Proposal: Surviving the Titanic

Pacific Lutheran University
Miguel Amezola, Nicholas Glover, and Quinton Teas

Abstract. We will use passenger data from the sinking of the RMS Titanic to predict survival. Creating a feature space we give us the opportunity to experiment with feature generation. Implementing C4.5, a decision tree algorithm for continuous features and binary classification, will deepen our understanding of machine learning models. Benchmarking with a bagging meta-estimator from scikit-learn will not only allow us to compare our implementation with another, but also give us the opportunity to familiarize ourself with a robust machine learning library

1. Introduction

The RMS Titanic sank on 15 April 1912, after colliding with an iceberg during its maiden voyage from Southampton to New York. Out of 2224 passengers and crew, 1502 lost their lives. Interestingly, some groups of people were more likely to survive, such as women, children, and the upper class. We would like to identify more factors that improved the likelihood of survival. Furthermore, we will map these factors to an appropriate feature space. And finally, we will implement and train a Decision Tree to analyze what sorts of people were likely to survive and compare its accuracy with that of a bagging-meta estimator from sciki-learn.

2. Implementation

For the passengers on the Titanic, there were only two outcomes, namely survive and not survive. Thus, they can be grouped into two disjoint sets or classes. We will use a decision tree to predict the class two which each passenger belongs.

2.1. C4.5 Algorithm. We will generate the decision tree using the C4.5 algorithm — an ID3 extension developed by Ross Quinton. Like its predecessor, C4.5 uses information entropy to perform recursive binary partitioning of a given feature space.

Definition 1 (Entropy [**3**]). Let S be a dataset, let X be the set of classes in S, and let $p(x)$ be the proportion of the number of elements in class $x \in X$ to the number of elements in S. We define $p(x)\log_2 p(x) := 0$ if $p(x) = 0$. Then **entropy** is the function $H : S \to \mathbb{R}$

defined by

$$H := -\sum_{x \in X} p(x) \log_2 p(x).$$

DEFINITION 2 (Information gain). Let $H(S)$ be the entropy of set S, let T be a collection of subsets created by partitioning S by feature F such that $S = \bigcup_{t \in T} t$, let $p(t)$ be the proportion of the number of elements in class $t \in T$ to the number of elements in S, and let $H(t)$ be the entropy of t. Then **information gain** is the function $IG : F \times S \to \mathbb{R}$ defined by

$$IG(F, S) := H(S) - \sum_{t \in T} p(t) H(t).$$

C4.5 uses a set of training data, a set of classified samples $S = s_1, s_2, \ldots$, to build decision trees. Each sample $s_i \in \mathbb{R}^n$, $s_i = (x_1, x_2, \ldots, s_{n-1}, y)$, where each x_i represents a feature of s_i and y represents the class label for s_i

At each node of the tree, C4.5 chooses the feature that most effectively partitions S into subsets $S_1, S_2, \ldots S_m \subset S$. This is done using the concept of information gain; that is, the attribute with the largest normalized information gain is used to partition S. This is done recursively until for each subset S_i of S,

$$s_1, s_2 \in S_i \wedge s_1 \neq s_2 \implies y_1 = y_2$$

where y_1 is the y entry in s_1 and y_2 is the y entry in s_2.

2.2. Learning Model. There are many decision-tree algorithms. Notable ones include ID3 (Iterative Dichotomiser 3) and C4.5 (successor to ID3) [**2**]. Since C4.5 made a number of important improvements to ID3, like the ability to handle both discreet and continuous attributes and allowing attributes to be marked ? for missing, we choose this algorithm.

2.3. Work Allocation. This implementation will require several subtasks, namely

(1) creating data structures for representing datasets and feature spaces,
(2) implementing the C4.5 algorithm, and
(3) training/testing the bagging meta-estimator.

Each task is assigned to a different group member: (1) Miguel, (2) Nicholas, and (3) Quinton.

3. Method

3.1. Data. The data have been partitioned into two disjoint subsets, E and F, such that the cardinality of F is about one half of the cardinality of E; that is, $|F| \approx \frac{1}{2}|E|$. As demonstrated in Table 2, the data have ten variables for each passenger. Note that **pclass** is a proxy for the socio-economic status of the passenger. Also, **age** is fractional if less than one, or in the form of $xx.5$ if the age was estimated. For **sibsp**, a sibling is defined as a brother, sister, stepbrother, stepsister, and a spouse is defined as a husband, wife (mistresses

TABLE 1. Tentative Schedule

Start Date	**Duration**	**Task**
April 10 or April 12	N/A	Project Proposal Presentation
April 3	1 week	Generate features
April 10	2 week	Implement model
April 24	1 week	Train/Test
May 1	1 week	Predict
May 8	1 week	Prepare final presentation
May 8	2 weeks	Prepare report
May 15, 17, or 19	N/A	Final project presentation
May 24	N/A	Report due

and fiancés were ignored). Similarly for **parch**, a parent is a mother or father and a child is a daughter, son, stepdaughter, or stepson. If children traveled with a nanny, then parch = 0 for them [**1**].

TABLE 2. Data Dictionary

Variable	**Definition**
survival	Survival
pclass	Ticket class
sex	Sex
age	Age in years
sibsp	number of siblings or spouses aboard the Titanic
parch	number of parents or children aboard the Titanic
ticket	Ticket number
fare	Passenger fare
cabin	Cabin number
embarked	Port of Embarkation

3.2. Train. The larger subset of data E will be used to build the machine learning model. We will also fit this training set to the bagging meta-estimator from scikit-learn. This bagging classifier in turn fits an ensemble of classification trees, each on random subsets of the original set. Such an estimator should reduce the variance that plagues decision trees, thereby outperforming our implementation of the C4.5 algorithm.

3.3. Test. We will use dataset F to see how well the model performs on unseen data. As with training, we will also test the bagging estimator with this dataset.

3.4. Predict. Predictions from our C4.5 implementation and the bagging meta-estimator will be recorded in distinct .csv files. Each file will have exactly 418 entries and a header row, as in Listing 1. Each file will have exactly 2 columns:

- PassengerId (sorted in any order), and
- Survived (contains binary predictions: 1 for survived, 0 for deceased).

```
PassengerId,Survived
892,0
893,1
894,0
Etc.
```

Listing 1. Prediction example

We will use the following metric to see how well our implementation of the C4.5 algorithm performs.

Definition 3 (Accuracy). Let $c, i \in \mathbb{N}$ such that $c + i \neq 0$ with c equal to the number of correct predictions and i equal to the number of incorrect predictions, and let $f : \mathbb{N} \to [0, 1] \cap \mathbb{R}$ be the function defined by

$$f(c, i) := \frac{c}{c + i}.$$

This function is the statistical measure commonly known as accuracy.

As for the bagging meta-estimator, scikit-learn provides a method for computing the mean accuracy on the given test data and labels.

4. Conclusion

Not only is the sinking of the RMS Titanic an interesting historical event, but an opportunity to design a machine learning model and test its performance. Since we are familiar with ID3, implementing its successor C4.5 will allow us to build on previous knowledge. The hands-on experience stemming from making our own implementation will deepen our understanding of machine learning models.

References

[1] Titanic: Machine learning from disaster. Web, April 2017. `https://www.kaggle.com/c/titanic`.

[2] Wikipedia contributors. Decision tree learning. Wikipedia, The Free Encyclopedia, April 2017. Retrieved 18:34, April 5, 2017, from `https://en.wikipedia.org/w/index.php?title=Decision_tree_learning&oldid=773803391`.

[3] C.E. Shannon. A mathematical theory of communication. University of Illinois Press, 1949.