Investigating the sinking of the Titanic

A proposal

Philip Hartout, Bas Straathof, Vincent Roest

Submitted on December 1, 2016

1 Context, research question and hypothesis

In this project, we would like to look at a dataset documenting several features of the passengers of the Titanic. These features include passenger number, whether they have survived or not, passenger class, the names of the passenger(s), their sex, their age, the number of siblings/spouses on board, the number of parents/children on board, their ticket number, the fare they paid, their cabin number (if mentioned) and where the passenger embarked. Regarding the disastrous aspect of the sinking of the Titanic, a limited number of rescue boats were available, which unfortunately didn't allow the rescue of all 2224 passengers present on board at the moment of the sinking. Subsequently, only 1502 passengers survived. Given this constraint, we formulate the following research question:

RQ1: What are the factors leading to an increased survival of certain passengers in the context of the Titanic disaster?

RQ2: How much does each factor contribute to the increased survival rate?

Intuitively, we make the following hypotheses:

H1: Passengers with a high class (which is a proxy for the socio-economic status), being female or a child will have an increased chance of survival.

H2: Given the events: C: the child survives; F: the female survives; H: a passenger from a high class survives and O: an ordinary passenger without any of the previous feature survives, we hypothesise that $P(C) > P(H) > P(H) > P(O)$.

2 Methods

In order to make a correct prediction taking into account all the features listed in the beginning, we intend to use the following algorithms to make predictions from our dataset[1]:

1. Decision trees
2. Non-linear (degree 3) support vector machine
3. 1-nearest neighbour

Other ?

We will evaluate the several aspects of the performance of each of these algorithms by constructing confusion matrices, learning curves, covariance matrices, computational power to perform the calculations, etc to assess which model is most successful in which respect(s) and to what extent.

[1] As suggested in the project description, we will use the learning algorithms drawn from `sklearn`