# Would you survive the Titanic?

**Machine Learning**

Claudio Cusano

A.A. 2020/2021

RMS Titanic was a British passenger ship that sank in the North Atlantic Ocean in the early morning hours of April 15, 1912, after striking an iceberg during her maiden voyage. The majority of its passengers died in the accident. As we will discover, not all of them had the same chance to survive.

Data about 887 passengers have been collected and randomly divided into a training and a test set. The training set includes 710 samples and is stored in the file `titanic-train.txt`, while the test set is composed of 177 cases and is stored in `titanic-test.txt`. Each row in the files represents a different passenger, and reports the following features:

- the ticket class ($1^{st}$, $2^{nd}$ or $3^{rd}$);
- sex ($0 \rightarrow$ male, $1 \rightarrow$ female);
- age, in years;
- number of siblings and spouses aboard;
- number of parents and children aboard;
- the passenger fare.

The last column reports whether the passenger survived (1) or not (0).

## 1 Train a model

Write a script that trains the logistic regression model for the Titanic problem. Make it draw the training curve (iterations vs. loss). Then answer the following questions:

1. which is a good value for the learning rate?
2. How many iterations are required to converge?

## 2 Analyze the model

Modify the script so that you can answer to the following questions:

1. what would be your probability to survive? (Make a guess about the ticket class, the fare etc.)
2. What is the *training accuracy* of the trained model?
3. Looking at the learned weights, how the individual features influence the probability of surviving?
4. What kind of passengers was most likely to survive? And what kind to to die?
5. Draw a scatter plot showing the distribution of the two classes in the plane defined by the two most influential features. Comment the plot.

## 3 Evaluate the model

Write an evaluation script that assesses the performance of the model on the test set. You can manually copy the parameters obtained with the training script, or you can use the `np.load` and `np.save` functions (`https://numpy.org/doc/1.18/reference/generated/numpy.load.html`). Then answers to the following questions:

1. what is the *test accuracy* of the model?
2. Is the model overfitting or underfitting the training set?
3. How can you increase the performance of the model?

## Assignment

Prepare a report of one or two pages with the answers to the questions (include a short comment for each question). The report must be in the PDF format. Include your name in the report and conclude the document with the following 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."

Make a ZIP archive with the report and the python scripts you used and send it by e-mail before the next Friday.