Contents

The goal of this paper is not to make predictions about the future. These results may teach us something about the circumstances during the time that the Titanic sank. Teaches us something about the civilization back in those days. (Women, children etc saved first?). Furthermore, this paper is written because I wanted to learn something about machine learning and programming using Python.

1 Introduction

In the year 1912 on the 15th of April one of the most infamous ships in history would crash into an iceberg and sink in the North Atlantic Ocean. During its maiden voyage from Southhampton to New York City on the 14th of April at 11:40 p.m. ship's time, the lookout sounded the alarm when a massive clump of solid ice caught his attention. The first mate had seen the iceberg before the lookout and tried to turn the ship around. Unfortunately, he was too late. Forty seconds later at a high speed the Titanic collided frontally with a huge rock made of ice with a weight of 30 million kilograms. The collision caused a series of holes along the side of the hull[1]. Six of the watertightcompartments were filled with water, whereas the ship could only carry on with a maximum of four compartments flooded. Consequently, the Titanic was doomed to sink. The crew understood they needed to act fast. They deployed the evacuation programm. The ship carried twenty lifeboats. In general the protocol "women and children first" was followed. However, this was not always the case. The chance of being saved was likewise dependent on the class in which one travelled and the place where one found itself during the evacuation. Around 2:20 p.m. parts of the Titanic broke off and sunk with one thousand people still on board.

[1] http://www.bbc.co.uk/history/titanic

On deck were some of the richest people in the world, including millionaires, silent movie stars, school teachers and emigrants, hoping to find a new life in New York City. A life that they would, therefore, never find. Two hours after the ship sank, the liner RMS Carpathia arrived and saved an estimated 705 survivors[2] The sinking of the RMS Titanic killed 1502 out of the 2224 people on board, crew members as well as passengers[3].

The RMS Titanic was the largest ship on water during that time and it was the second of three Olympic-class ocean liners operated by the White Star Line [4]. The ship consisted of nine decks, the boat deck, seven decks labelled from A to G which carried the passengers and the Orlop Deck which was below the waterline. The liner had a height of 175 feet and a breadth of 92 feet. Furthermore, the Titanic was designed and constructed in Belfast, Northern Ireland by Alexander Carlisle and Thomas Andrews[5]. *Insert image / map of Titanic with decks displayed*

The Titanic may be one of the most iconic ships in history, its story known the world over[6] . The tragedy has led to better safety regulations for ships and has inspired numerous expeditions, movies, books, plays and characters. *Insert image movie?.*

So many passengers have lost their life due to the fact that there were not enough lifeboats for everyone. Luck has played a part in surviving this disaster. Moreover, some groups had an advantage as opposed to other groups. For instance, the "women and children first" policy left a larger number of men aboard. In the same way as some people of the upperclass might have had a better chance at surviving as well.

In this paper we will take a look at what people were more likely to survive the demise of the Titanic with the help of machine learning. We will predict the chances of survival of certain groups of passengers. In addition to, we will see if the expectations that children, women and rich people were indeed benefited are correct.

The wreck of the Titanic

Dimensions of the Titanic.

[2] https://en.wikipedia.org/wiki/RMS_Titanic#Maiden_voyage (consulted 5th of August, 2018)

[3] https://www.kaggle.com/c/titanic

[4] https://en.wikipedia.org/wiki/RMS_Titanic#Maiden_voyage

[5] https://www.encyclopedia-titanica.org/titanic/ (consulted on 5th of August, 2018)

[6] http://www.bbc.co.uk/history/titanic

2 Different types of machine learning

Machine learning consists of giving computers the ability to learn and make decisions from data without being explicitly programmed. Using machine learning techniques to build predictive models.

A well-known example is detecting spam in your email inbox. Face recognition or teaching computers how to play chess, AlphaGo. A lot of different types of machine learning exist. Two examples will be discussed.

Unsupervised learning This is a version of machine learning where the computer has to uncover hidden patterns from unlabeled data.

For instance, grouping customers in categories based on buying behaviour without knowing in advance what these categories might be.

Supervised learning Where unsupervised learning has to make decisions from data that isn't labeled, supervised machine learning deals with labeled data.

Data points. These are samples described using predictor variables and a target variable. Organised in a table with rows and columns. The goal is to predict the target variable, in this case 1 or 0 representing survived or not survived respectively in our Titanic dataset, given the predictor variables.Such as / examples of our predictor variables: class, gender, age, siblings etc.

Two different types of supervised learning.

- Classification : target variable consists of categories.
- Regression : target variable is continuous.

Predicting survival on the Titanic is a classification problem. We have to classify, based on our predictor variables, if a person belongs to the class of survived (1) or not survived (0).Titanic using labelled data. More specifically historical data with labels. Data can be collected by experiments or crowd-sourcing.

Classification Titanic is a binary classification problem.

Goal is to learn from data for which the right output is known so we can make predictions on new data for which we don't know the output. In order to do this, we will use scikitlearn. This is a popular machine learning library for Python. Integrates well with numpy libraries.

Regression

Algorithms / programming Couple of libraries we will use:

- sklearn

- numpy
- pandas
- matplotlib.pyplot

Short description of each package.

Common used algorithm for classification problems is KNearestNeighbours. Predict label of a datapoint by looking at the 'k' closest labeled data points. Taking majority vote on what label an undecided point has to have. Creates a set of decision boundaries.

