Titanic: Disaster Prediction

Rajesh Nagula

October 2019

1 Objective

The objective of this assignment is to predict the survival of the passengers aboard the Titanic based on the information available about them using the data regarding the survival of other passengers and their information.

1.1 Functions

The first part of the program is to pre-process the data using pandas. Then we will run 100000 iterations, each requiring 3 functions: gradient, hypothesis and sigmoid. We will now make predictions using trained weights and bias.

2 Methodology

This predictor works on the basis of logistic regression. First of all the data is pre-processed. The null values of a feature are filled with the mean values from the non-null values of that feature. The string features that can be mapped to integers are all mapped to integral values. The features that cannot be mapped to integers and do not show a pattern are dropped from our data. This completes the pre-processing of the data. Now we are ready with non-null integral or float values in our information/features.

Each feature/piece of information regarding the passenger will have a corresponding weight. A product of each feature is taken with it's corresponding weight and is summed up along with the bias to give us the hypothesis of each passenger. This hypothesis is passed through an activation function called as sigmoid function

$$\frac{e^x}{1+e^x}$$

which converts the hypothesis to a float value between 0 and 1. If this activation value yields a value ≤ 0.5, the predicted output is said to be 0(did not survive) and 1(survived) otherwise.

Initially, the weights and bias are initialized to 0. The gradient function computes the hypothesis and output for each passenger in our training data and compares it to the output and returns the small change in weights and bias that can improve the hypothesis and output in accordance with given output. These small changes are made to the weights and bias repetitively(100000 times) in our program. Finally what we will have are the ideal weights and bias required to calculate hypothesis and output given the features.

Finally, we will use the features of each passenger from the test data and along with the weights and bias we arrive at, we will calculate the hypothesis and the outputs. Now the passengerIDs of the passengers in the test data are stored along with output for each passenger in a new CSV file called TitanicSubmission.csv

3 Results

Since we don't know the test outputs, the weights were trained on a random 40% of the training set and applied on the remaining 60% of the training set and was found to have 81% accuracy in predicting the output.

4 Discussion

The code can be made more accurate by generating features from the existing features or by changing the algorithm from logistic regression to random forest classification.

5 Conclusion

The code can take in any training data and learn the weights and bias to classify any test data into two classes i.e. 0 or 1 using logistic regression.