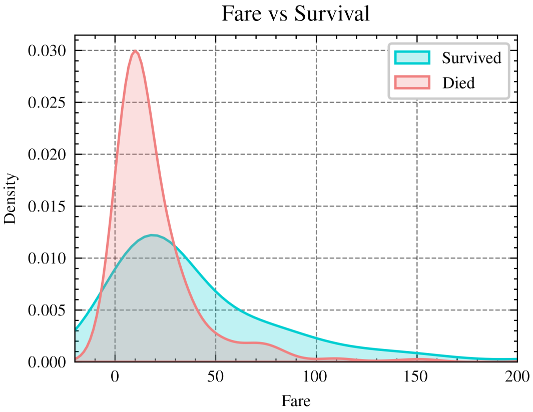


3) Fare: Again following a linear regression based imputation would work well compared to a constant imputation method due to skewed distribution.

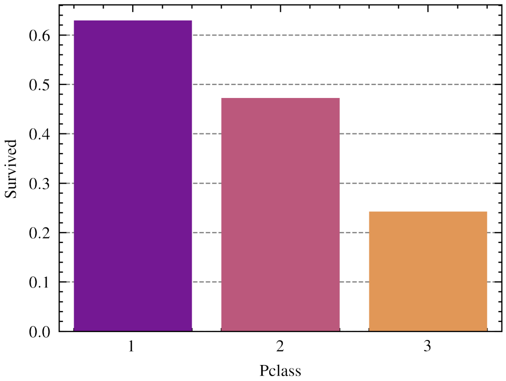


B. Exploratory Data Analysis

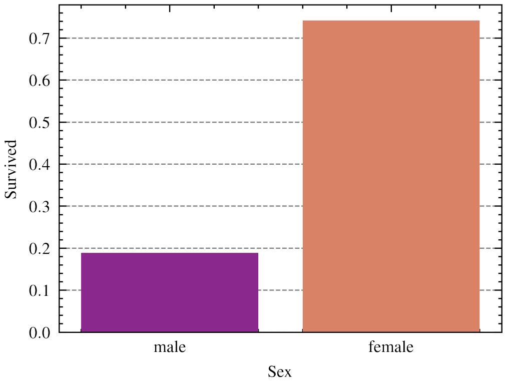
1) Age versus Survival: The distribution of the survived and dead are similar. One notable difference is that a large proportion of children survived. This shows evident attempts were taken to save children by giving them a place on life rafts.



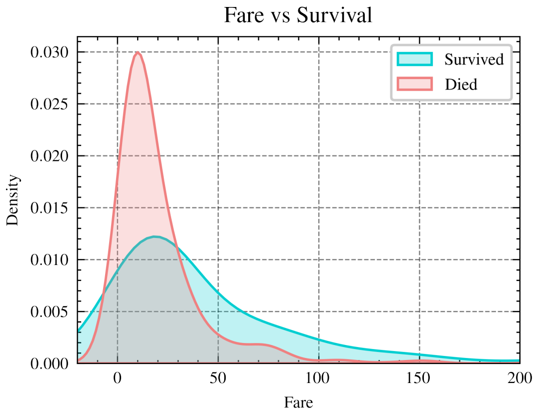
2) Ticket Classes versus Survival: The first class was the safest and the third class was the unsafest travelling option. Since there is a inherant order in the nature of class, let's stick with label encoding for ticket class.



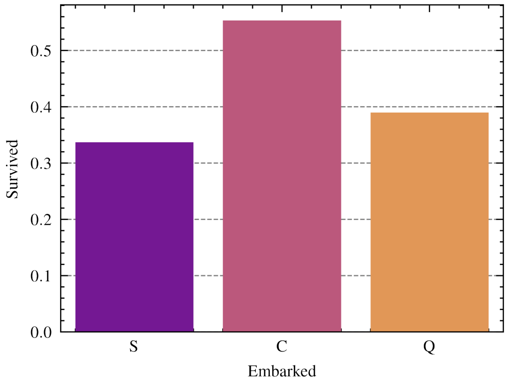
3) Gender vs Survival: Females had more chance of survival. This could be because more importance has been given to them during evacuation (in accordance with the movie).



