Bern University
of Applied Sciences

# Titanic-Matlab

Using Classification Learner in Matlab

BZG1308ab

Till Moser and Sebastian von Allmen

Version 1.0 of January 16, 2022

## Abstract

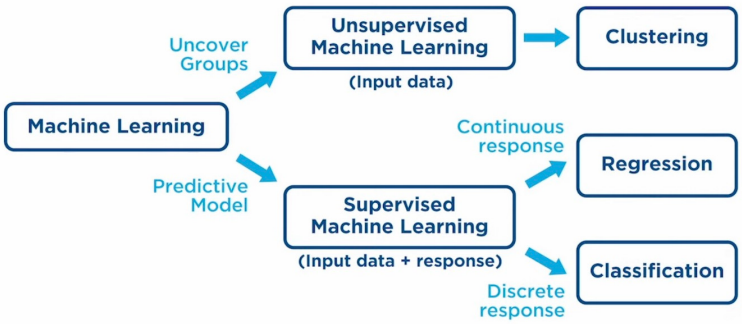
In this documentation we will give you a brief overview of how Machine Learning works and how you can use it in Matlab. More specifically we will talk about the classification learner in Matlab. This will give you a powerfull tool for data analysis and you can use the gained knowledge with the Regression Learner.
To do this we used a online Machine Learning competition. Kaggle-Titanic. It is a very popular competition for people who are trying to get into data science. Therefore there are a lot of online guides for it. We combined the knowledge we gained from them for this report. The goal of the competition is to guess who survived the sinking of the Titanic and who did not. To help you guess you get a train.csv file with information about the people on board.

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
| 1 | 0 | 3 | 'Braund, M... | male | 22 | 1 | 0 | 'A/5 21171' | 7.2500 | '' | S |

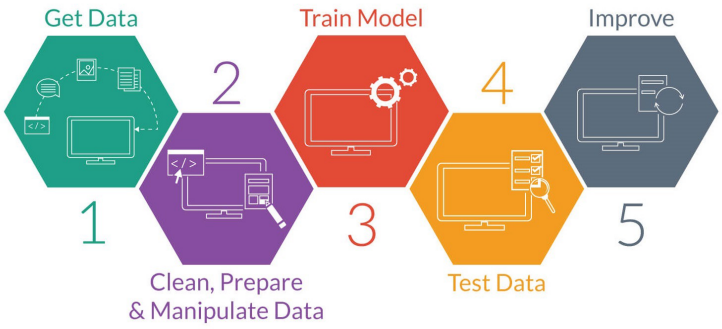
You use this data to train your model. After developing your model the true test comes. You run it using the test.csv file. This contains the same information but not whether or not the Passenger survived. To guess this you use your model. How accurate this guess is, shows how good your model was developed

## Introduction

You can group the Machine Learning the following ways.



We will focus on the classification of Supervised Learning. This means the output comes out to be in one class. I.e. survived or did not survive. It could also be Apple, Banana or Orange. A regression problem would f.e. be how much taxes you have to pay. I.e. 698.5 chf or 1045.7 chf. To do Supervised Machine Learning in Matlab we used the Classification Learner. Before diving in to that, we want to explain how the general Process of Machine Learning works. This graphic show the process.




## Get Data

First you have to get your data. This will most of the time be the easy part. In our case we just have to read in the data from the provided .csv File.

```
Test = readtable('test.csv','Format','%f%f%q%C%f%f%f%q%f%q%C');
```

## Clean Data

It is highly likely that you will have some blank spots in your data. To fill them is your responsibility. You will also likely have some data that is non relevant to solving your Problem. To remove them is also your job. In code this would look something like this.

```
avgAge = nanmean(Train.Age)
Train.Age(isnan(Train.Age)) = avgAge;
Test.Age(isnan(Test.Age)) = avgAge;
```

## Train Model

We train the Model using the Classification Learner from the Matlab Toolbox. You can read more about it in its own section.

## Test Data

To test the Data we have two options. We can test run it against the training data. This will give us a good estimate of how good the model performs. It should be pretty good since it is the same data we used to train the model. One problem with testing out models like this is overfitting. Our model could theoretically reach 100% accuracy agains our training data if it remembers for each passenger if he survived or not. The problem is that such a model would perform very badly oustside its trainign enviorement. if If we want to do a real test, we should run it against the test data. We do not know the outcome for this data. To check how good we were we can upload this data to Kaggle. It will tell us our success rate.

### Improve

Now we get to the interesting part. To improve our model we have two options. The first option is we find a better model and train it more. The other option is to clean, prepare and manipulate the input data in a smarter way, by including intelligent features and removing unimportant Data.

## Methology

To create your own Machine Learning Model you first have to learn the basics. We used this tutorial for that Matlab-Tutorial. This will teach you the basics and you will be able to develop your own models after completing it.