LogisticRegression

Other things I will not use, but are worth mentioning because they play a big part in the world of machine learning.

All machine learning models implemented as python classes.

- Implement algorithms for learning and predicting
- Store the information learned from the data.

Training a model on the data is called 'fitting' a model to the data using the .fit() method. Predict labels of new data using the .predict() method. Don't mention method. Explain what fitting is, error function. This is what you do working with Logistic Regression, not KnearestNeighbours.

At the end you can measure model performance. Want to know how well our model has performed. Metrics such as accuracy. Which data to use to compute accuracy, which is the fraction of correct predictions.

How well will model perform on new data that the algorithm has never seen before. Splitting of your dataset.

Fitting actually means that you tell your computer to find a curve that is as close to as many datapoints as possible. y = ax+b

$$y = ax + b \tag{1}$$

In this case there is only one predictor variable. But we have more than one predictor variable in our dataset of the Titanic. a and b are parameters of our model. We want to fit a line to the data. Our Titanic dataset has more dimensions. Our line will look something like this, where each x is a different predictor variable.

$$y = a_1x_1 + a_2x_2 + a_nx_n + b \tag{2}$$

We must specify a coefficient for each feature and a variable b. This is the fitting process.

Fitting consists of choosing your a and b. Define an error function for any given line. Choose the line that minimizes the error function / loss function. What is an error function? Explain.

Line has to be as close to the actual data points as possible. We have to calculate vertical distance between data point and the line. This is called the **residual**. Minimizing the sum of the residuals will not work because very large positive values will cancel out large negative values. Solution $\rightarrow$ minimize sum of the squares (lossfunction) of residuals. OLS = ordinary least squares. Same as minimizing the mean squared error of the predictions on training set. When you call fit on logistic regression model in scikitlearn, it performs this OLS under the hood.

2.1 Logistic Regression

In this paper we will use Logistic Regression as our algorithm. The name is misleading because logistic regression is commonly used for classification problems. Logreg outputs probabilities. If p is larger than 0,5 , classify as 1. $p < 0.5$, classify as 0 (not survived). Larger area under ROC curve = better model. Area is called AUC. Popular metric for classification models. AUC using cross validation. If AUC is greater than 0,5, the model is better than just random guessing.

`matplotlib`

```
1 import numpy as np
```

Choosing your parameters is called hyperparametertuning.

- Try different values
- Fit all of them separately
- See how well each performs
- Choose best performing one

Important to use crossvalidation! Otherwise, overfitting parameter.

1,2,3 - steps Introduction

1. Split dataset into a training set and test set, new dataset.
2. Fit/train classifier to the training set, what is fitting? Difference Knearest and Logistic

3. Predict on the test set

4. Print the prediction

5. Compare predictions with known labels

Test_size?

Perform your split so that your split reflects labels on your data. You want labels to be distributed as they are in the original dataset.

Problems KNearestNeighbours Overfitting: smaller k, more complex model, erratic pattern Underfitting: smoother decision boundary, larger k, less complex model. Generalizing too much, you use too little information.

Model performance is dependent on the way our data is split. Results are not reliable because of this. We solve this by using cross-validation. *insert image of folds*. Second fold as test set, fit on remaining data, predict on test set and compute metric of interest. 5-fold cross-validation. k-fold cross validation. More folds is more computationally expensive.

Measuring model performance using accuracy. This is a fraction of correctly classified samples. However, this is not always a useful metric. For instance, if we take a look at spam classification. 99% of your email is real and 1% is spam. We instantiate a classifier which classifies all emails as real. Computing the accuracy will give us a score of 99%, which is pretty high. But our classifier is horrible at predicting spam. **Class imbalance**. We have to use more nuanced metrics, such as the confusion matrix. *insert image of confusion matrix*. Accuracy, precision, recall, F1 score. High precision → not many real emails are predicted as spam. High recall → predicted most spam emails correctly. Confusion matrix in N dimensions?

- Underfitting and overfitting
- Train-test split
- Cross-validation
- GridSearch

Overfitting als je te veel variabelen toevoegt, LogisticRegression.

3 Preparation

3.1 A first look at the dataset

First we perform some numerical EDA. EDA stands for exploratory data analysis. This will help us explore our dataset and get a first impression of

the information. Not necessary to build a dataframa, for the information is already organised in a table.

code with describe etc

Next we perform some visual EDA. Scatter matrix, plotting, binary Seaborn's countplot. Possible correlation? Explain / describe diagrams.

3.2 Preprocessing techniques

How to deal with missing values, dummies, place of boarding, gender, cabin numbers. Map of Titanic? Need to encode categorical features numerically → convert to dummy variables. 0 = not that category.

Missing data

- NaN replace
- drop missing data
- impute missing data: make an educated guess

Centering and scaling

- Features on larger scales can unduly influence the model.
- We want features on a similar scale. **Normalizing**
- Standardization: substract the mean and divide by variance.
- Substract minimum and divide by the range
- Normalize so that data ranges from -1 to +1

We have to build a classifier that needs to learn from already labeled data. Training data = already labeled data.

Using GridSearchCV or RandomizedSearchCV, we can choose our parameters for KNearestNeighbours (K) and LogisticRegression (C). Large C kan lead to overfitting, small C kan lead to underfitting.

4 References