Assignment 2: Data Lab, Logistic Regression

- Due by Friday 1st October, 2021 by 5pm IST.
- To be submitted to the following email address: office.of.gr@gmail.com
- The subject of the email should be: Assignment Number [2]: Data Lab, 2021
- Please clearly mention your name and roll number.
- Submit your work as a single pdf file. Additional material, code, etc can/should also be submitted, but there should be atleast 1 pdf, which has the entire assignment.
- Wherever there is code, in the assignments, the code should be well documented and easy to understand / follow.

The objective of the assignments is three fold. One is to be able to develop expertise in writing and communicating about technical topics. This will be done by using the IEEE conference style format for all assignments. The other is to explain, in your own way, the mathematical ideas that are embedded within the technical topic of interest. For example, in this case it is logistic regression. The third is to use the topic, in this case of logistic regression, to understand a problem from the real world. So in a sense the objective is to write what one may call a mathematical essay on Logistic Regression.

Title could be: Assignment 2: a mathematical essay on logistic regression.
Abstract. Give a brief overview of your assignment.
Author: Name, Department, Institution, Email

Section 1: Introduction

In this section, the 1st paragraph should be on a broad overview of the topic. The 2nd paragraph should be an overview of the technical aspects (i.e. in this case it is logistic regression). The 3rd paragraph should be about the problem that you are aiming to solve/understand using logistic regression. Finally, the 4th paragraph should give an overview of the paper.

Section 2: Logistic regression

This section should outline the key principles underlying Logistic regression.

Section 3: Data

The sinking of the Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the widely considered “unsinkable” RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren’t enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew. While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others. In this challenge, we ask you to build a predictive model that answers the question: “what sorts of people were more likely to survive?” using passenger data (ie name, age, gender, socio-economic class, etc).

Two similar datasets include passenger information like name, age, gender, socio-economic class, etc. One dataset is titled “train.csv” and the other is titled “test.csv”. The train file will contain the details of a subset of the passengers on board (891 to be exact) and importantly, will reveal whether they survived or not, also known as the “ground truth”. The test file dataset contains similar information but does not disclose the “ground truth” for each passenger. It’s your job to predict these outcomes.

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	M / F
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Section 4: The problem

(a) Outline the problem, and plot/visualize the data.
(b) Make progress on the problem, by applying the techniques of logistic regression to the problem at hand.
(c) Discuss any insights and observations.

Section 5: Conclusions

Write about 1 paragraph on the key insights that were obtained from your study and also outline any further avenues for investigation.

References

Please put in all the references that you have used during the assignment. The format should be the same as the IEEE conference format.