Machine Learning Homework 1

Malaspina Francesco

8/04/2021

Abstract

The following report analyzes and evaluates a script that trains the logistic regression model for the Titanic problem.

1 Training the model

I found that **0.001** was a good value for the **learning rate** because with a larger value the algorithm was more unstable while with a smaller one it took to long for it to converge.

With this value **100000 iterations** were required for the model to converge. In fact, as it can be seen from the following Figure, the Training Curve at this point is already enough flat both for the accuracy and the losses.

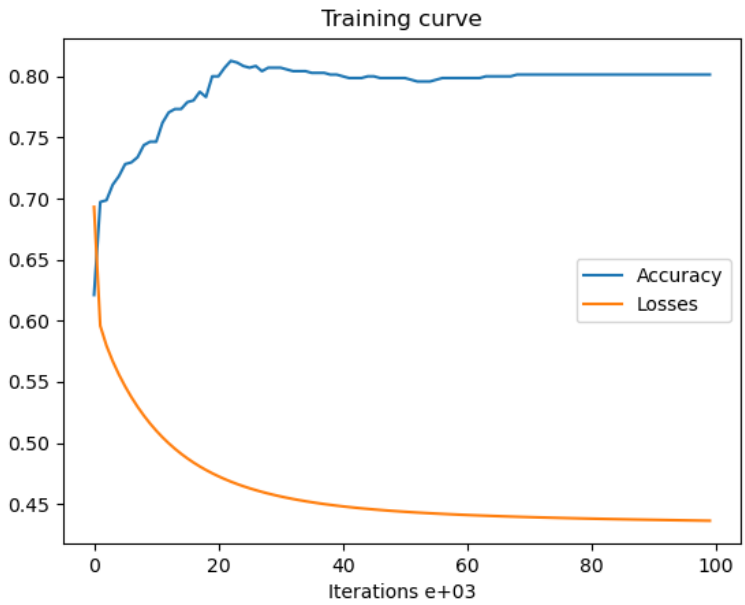

Figure 1: Training curves for Accuracy and Losses

2 Analyzing the model

2.1 Trying out the model

I assumed as my parameters the following: 2nd class, male, 21 years old, 1 spouse aboard, 0 parents or children aboard and 35 as ticket fare; I used the model to estimate what would have been my probability of surviving the disaster and the result was 21%.

2.2 Training accuracy

After the training was complete I computed the probability of surviving of each passenger in the training set according to the model, then I used a threshold of 0.5 to classify each one of them and compared this predictions with the actual data.

The resulting **training accuracy** was 80.14%.

2.3 Weights analysis

The trained model had the following **learned weights**: [-0.79789356; 2.67170526; -0.02611926; -0.28575255; -0.09190851; 0.00533742].

By looking at them we notice that the **biggest** one is the second, representing the **sex** of the passengers, and from it being positive we understand that the women had a much greater chance of surviving that disaster.

The **second** most influential feature is the first one, that is the **class** in which the passengers traveled, and because it is negative we can infer that the people in the first class were more likely to remain alive than those in the second and in the third.

We also notice that the **Age** (the 3rd one) weight is quite small but this is a consequence of the different metric used for this feature, most of the others were classifiers identified with 0 and 1 while age is a number from 0 to around 70. If we take this into consideration we can say that also the age was an important feature of our model and that the youngsters had a greater chance of pulling through.

In the end we can affirm that young women in the first class had the greatest chance to live on, while old men in the third had the lowest.

Figure 2: Scatter plot of class and sex

In the previous Figure I draw a scatter plot of the sex and class with a different color depending on the outcome (yellow dots for the survivors and blue dots for the others), introducing some random noise to separate the dots, which otherwise would have been overlapped. And we can notice that it agrees with the assumption made in the previous paragraph.

3 Evaluating the model

In a different script I assessed the performance of the model on the data taken from the **test set**. The resulting **test accuracy** was 79%. We can say that the model is not overfitting given that the accuracy is very close to the train one. And because of this is possible that it is slightly underfitting. To improve the model we could try exclude one or more of the less influential feature and check if this increases the accuracy or not.

4 Final statement

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.