# Titanic Logistic Regression

Daniele Gilio

April 6, 2020

## 0.1 Introduction

In this assignment we were asked to train a logistic regression model in order to predict the survival probability given the following features:

- Class
- Gender
- Age
- Siblings/Spouses on board
- Parents and Children on board
- Passenger Fare

The labels associated with those features are binary values indicating whether or not a given passenger survived the disaster. The Training Set is comprised of 710 samples whereas the Test Set is made of 177.

## 0.2 Data Analysis

The first thing we did was an analysis of the data in order to have an idea about how they are distributed and what features might be the most relevant. This was done by plotting the features pairwise and see if a possible decision boundary could be identified. This initial process hinted that the most influential features might be Class and Gender, as we can see in Figure 1. To further investigate this we plotted the "Empirical" chance of

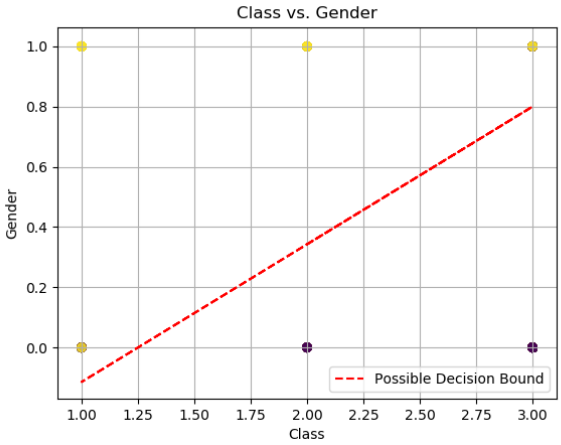


Figure 1: Class vs. Gender Scatter plot

survival, which is basically the data relative frequency, against the features we spotted. The graphs in Figure 2 tell us that we might be on the right track.

## 0.3 Training

The training was performed for $2 \cdot 10^5$ steps and we used a Learning Rate equal to 0.001. Increasing the Learning Rate lead to oscillating values in the loss functions whereas decreasing it worsened the final training loss value and consequently the training accuracy. The plots in Figure 3 show both the training and the test loss and their matching accuracy. If we base our model evaluation purely on the loss function, we can say that it may be enough to perform between $5 \cdot 10^4$ and $7.5 \cdot 10^4$ steps to reach convergence as these values are the ones in which the test loss settles on a constant value. Looking at the accuracies though shows that the best step value is probably around $1.5 \cdot 10^5$, because the test accuracy has a peak and it decreases shortly after indicating a possible overfitting issue. The final values for loss functions and accuracies can be found in Table 1.

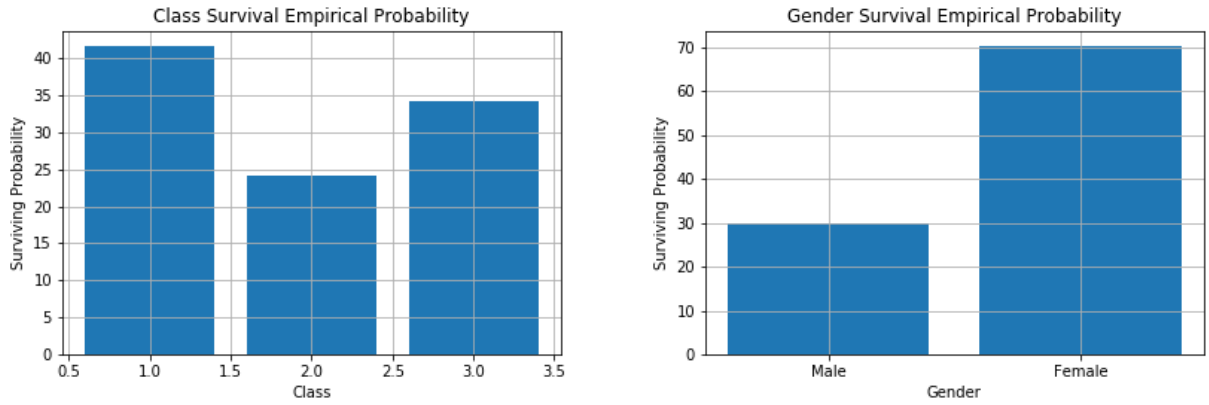


Figure 2: Empirical Probabilities for Class (Left) and Gender (Right)

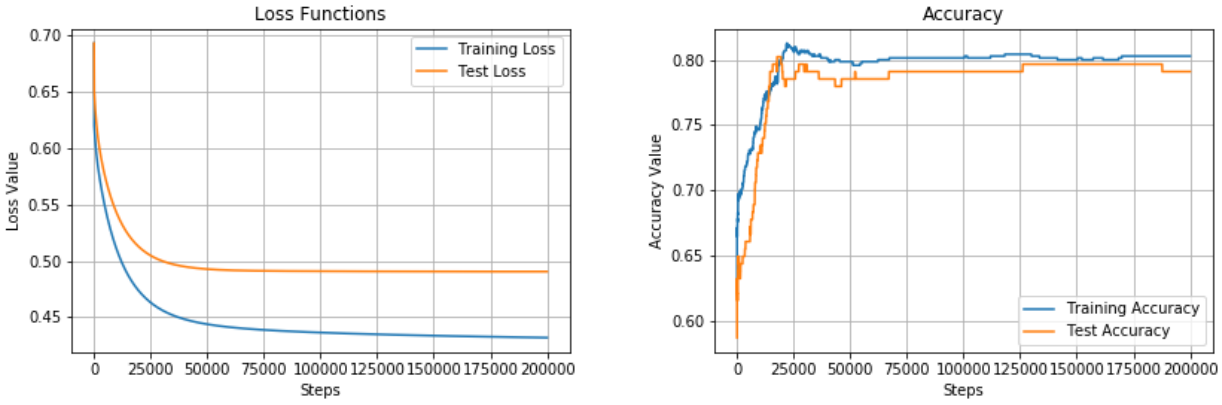


Figure 3: Loss (Left) and Accuracy (Right)

| | Training | Test |
|---|---|---|
| Loss | 0.432 | 0.490 |
| Accuracy | 80.281 | 79.096 |

