School of Data and Computer Science

Sun Yat-sen University

ARTIFICIAL NEURAL NETWORKS AND PRINCIPAL

Titanic: Machine Learning from Disaster

Mid-term Report

Author:

Zicong Hong 15344015

Jiqi Zhang 16340286

Jinliang Zhang 16340288

Teacher:

Ruixuan Wang

Academic year 2018-2019

Contents

1

Introduction

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. In this lab, we will give out an analysis of what sorts of people were likely to survive and apply several tools of machine learning to predict which passengers survived the tragedy. The model has been tested by the Kaggle platform and the code has been upload to https://github.com/hongzicong/NN_Kaggle_Project.

2

Data set description

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
Name	Name of Passenger	
sex	Sex	
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

3

Data Preprocess

3.1 Categorical Data Transformation

First, we find there are many categorical variable in the data set provided by Kaggle, such as Sex, Embarked, Name and Ticket, according to Figure 1. Thus, we need to transform them into indicator variables or just drop them. However, the Name and Ticket data have many categories, so we decide to drop them for convenience.

```
In [8]: train.head()
Out[8]:
   PassengerId  Survived  Pclass  \
0            1         0       3
1            2         1       1
2            3         1       3
3            4         1       1
4            5         0       3

                                                 Name     Sex   Age  SibSp  \
0                             Braund, Mr. Owen Harris    male  22.0      1
1  Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0      1
2                              Heikkinen, Miss. Laina  female  26.0      0
3        Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0      1
4                            Allen, Mr. William Henry    male  35.0      0

   Parch            Ticket     Fare Cabin Embarked
0      0         A/5 21171   7.2500   NaN        S
1      0          PC 17599  71.2833   C85        C
2      0  STON/O2. 3101282   7.9250   NaN        S
3      0            113803  53.1000  C123        S
4      0            373450   8.0500   NaN        S
```

Figure 1: Overview for the data set provided by Kaggle

3.2 Missing Data Detection

Before selecting model, we need to read the data set and then find out whether some missing value exist in the data set or not. According to the Titanic data set given by Kaggle, we list the variables with missing value and the percentage of the missing value for them as follows.

Variable	Percentage for missing value
Age	19.86%
Cabin	77.10%
Embarked	00.22%

3.3 Missing Data Filling

For these missing value, we can just drop them or fill them based on other correlated data. Although the latter one will be more difficult, it can preserve more details of the data set.

Because there are about 20 percent of the Age data is missing, we decide to use the most correlated variable to fill the missing Age data. As for Embarked, there are also a few missing data. By contract, the major part of the Cabin data is missing. Even though we use other variables to fill the missing Cabin data, the model may become worse. Therefore, we decide to drop the Cabin data. Next, we show the correlation matrix for all variables in Figure 2.

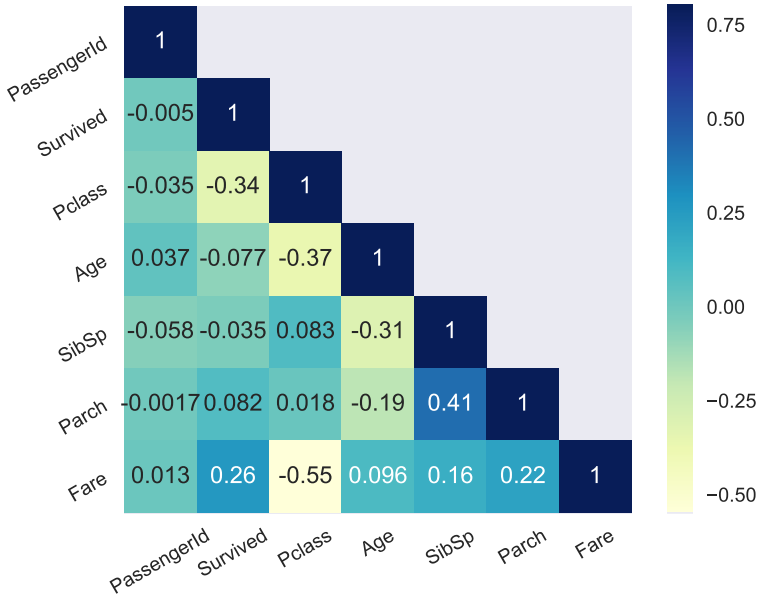

Figure 2: Correlation matrix for all variable

From Figure 2, we can see that the correlation between Age and Pclass is strongest. Furthermore, some references also suggest that there are strong correlation between the age of passengers and their ticket class in Titanic. Therefore, for the passengers with missing Age data, we will use the average age values in their respective Pclass as their age.

