CS229 Titanic – Machine Learning From Disaster

Eric Lam
Stanford University

Chongxuan Tang
Stanford University

Abstract

In this project, we see how we can use machine-learning techniques to predict survivors of the Titanic. With a dataset of 891 individuals containing features like sex, age, and class, we attempt to predict the survivors of a small test group of 418. In particular, compare different machine learning techniques like Naïve Bayes, SVM, and decision tree analysis.

1. Introduction

Using data provided by www.kaggle.com, our goal is to apply machine-learning techniques to successfully predict which passengers survived the sinking of the Titanic. Features like ticket price, age, sex, and class will be used to make the predictions.

We take several approaches to this problem in order to compare and contrast the different machine learning techniques. By looking at the results of each technique we can make some insights about the problem. The methods used in the project include Naïve Bayes, SVM, and decision tree. Using these methods, we try to predict the survival of passengers using different combinations of features.

The challenge boils down to a classification problem given a set of features. One way to make predictions would be to use Naïve Bayes [1]. Another would be to use SVM to map our features to a higher dimensional space. Our approach will be to first use Naïve Bayes as a baseline measure of what is achievable. Once this is complete, we use SVM [2] on our data to see if we can achieve better results. Lastly we use decision tree analysis [3] and find the optimal decision boundaries.

2. Data Set

The data we used for our project was provided on the Kaggle website. We were given 891 passenger samples for our training set and their associated labels of whether or not the passenger survived. For each passenger, we were given his/her passenger class, name, sex, age, number of siblings/spouses aboard, number of parents/children aboard, ticket number, fare, cabin embarked, and port of embarkation. For the test data, we had 418 samples in the same format.

The dataset is not complete, meaning that for several samples, one or many of fields were not available and marked empty (especially in the latter fields – age, fare, cabin, and port). However, all sample points contained at least information about gender and passenger class. To normalize the data, we replace missing values with the mean of the remaining data set or other values.

3. Data Analysis

In order to prepare our data for training in our Naïve Bayes classifier, we remove or replace blank values and determine bin sizes for each feature. For instance, fare can range a large number of values, so we group fares together. The same is done for cabin data since the data is divided into cabin sections (A,B,C,D). A similar grouping is done to the ticket data.

For our SVM, we do not need to bin values together. Instead we simply turn all the values into numerical values. In order to do this, we'll interpret the bit representation of strings and characters as float represented numbers.

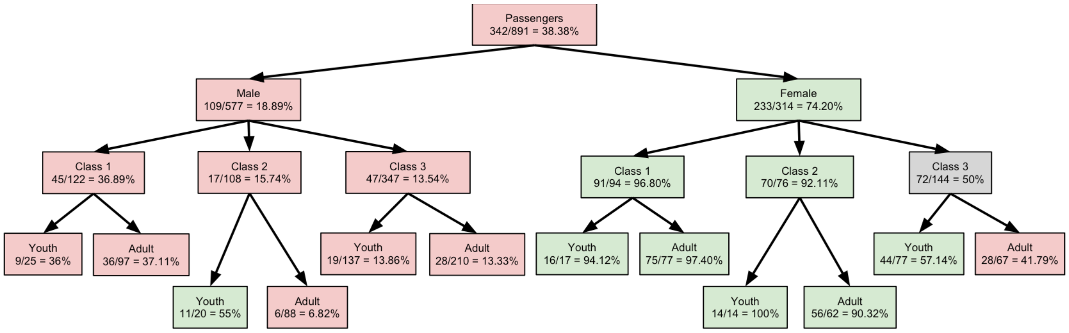

Figure 1. Data breakdown by sex, class and age. Percentages are percent that survived. Red boxes indicate mostly died, green boxes indicate mostly survived, and grey indicates 50% survival

In figure 1. of the previous page, we see the breakdown of the data to get a better sense of what features might be good indicators of our classification problem. First, we notice that out of all the passengers in the test data, 36.38% survived. If we breakdown the group into sex, we see that a significant difference in survival between females (74.20%) and males (18.89%). This is a strong indicator that sex would probably be a good feature to use. Continuing our analysis, we see that of the females in first class and second class (first class can be thought of as upper class, second class as middle class, and third class as lower class), more than 90% survived. Third class fared much worse with a 50% survival rate. Of the males, first class had a much higher survival rate (36.89%) than second (15.74%) or third (13.54%). Interestingly, there was not significant variation in survival given a person's age in any subgroup except for youths in second class.

