2020年1月29日 星期三

Titanic summary

Flowchart

1. Introduction

2. Analyzing data

1）Data importing and brief analyzing

Import the package and datasets and have a brief idea of the features, like describe function, info, shape, head, columns and stuff. We can also check the missing values. (just check is ok. We can deal with the missing value later)

We can import the train set and test set at the same time and explore the datasets together. In this way, we don't need to explore the test after training the model.

2) Data analyzing and visualization

We can visual every feature and the relationships between the features and the "survived" feature. Describe the relationship between each feature and survived feature separately. If some features don't have obvious relationship with the 'Survived' feature, we should explain it, too.

3. Feature Engineering

1) Dealing with the missing values. As we mentioned before, some data are missing, so in this section we'll decide what to do with the missing data: either fill them with some predictions or just drop them.

2) Change feature's type. Some features may be string, so we should transform these features' type into numeric

3) Drop the irrelevant feature

In this part, we should also analyze every feature. In part 2's 2), we just visual the data, but not deal with data. So, in this part, we should deal with these data, like we can fill the missing values, transform type and decide which feature should be dropped

4. Modeling

1) splitting the train data

We will use part of our training data (20% in this case) to test the accuracy of our different models.

2) test different models

For every model, we calculate the accuracy. (from sklearn.metrics import accuracy_score)

5. Cross validation

In the previous part, the accuracy actually highly depends on the splitting of the training set (we can try different separation percentage, then results change a lot). To avoid that, we will use cross validation. Here, K-fold cross validation is used.