4

Logistic Regression

4.1 First trial

Logistic regression model is a statistical model that uses a logistic function to model a binary dependent variable. It is suitable for our problem because the predicted result of problem is also binary variable, i.e., survived/dead. Therefore, depending on LogisticRegression in sklearn, we first train a prediction model based on the default parameters.

	precision	recall	f1-score	support
0	0.79	0.89	0.83	99
1	0.84	0.70	0.76	80
avg / total	0.81	0.80	0.80	179

Figure 3: Report for first trial of logistic regression

Figure 4: Result for first trial of logistic regression

4.2 Improvement

However, the result is not very well, thus we decide to adapt the parameters of the model for a better score. Unfortunately, after several trials shown in Figure 5, the predicted result doesn't have any improvement and even becomes worse.

Therefore, we decide to utilize Name data instead of dropping it as above. After checking the Name data again, we find that the social status can be found out from the Name data, such as Master in Figure 6. Then, we append a new variable isMaster into the train data and train

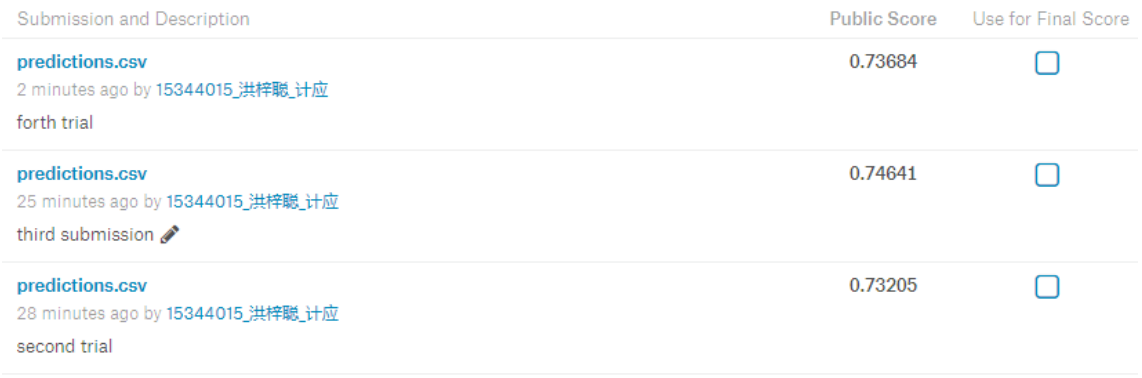

Figure 5: Result for several trials of logistic regression

the model again. The result is much better than before and get a score of 0.75598 as shown in Figure 7.

	A	B	C	D	E	F
1	Passenger	Survived	Pclass	Name	Sex	Age
2	1	0	3	Braund, Mr. Owen Harris	male	
3	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	
4	3	1	3	Heikkinen, Miss. Laina	female	
5	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	
6	5	0	3	Allen, Mr. William Henry	male	
7	6	0	3	Moran, Mr. James	male	
8	7	0	1	McCarthy, Mr. Timothy J	male	
9	8	0	3	Palsson, Master. Gosta Leonard	male	
10	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	
11	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	
12	11	1	3	Sandstrom, Miss. Marguerite Rut	female	
13	12	1	1	Bonnell, Miss. Elizabeth	female	
14	13	0	3	Saundercock, Mr. William Henry	male	
15	14	0	3	Andersson, Mr. Anders Johan	male	
16	15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	
17	16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	
18	17	0	3	Rice, Master. Eugene	male	

Figure 6: Surname in the Name Data

Figure 7: Result for improvement trial