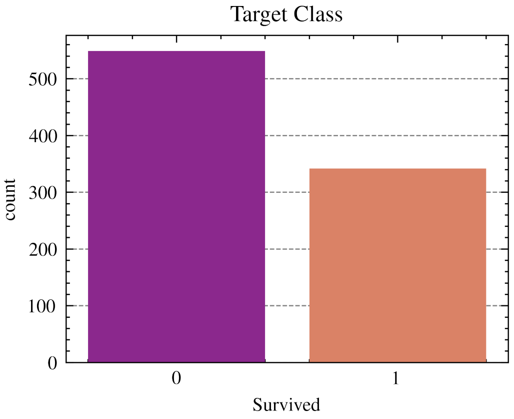
4) Fare vs Survival: Passengers who were able to afford more had a higher survival rate compared to those who couldn't. This could also be related to the location of rooms offered and the ease of access to lifeboats during the tragedy.



5) Embarkation vs Survival: People who boarded from Cherbourg, France have 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 boarded from Cherbourg, on average, were significantly richer and influential and had better chances of survival.



6) Imbalancedness: There is a slight imbalancedness in the dataset. There are more passengers who have not survived (61.6%) and less who have survived (38.4%). Though this is not significant, we will grid search if we need to apply any class weights to correct this measure.



C. Feature Engineering

1) Honorifics: The name attribute along with first and last name contained titles. Honorifics (Titles) were quite commonly used during the era in which the tragedy happened. This would be a useful feature to extract which could determine how rich and influential a person was. It is also interesting to see from the training dataset that the captain didn't survive and went down with the ship.

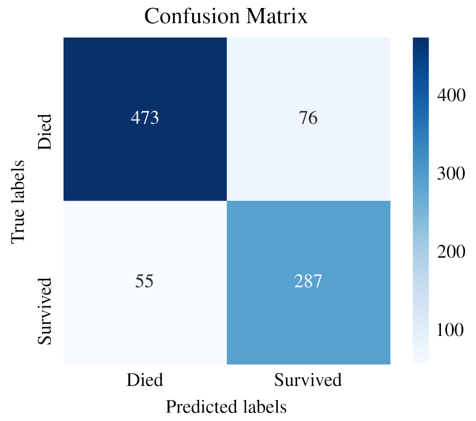
2) Cabin Group: Even though Cabin contained a lot of missing values, considering only the cabin group of the available data and taking missing values as a separate category showed some trends. The cabin values are also representative of the ticket classes and the closeness to lifeboats. Hence, a new feature *Cabin Group* was constructed.

3) Ticket shared: The ticket attribute consists of ticket numbers with or without some added prefix. Without available information, it's difficult to establish a relationship about the prefixes. Dropping prefix, a closer analysis on tickets showed that the tickets id were not unique. Survival rates of passengers could be predicted upon knowing the survival of candidates who had shared the same ticket number. A new attribute was constructed which mapped the number of people who shared common ticket numbers.

One-Hot encoding was used to preprocess data without inherent order and label-coding was used with data that possessed some order (eg. ticket class)

D. Model Evaluation

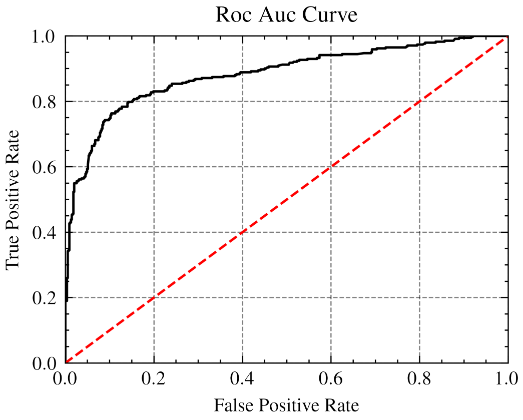
After Data cleaning, Preprocessing and Feature Engineering, Using the concepts learnt through SVM, a kernel based Logistic regression model with Gaussian kernel was trained and tuned over hyper-parameter λ and C. This gave a significant performance boost compared to previous model due to the fact that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



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

In a classification setting, accuracy may not be the best score to consider especially in a skewed class situation. 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 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 approximately two percent which suggests an improvement over linear logistic regression model

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.



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*.

IV. Conclusions

While there was some amount of luck involved, Socio-Economic factors such as incomes, influences, class tickets and factors such as gender, travelling with children seems 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 population.

Extensions to traditional Logistic Regression like use of Gaussian Kernels and the kernel trick usually applied to SVM's and better imputation gives improvements in the prediction of logistic regression classifiers.

References

[1] Aurelian Geron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools and Techniques to Build Intelligent Systems, pp. 112–180

[2] Christopher M. Bishop, Pattern Recognition and Machine Learning

[3] Joanne Peng, An Introduction to Logistic Regression Analysis and Reporting