Titanic - Machine Learning from Disaster

Project Proposal for Machine Learning(91.545) at University of Massachusetts, Lowell

Kunal Vyas
kunal_vyas@student.uml.edu

Zeshi Zheng
zeshi_zheng@student.uml.edu

Lin Li
lin_li@student.uml.edu

1 Team Members and Roles

Kunal Vyas - Problem Analysis, Research, Code implementation, Dataset Analysis, Testing and Debugging.
Zeshi Zheng - Dataset Analysis, Code Implementations, Visualization development, Model Building, Algorithm Development, Testing and Debugging.
Lin Li - Problem Analysis, Code Implementations, Visualization development, Algorithm Development.

2 Goal

RMS Titanic sinking, known as one of the deadliest commercial peacetime maritime disaster in modern history, happened after colliding with an iceberg. Only 706 passengers and crew out of 1502 lost their precious lives in the infamous sinking. Among the surviving people, women, children and upper class had larger probability to survive.
We aim to analyse what sort of people were likely to survive. In particular, we will apply the tools of machine learning to predict which passengers survived the tragedy.

3 Motivation

Learning answers to similar problems can provide insights into the cause of those events happened and what we can do to avoid the wrongdoings, if any, in the future. Thus, knowing this problem's solution and successful techniques to be used to solve it will be very valuable to the field of machine learning.
There are existing implementations that serve the same purpose:

3.1 CS229 Titanic – Machine Learning From Disaster

Eric Lang and Chongxuan Tang discuss three approaches to solve this problem in their paper. In particular, they have compared different machine learning techniques like Naïve Bayes, SVM, and decision tree analysis.

Conclusion: Although this and many more such implementations exist that have solved the problem with many approaches, they are not accurate enough and we aim to improve that in our approach.

4 Approach

We will process the dataset and make a prediction by spiral development, which means we will build a sequence of rough primitives at the very beginning by adding a few features or functions. Then, we increase more features and improve the whole program gradually. At each iteration, we analyze the problem by following respects:

- Pick features: we discuss and add a few number of features to our model, meanwhile, delete some worthless features from our training set;

- Program preparation: prepare for incoming features, and do some changes for last construction;
- Build model: decide if our old equation/model need to update, and make changes;
- Train: combine our model and new set of examples;
- Predict: judge if our model is overfitting or underfitting.

We will develop the machine learning algorithms in Python language and we might also use JavaScript to develop visualizations, among other tools and utilities to process data.

5 Dataset

Our dataset is provided by Kaggle, where all of historical data we get is used for training and build models. In the training set, it provides the outcome ('ground truth') for each passenger, that is, for each passenger in the test set, we will predict whether or not they survived the sinking (0 for deceased, 1 for survived) by a bunch of features, these are:

survival	Survival	(0 = No; 1 = Yes)
pclass	Passenger Class	(1 = 1st; 2 = 2nd; 3 = 3rd)
name	Name	
sex	Sex	
age	Age	
sibsp	Number of Siblings/Spouses Aboard	
parch	Number of Parents/Children Aboard	
ticket	Ticket Number	
fare	Passenger Fare	
cabin	Cabin	
embarked	Port of Embarkation	(C = Cherbourg; Q = Queenstown; S = Southampton)

6 Expected Outcome

Classifiers are created based on passengers' gender, age and passenger class. After the creation, we can predict if a certain passenger could be survived or not in the sinking.