4. Approach/Method

Basic Naïve Bayes classification in [1] is used as a baseline to see what is achievable. More sophisticated techniques like SVM in [2] and decision tree analysis is used [3] to see if improvements can be made in the classification test. We experimented with using different feature sets of each method and found the optimal feature combination on the test group.

5. Naïve Bayes

As a benchmark, we first implemented Naïve Bayes [1]. For the Naïve Bayes model, we considered the following features: 1) sex, 2) passenger class, 3) age, and 4) fare. We chose to use the multinomial event model and Laplace smoothing. First we had to find *P(died)* and *P(survived)* by tallying up the number of passengers that survived and dividing by the total number of samples. Both gender and passenger class took on discrete values. Gender has two values, male and female, and passenger class has three values, 1st, 2nd, and 3rd class. Next, we had to estimate the conditional probabilities of these features given whether a passenger survived. For example, to find $P(male|survived)$ we had to count the number of male survivors and divide that by the total number of survivors. The same can be done for other features and values.

Before computing the parameter estimates for the features age and fare, they first need to be discretized. For age, we grouped the ages into buckets of size 5. We used a default value of -1 for samples for which age information was not provided. Similarly, for fare, we grouped the fare prices into buckets of size 20 and used -1 for samples with no fare information. After discretizing age and fare, we found the estimate for conditional probabilities of those features given whether a passenger died or survived the same way as before.

The conditional probability computed from the training set is given by

$$p(F = f|S = s) = \frac{\sum_j^n 1\{f_j = f, s_j = s\}}{n}$$

Where f is a particular feature and s is survival. We iterate through all the training examples $j.$ This gives us a conditional probability distribution based on survival. Using this conditional probability distribution, we can compute the probability that a test point survives given the feature set. We use the maximum a posteriori or MAP decision rule as shown below.

$$classify(f_1, .. f_n) = \arg\max_s p(S = s)\Pi_i^n p(F_i = f_i|S = s)$$

Where f_i are our features, s is true if survived and false if not survived. We multiply the probability of each feature given a negative outcome and compare that with the probability of each feature given a positive outcome. We make a prediction based on which probability is greater.

Table 1. Naïve Bayes Accuracy – Using Different Features

Pclass	Sex	Age	Fare	Accuracy
Yes	Yes	Yes	Yes	76.79%
Yes	Yes	No	Yes	75.36%
No	Yes	Yes	Yes	74.40%
Yes	No	Yes	Yes	65.79%
No	Yes	No	No	76.79%

We tried several different combinations of features to use, shown in the table above. And the gender of the passenger seemed to be the strongest indicator of whether a he/she survived. Adding the other three features to our Naïve Bayes model did not improve the prediction accuracy. This is probably because gender has the most correlation to survival and dominates the Naïve Bayes classifier. However, without considering the gender feature, the addition of other features did improve the performance of the classifier, which shows that although the other features are not as strong an indicator of survival, they still do have a small correlation to a passenger's chance of survival.

6. SVM

To improve our classification, we used support vector machines [2]. We considered the following features: 1) passenger class, 2) sex, 3) age, 4) number of siblings, 5) patriarchal status, 6) fare, and 7) place of embarkation. We used a Gaussian radial basis function as our kernel and set the tolerance to $\varepsilon = .001$.

