COMPUTATIONAL INTELLIGENCE (CI-MAI) - 2013-2014 Project proposal: Survival in the Titanic

Lluís A. Belanche, David Buchaca, Àngela Nebot

Abstract

The sinking of the Titanic is one of a number of famous shipwrecks in history. During her maiden voyage, April 15, 1912, the Titanic sank after colliding with an iceberg, about 375 miles (604 km) south of Newfoundland, killing $1,502$ people out of a total of $2,224$ (passengers and crew). This tragedy shocked the world at the time and ultimately led to better safety regulations for passenger ships. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.

The project amounts to build a prediction model for survival (or survival probability) out of the real data.

1 Introduction

The Titanic dataset is a classical benchmark in the machine learning community. The data you need to used can be found at:

`https://www.kaggle.com/c/titanic-gettingStarted/data`

In particular, you are asked to download the `train.csv` and `test.csv` files. In order to do so, you need to register on www.Kaggle.com[1]. Each row in the datasets represents a person.

Once you have the data you are expected to predict the labels of the test set using a classifier that you will have to train in the train set.

2 A look at the data

In the webpage above there is information about the data, which is reproduced below. The left column has the names of the variables (predictors) and the right column their description:

```
survival        Survival
                (0 = No; 1 = Yes)
pclass          Passenger Class
                (1 = 1st; 2 = 2nd; 3 = 3rd)
name            Name
sex             Sex
age             Age
sibsp           Number of Siblings/Spouses Aboard
```

[1]Kaggle is a company that shares data mining problems and expects people to compete with each other to solve them.

```
parch           Number of Parents/Children Aboard
ticket          Ticket Number
fare            Passenger Fare
cabin           Cabin
embarked        Port of Embarkation
                (C = Cherbourg; Q = Queenstown;
                 S = Southampton)
```

Read the notes on the web and make sure you understand what the variables mean. There are some special notes about some of the variables:

- Pclass is a proxy for socio-economic status (SES)
  1st = Upper; 2nd = Middle; 3rd = Lower.
- Age is in Years; Fractional if Age less than 1. If the Age is estimated, it is in the form xx.5
- With respect to the family relation variables (i.e. sibsp and parch) some relations were ignored. The following are the definitions used for sibsp and parch.
  - Sibling: Brother, Sister, Stepbrother, or Stepsister of Passenger Aboard Titanic
  - Spouse: Husband or Wife of Passenger Aboard Titanic (Mistresses and Fiances Ignored)
  - Parent: Mother or Father of Passenger Aboard Titanic
  - Child: Son, Daughter, Stepson, or Stepdaughter of Passenger Aboard Titanic

  Other family relatives excluded from this study include cousins, nephews/nieces, aunts/uncles, and in-laws. Some children travelled only with a nanny, therefore parch=0 for them. As well, some travelled with very close friends or neighbors in a village, however, the definitions do not support such relations.

3 Evaluating your models

The data has been split into two groups, a 'training set' and a 'test set'. For the training set, you are provided with a label for each passenger (survived or not). You will use this set to build your model to generate predictions for the test set. Your score is the percentage of passengers you correctly predict. In order to know your score you need to send (submit) a vector of predictions to Kaggle; this vector will have the same length as the number of elements in the test set. Your submission should be in csv format.

Once you submit you model you will know the error that you get and you will be placed in a list with all the other participants. This list is called the *Kaggle leaderboard* and has a public and private component: 50% of your predictions for the test set have been randomly assigned to the public leaderboard (the same 50% for all users). Your score on this public portion is what will appear on the leaderboard. At the end of the contest, Kaggle will reveal your score on the private 50% of the data, which will determine the final winner. This method prevents users from 'overfitting' the public leaderboard. Anyway, once you deliver your project, we can give you your score on the full test set if you are interested.

4 Pre-Processing the data

The first line in the `train.csv` file contains the names of the variables in the training set, which coincide with those in the test set, with the exception of Survived (that is absent from the test set). Important information:

- There is a column, PassengerId, which is probably irrelevant for the task.
- There are some columns that correspond to strings, such as "Name", while others are categorical, as "embarked".

Question 1 *Are all the variables relevant for prediction purposes? In case you remove some columns you have to explain the changes in performance.*

You are expected to complete the following table in the report with the type of the columns of the initial data.

Variablename	*Type*
PassengerId	*Integer*
Pclass	*Integer*
Name	*String*
Sex	
Age	
SibSp	
Parch	
Ticket	
Fare	
Cabin	
Embarked	

Question 2 *Are all these types the ones you would like to have in order to develop your classifier? How do you represent non-numerical variables, if necessary?*

5 Different Models

In order to predict passenger survival you might want to use different techniques. Please consider using different neural networks and/or classical statistical methods (discriminant analysis, logistic regression).