Titanic Logistic Regression

Daniele Gilio

April 4, 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
- Sex
- 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 dataset 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 Sex, as we can see in Figure (). To further investigate this we plotted the "empirical" chance of survival, which is basically the relative frequency, against the classes we spotted. The graphs in Figure() 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 0.005. 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() show the both the training and the test loss and their relative accuracy. As we can see 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 (and the test accuracy) settle on a constant value.

0.4 Model Analysis and Evaluation

0.5 Final Thoughts