### Classification Learner

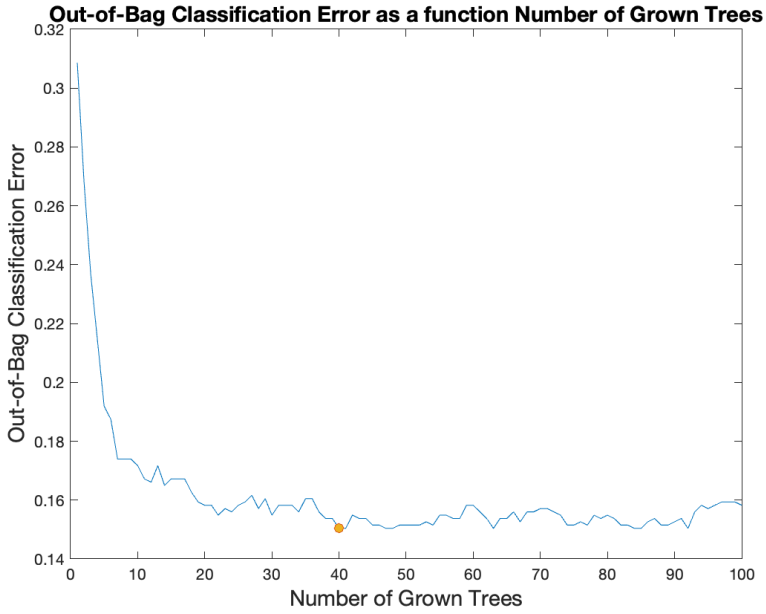
### Sebasmodell

In my Model I decided to use the boosted Tree Algorithm. It works like..... I tried to improve the performance of the Model by including some clever feature Datatypes and by improving how I filled out blank spaces.

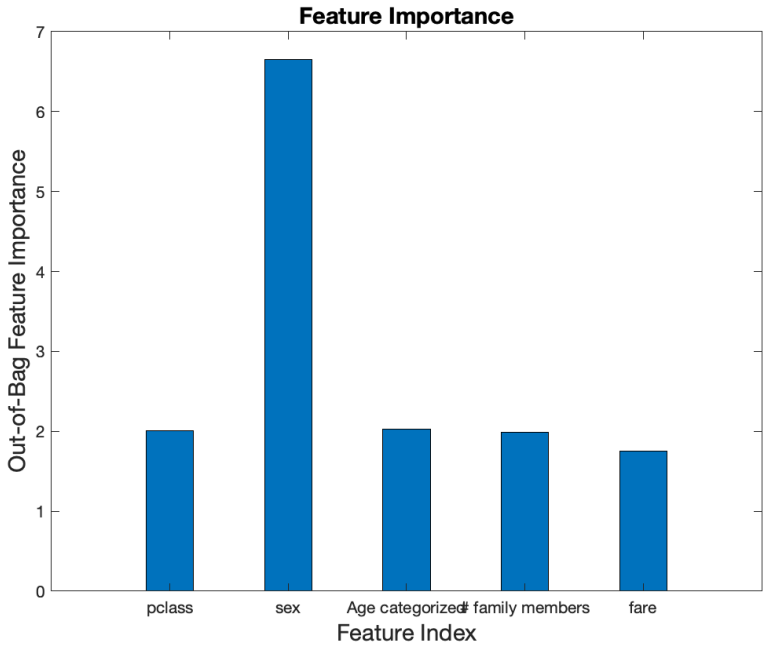
### Tillsmodel

Boosted trees are often prone to overfitting. This means, that an individual decision tree takes decisions witch work very good for the test data but sadly loose accuracy on the general problem. To combat this effect we can use an Random Forest Classifier. A random forest combines multiple deep decision trees into a forest which produces a result with a much lower variance.
For implementing this model in Matlab we can use the TreeBagger class from the Classification Learner toolbox.
For each person with an missing age value I decided to put in the median age of all other passengers. This will likely worsen the accuracy on children with missing age data but will be enough for the moment. The ages then were categorized into six groups.
The values for siblings and spouses ”sisps” and parent and child ”parch” were put together into an single variable family.

The variables: class, sex, categorized age, family, fare and survived were put together into a table with wich the classification learner was trained. First I plotted the Out-of-Bag Classification Error as a function of number of Grown trees in the random forest. Out of this visualisation i could read out the optimal forest size for our model. I choose the first minimal bevor the graph began to plateau.




Furthermore you can also show the importace of each category. You can use the "OOBPermutedDeltaError property for this. I find it a bit confusing that the age does not play a bigger role in this model.




The final steps where now to train a random forest model with the optimal size and evaluate the test data with it to commit the result onto the kaggle website for evaluation.
To train the model I used the Tree Bagger function from the classification learner toolbox. This would look a bit like this:

```
b_optimal = TreeBagger(optimal_forest_size,tbl(:,1:5),tbl(:,6),'
    OOBPredictorImportance','off','OOBPrediction','on');
```

## Results

In the end we developed two individuell models that worked better than primitive guesses. We are really happy with that. A primitive guess would be to say every woman survived and every man died.

| Method | Acuraccy in [%] |
|---|---|
| primitve | 74.2 |
| sebasmodel | 77.2 |
| tillmodel | 78,2 |

We think that Till's model was marginally better becuase he used the random forest function instead of the single decision tree function for his model. But the difference is so small becuase he had a better data preparation with his age average for sex and class.

## Discussion

The prediction of survival for each passenger of the Titanic was a very interesting project to learn the basics of data science and machine learning. It also provided a good framework to test the Classification Learner Toolbox in MatLab and use it on our own. Our Struggle was eased with an good online tutorial for our use case. We think without this tutorial it would have been virtually impossible to complete this task in a justifiable timeframe.
Our models could furthermore be optimized by combining them into a single model with also the data preparation combined. But in the end we think we can be proud of our achievements.