Table 1: Losses and Accuracies

## 0.4 Model Analysis and Evaluation

If we take a look at the weights in Table 2, obtained at the end of the training, we can see that our first intuition was right. Gender is the most influential feature, followed by Class, Siblings/Spouses, Parents and Children, Age and Passenger Fare. We can conclude that the people which were most likely to survive were women travelling alone in first class. The opposite is also true: men in third class travelling with other people were most likely to die. A series of $10^5$ "Educated Guesses" (Figure 4) proves our theory and enlightens the fact the "Women and Children First" is not a fictional saying but a real world rule of navigation. Given these premises we are glad we were not on the ship since the model predicts a probability of 27.77% in a best case scenario ($1^{st}$ Class) and a 1.92% in the worst one ($3^{rd}$ Class).

| | Class | Gender | Age | Siblings/Spouses | Parents/Children | Fare |
|---|---|---|---|---|---|---|
| Weights | -0.974 | 2.768 | -0.033 | -0.302 | -0.099 | 0.003 |
| Bias | 1.56 | | | | | |

Table 2: Weights and Bias Table

## 0.5 Final Thoughts

Plotting the contour of decision bound in the Class vs. Gender graph (Figure 5) tells us that our model is doing a good job in recognising those two features as the most important ones. In our opinion the model is not

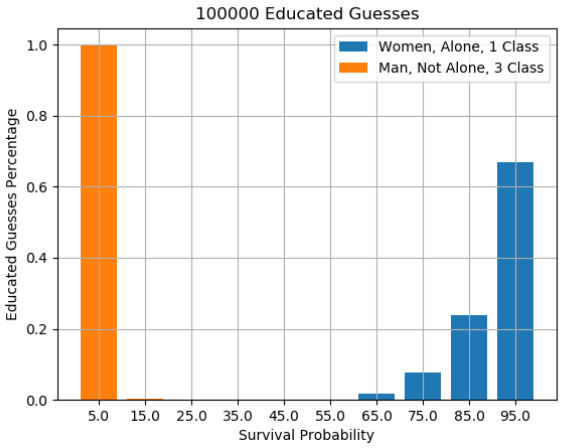


Figure 4: "Educated Guesses"

overfitting the data since the test accuracy is very close to the training one. Plotting the contour for Class vs. Siblings/Spouses shows, on the other hand, the limitations of the model (Figure 5). This is probably due to underfitting, the bias value is in fact rather high compared to the second most influential feature. The model might

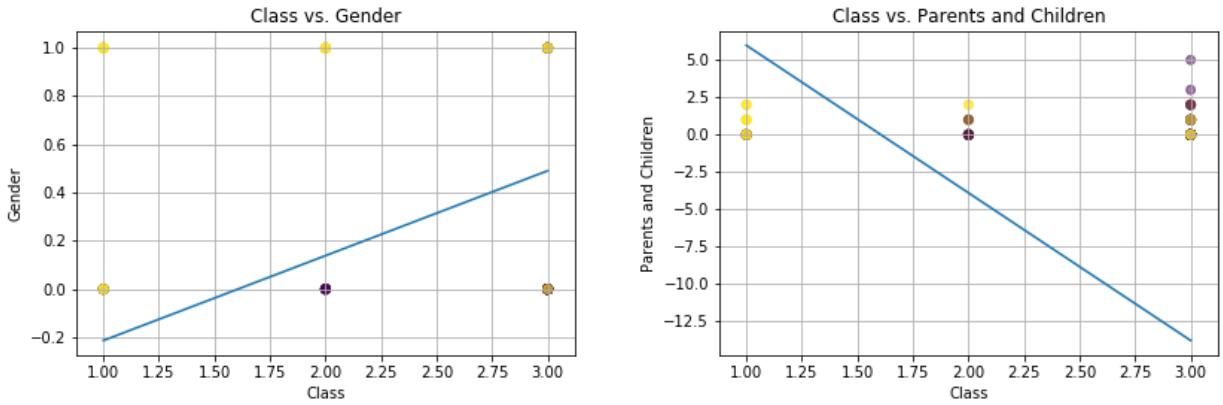


Figure 5: Class vs. Gender (Left) and Class vs. Siblings/Spouses (Right). The contours are plotted taking all the non displayed features equal to 0; it is not ideal but it still works for the purposes of this report.

be improved by adding regularization. We tested both L1 and L2 regularizations with $\lambda = 0.005, 0.01, 0.015$. L1 regularization decreased the accuracy by $\sim 0.6\%$ with $\lambda = 0.015$ whereas L2 increased it by $\sim 0.4\%$. These slight variations in accuracy may indicate that the model is too simple for the problem given, testing a non-linear model might be a better way to represent the data.

I affirm that this report is the result of my own work and that I did not share any part of it with anyone else except the teacher.