Table 2. SVM Accuracy – Using Different Features

Pclass	Sex	Age	Sibsp	Parch	Fare	Embarked	Accuracy
Yes	Yes	No	No	No	No	Yes	77.99%
No	Yes	Yes	Yes	Yes	No	No	74.88%
No	Yes	No	Yes	No	Yes	No	73.21%
No	No	No	No	Yes	Yes	Yes	67.70%
No	No	Yes	No	Yes	No	Yes	64.83%
No	Yes	Yes	Yes	No	Yes	Yes	61.96%
No	No	Yes	No	No	Yes	Yes	58.13%

Unlike Naïve Bayes, no extra data cleaning was needed. Iterating through all possible feature combinations, we were able to achieve an accuracy rate of 77.99% on the test data set using only three features. The three features that achieve this rate were class, sex and place of embarkation. Using age, fare, and place of embarkation resulted in the worst accuracy of 58.13%. It is interesting to note that this accuracy would be less if we had just guessed that all test points died (accuracy of 63.23%). This suggests that perhaps class and sex are strong indicators of survival whereas age and fare are weaker indicators of survival. In figure 2, we see the SVM learning curve using the features class, sex and place of embarkation. At around 400 samples, the training curve has reached its asymptotic value of 77.99% and any additional sample does not improve the accuracy.

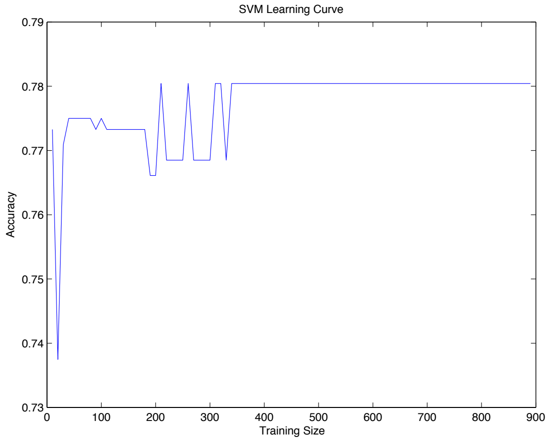

Figure 2. SVM Learning curve using class, sex, and place of embarkation

7. Decision Tree

We built our decision using the following features -- gender, passenger class, age, and fare. We first split the data into males and females because it was most correlated with the chance of survival. From just using a single feature, we achieved an accuracy of 76.79%, which is the same percentage as Naïve Bayes with just the gender feature. This is expected since with just the gender feature, both classifiers are labeling test samples the same way (which is marking all females as survived and all males as died).

Then we split both males and females into passenger classes. Even after splitting the data into passenger class, males in each class are more likely to die, and passengers in each class, other than class 3, are more likely to survive. If we choose the hard decision that female passengers in class 3 all survive, it will still produce an accuracy of 76.79% because the classifier hasn't changed from the earlier process of labeling all males as died and females as survived. However, if we choose the hard decision that all females in class 3 will die, our accuracy improves to 77.27% on the test data.

Next, we look at the feature age. Since the domain of age is continuous, we have to find a good decision boundary to split our data. After plotting the age and survival of passengers in each gender and passenger class, we decided to use a binary decision because in most cases, older passengers were more likely to die than younger ones. Instead of using the same age boundary for each gender and passenger class, we considered each gender and passenger class, case by case and found different boundary thresholds for each. To find our boundary threshold, we tried to minimize the classification error on our training set. This means that we chose the age boundary for each gender and passenger class such that if we classify all samples below the age boundary as survived and all above as died, we minimize the classification error on the training set. After including age in the decision tree, we achieve a classification error of 78.94%.

