Titanic ML Project

Bogdanowicz Michal Kamil, Geraci Luca

May 9, 2020

Abstract

The abstract text goes here.

1 Proposition

The proposition is using a set of machine learning methods to predict the people that would survive the Titanic sinking of 15 April 1912.

The methods that have been used are:

1. Logistic Regression TODO
2. Random trees TODO
3. Neural Networks TODO

2 Data

2.1 Incomplete Data

The Data propsed to the professor was incomplete. It heavily reduced the avialble data to perform the best practices for evaluating a model. The test and training were already separated and ground truth was missing. So the complete dataset has been taken from the internet. In addition the complete list of surviors can be found at the wikipedia page here (not a ML-friendly format). This data sadly is used to cheat on the kaggle competition. Making it a quite infamous one.

2.2 Data Format

- Ticket class
- Survival
- Name
- Sex

- Age in years
- sibsp : number of siblings / spouses aboard the Titanic ¡– This might be problematic as it doesn't seem to make much sense.
- parch : number of parents / children aboard the Titanic¡– This might be problematic as it doesn't seem to make much sense.
- Ticket number
- Fare
- Cabin/s assigned
- Port of Embarkation
- Rescue Boat
- Body
- Destination

With additional notes of : Ticket class: A proxy for socio-economic status (SES) 1st = Upper 2nd = Middle 3rd = Lower

age: Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5

sibsp: The dataset defines family relations in this way Sibling = brother, sister, stepbrother, stepsister Spouse = husband, wife (mistresses and fiancs were ignored)

parch: The dataset defines family relations in this way Parent = mother, father Child = daughter, son, stepdaughter, stepson Some children travelled only with a nanny, therefore parch=0 for them.

2.3 Data Preparation

The data cannot be used as it is. The names are going to be deleted, as that should not inlfuence the resut.

3 Logistic Regression

This is the procedure that has been used to select the model.

1. Choose the maximum polynomial to use by comparing the measures with CV (Cross Validation)

   This

2. Check for Bias and variance by visualizing the Error in CV and Error on Training set on a graph with X as the degree of polynomial

3. Choose the regularization parameter that reducces the CV Error the most by using the polynomial chose and Lambda going from 0.01 to 5.

4 Conclusion

Write your conclusion here.