Machine Learning Project

Xin Li, xxl162230

Xiaoyu Zhang, xxz173130

1. Introduction

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. 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 problem is to complete the analysis of what sorts of people were likely to survive. In particular, apply the tools of machine learning to predict which passengers survived the tragedy.

Basic approaches include missing data imputation, feature engineering, logistic regression and support vector machine.

2. Problem Definition

Task Definition

The data which problem provides has been split into training and test set. For the test set, problem does not provide the ground truth for each passenger. It is our job to predict these outcomes. In data set, we have some features as below.

|  |  |
| --- | --- |
| **Variable** | **Definition** |
| survival | Survival |
| pclass | Ticket class |
| 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 |

3.Experimental Evaluation

3.1 Methodology

1.The flowchart of data processing is shown below.

START

DROP FEATURES: ‘PassengerId’, ‘Ticket’, ‘Cabin’

GET EACH PERSON’S TITLE FROM FEATURE ‘Name’

CREATE A NEW FEATURE ‘Title of person’

DROP FEATURE: ‘Name’

FILLING MISSING DATA IN ‘Age’

MAPPING ‘Title of person’

TRANSFORM FEATURES ‘Sex’, ‘Embarked’, ‘Title of person’ INTO CATEGORIAL

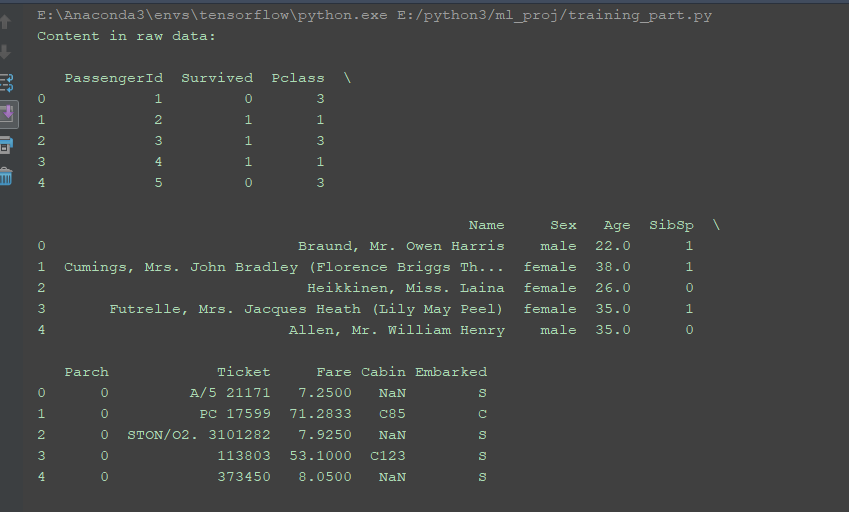
DROP DUMMY FEATURES

MERGE ‘SibSp’ AND ‘Parch’ INTO ‘Family size’ AND DROP

END

flowchart of data processing

2.In the first step, we use pandas.read\_csv to read training data set, see what features are included and what the type they are.

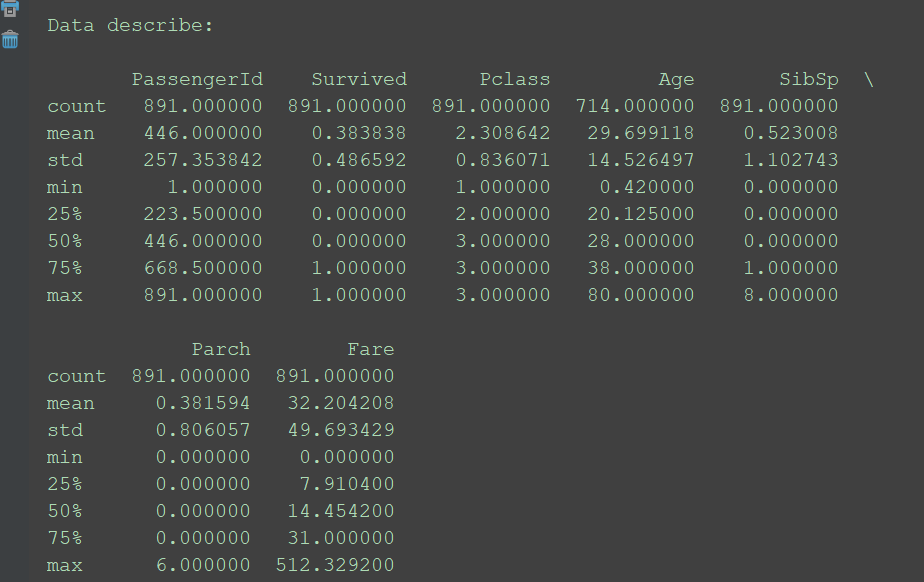


Overview of data set

From the result, we know:

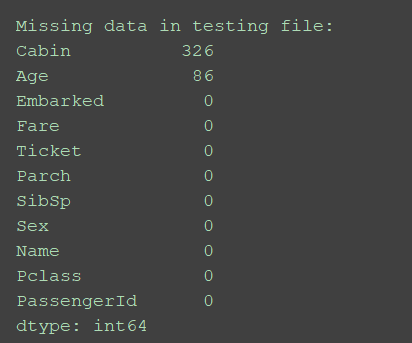
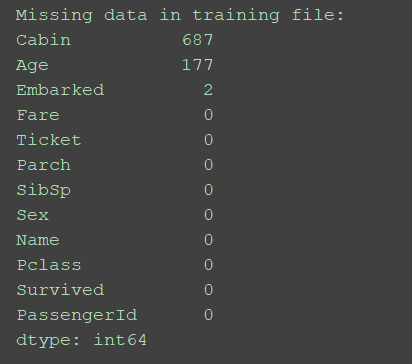
‘Sex’ and ‘Embarked’ features are categorical. Categorical feature should be encoded.

3.Then we use describe() function to generate the descriptive statistic to get the information about features of data set.



Descriptive statistic

4.To prepare the data, we need to deal with missing data. We use draw\_missing\_data\_table function to see the missing data. Sum up all missing date and show the number of missing data of each feature. The results are shown as below.



Missing data

From the result, we get:

(1)'Cabin' has too many missing values. We decide to delete this feature.

(2)'Age' can be imputed. For now, we'll use data\_proc(object) function in data\_processing step. Object varys 0,1,2,3,4,5 and each of them represent a different method to filling gaps in a feature. Later, we will revise this function.

5.Drop features

From above result, we drop ‘Cabin’ feature.

6.Based on our experience, some feature, like ‘Ticket’ and ‘PassengerId’, have no correlation with the outcomes. So we decide to delete these two features.

7.To complete the imputation of ‘Age’ missing data, our method is data\_proc(object) function. Different object represents different method as shown below.

'0' -- drop vacant value

'1' -- using mean value

'2' -- using median value

'3' -- using a previous value

'4' -- using a next value

'5' -- using the average age of corresponding title

When ‘object’ equals 5, method is to estimate the missing values based on known relationships. In this case, we can do this by using the information in the variable ‘Name’. Looking to 'Name' values, we can see person's name and title. Person's title is a relevant information to estimate ages. To give an example, we know that a person with the title 'Master' is someone under 13 years old, since ['a boy can be addressed as master only until age 12'](http://bit.ly/2HfFHZr). Therefore, employing the information in 'Name' we can improve our imputation method.

Method step are:

Extract titles from ‘Name’. Decrease title categories into 5(Mrs., Ms., Mr., Mrs., Other)

For each title, get people’s average age and use it to fill missing values.

8.Then transform feature ‘Sex’, ‘Embarked’, ‘Title’ into categorical. Change 'male' and 'female' in Sex into 1 and 0, 1 means 'male', 0 means 'female'. Delete ‘Embark\_C’ and keep ‘Embark\_Q’ and ‘Embark\_S’.

9.Drop dummy features. We merge ‘SibSp’ and ‘Parch’ into new feature ‘Family size’.Then delete ‘SibSp’, ‘Parch’ and ‘Name’ features.

So far, we finish the data processing step.

For training step, we plan to use two methods: Logistic Regression and SVM. Use two different machine learning algorithm and use different data processing methods (object varies 0,1,2,3,4,5) to train the model. We use cross-validation to get the accuracy of each case.

3.2 Results

Here are the accuracy of Logistic Regression and SVM when using different data processing method.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Object  Method | 0 | 1 | 2 | 3 | 4 | 5 |
| Logistic Regression | 0.8111 | 0.8212 | 0.8212 | 0.8212 | 0.8212 | 0.8156 |
| SVM | 0.8321 | 0.8268 | 0.8268 | 0.8268 | 0.8044 | 0.7932 |

4.Conclusion

After analysis of problem and compare the accuracy, using SVM method and data\_proc(object=0) which means drop vacant value is the best way to predict outcomes of test data set.

5.References

Problem link:

https://www.kaggle.com/c/titanic