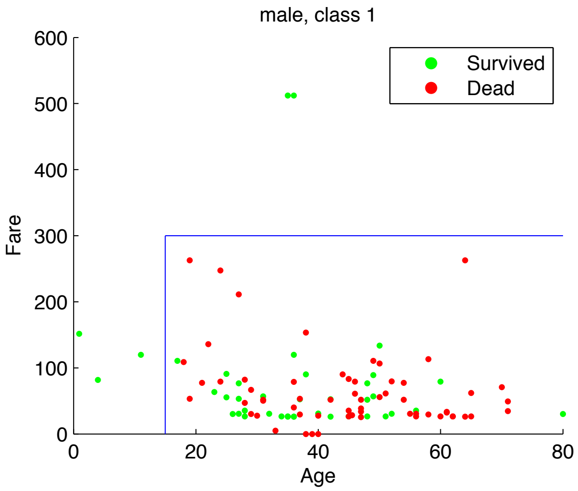

Figure 3. Optimal decision boundaries for the subgroup male, class 1

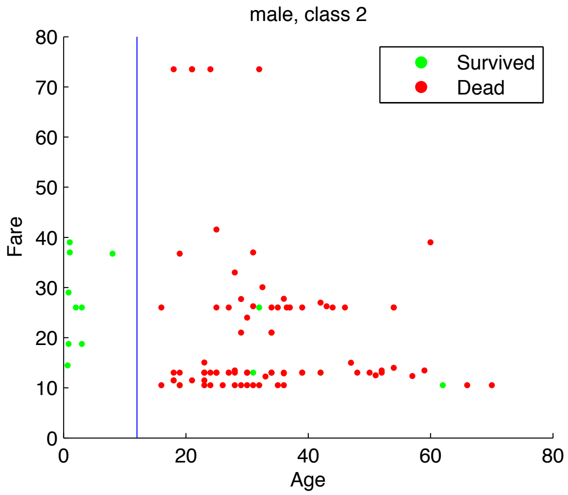

Figure 4. Optimal decision boundaries for the subgroup male, class 2

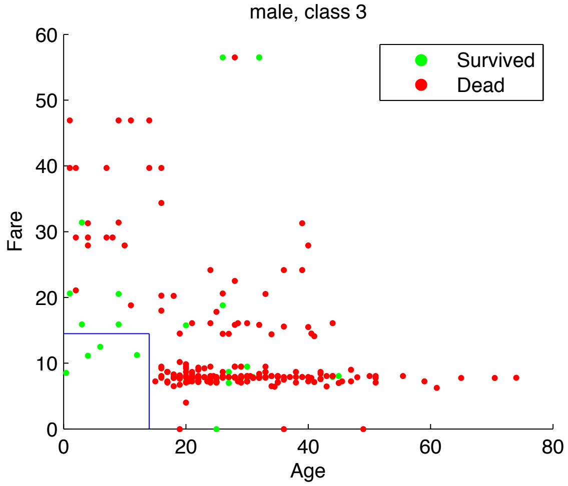

Figure 5. Optimal decision boundaries for the subgroup male, class 3

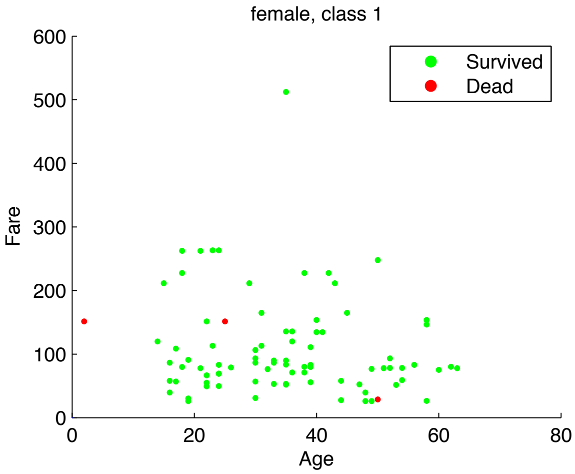

Figure 6. Fare vs. Age and survival in the subgroup female, class 1

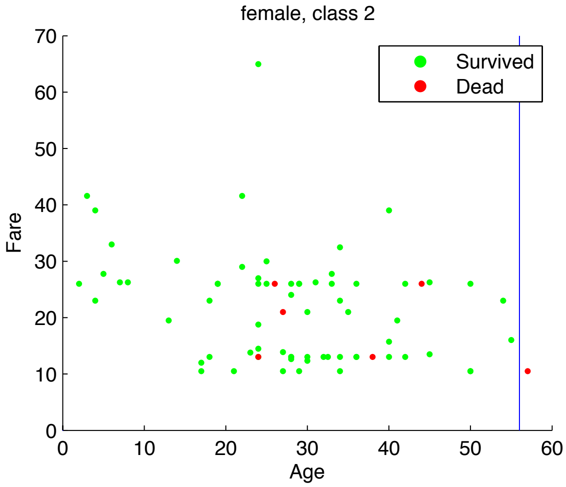

Figure 7. Optimal decision boundaries for the subgroup female, class 2

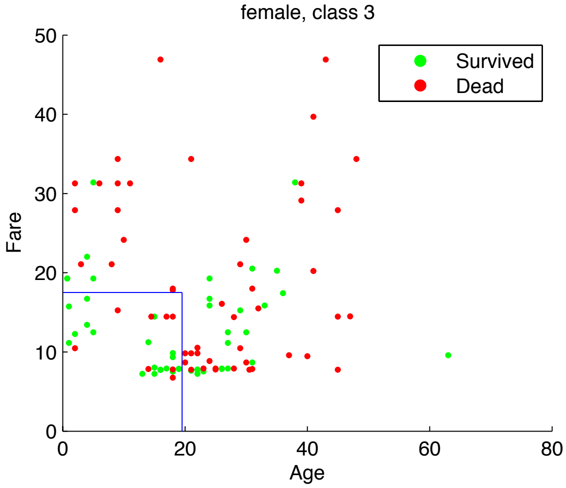

Figure 8. Optimal decision boundaries for the subgroup female, class 3

The fare feature itself is not well correlated with survival rate so we decided to consider fare together with age. For males in class 1, the training data suggests that young passengers and passengers who pay a high fare are most likely to survive. For both males and females in class 3, the training data suggests that passengers who are both young and paid low fare are most likely to survive. For passengers in class 2 and females in class 1, fare didn't seem to be a factor in determining survival rate. So after finding the boundary thresholds to use for male in class 1 and passengers in class 2, that minimizes the error in the training set, we achieved a prediction accuracy of 79.43% on the test data.

We also tried separating the data based on other based such as the number of siblings/spouses and the number of parents/children aboard. However, these features did not provide good insight into survival rate.

8. Results

Of the three methods, Naïve Bayes performed the worst and decision tree performed the best. However, the best and worst performance only differs by 2.64% so all the methods have roughly the same performance on our data set. This is probably because there was one feature that was strongly correlated with whether a passenger survives. The Naïve Bayes model assumes that all features are independent but the decision tree does not make this assumption. Even though the decision tree considers correlation between features, it only performs marginally better than Naïve Bayes. So this shows that assuming that features are independent is not necessarily a bad assumption for our problem. Table 3 offers a summary of the achievable accuracy using Naïve Bayes, SVM, and decision tree analysis.

Even though we were given many features of passengers in our data, we found that most of the features were not useful in classification. For example, the number of sibling/spouses and the number of parents/children did not help with classification in any of the three models. Knowing the number of relatives aboard did not help with classification, but perhaps, if we were given the links between passengers then we'd be able to infer more about the survival rate. Since family units tend to all die or all survive, knowing the family links would have been useful.

Table 3. Comparison of Performance

Naïve Bayes	76.79% Accuracy
SVM	77.99% Accuracy
Decision Tree	79.43% Accuracy

9. Conclusion/Future Work

There were not significant differences in accuracy between the three methods we experimented with. Even using every combination of features, we were still not able to produce an accuracy rate that was much different than simple Naïve Bayes classifier using only sex as a feature. It appears that the other features were only weakly indicative of survival, as sex seemed to dominate the others in terms of being able to accurately predict survival. Even with more sophisticated algorithms, we were not able to achieve much improvement. This shows the importance of choosing important features and obtaining good data.

It would be interesting to continue this analysis with other possible features or with other machine learning algorithms like random forests or other variants of Naïve Bayes.

10. References

[1] A. Ng. CS229 Notes. Stanford University, 2012

[2] Cortes, Corinna; and Vapnik, Vladimir N.; "Support-Vector Networks", Machine Learning, 20, 1995.

[3] Stuart J. Russell , Peter Norvig, Artificial Intelligence: A Modern Approach, Pearson Education, 2003 pg 697-702

Future Distribution Permission