SVM algorithm to predict Titanic survivors.

Alexandre Cazé

September 2, 2015

Abstract

Given data on 891 passengers of the Titanic (age, class, sex, ...), the Kaggle tutorial "Titanic: Machine Learning from Disaster" (`http://www.kaggle.com`) asks to predict the survival of 418 other passengers. In this note, we present our results using the Support Vector Machine (SVM) algorithm LinearSVC provided by the library sklearn in python. After a raw overview on the data, we only take into account two features, that are Gender and Class. We show that this algorithm leads to a model where all women survive and all men die, independently on the class. This leads to a prediction efficiency of 0.76555 in the Kaggle leaderboard, which corresponds to the benchmark "Gender based model".

1 Overlook on the data and first conclusions

1.1 Features

We restrict ourself to three features of the data: Gender, Class and Adult.

- Gender: Male or Female.
- Class: 1st, 2nd or 3rd class.
- Adult: Adult (over 18) or Child (under 18).

1.2 Statistics

Some first statistics describing the data are presented in the following tables:

- Table 1 sums up the general statistics of the training and test sets.
- Table 2 sums up the survival rates depending on each feature individually.
- Table 3 describes the interaction between the features Gender and Adult for the survival rate.
- Table 4 describes the interaction between the features Gender and Class for the survival rate.
- Table 5 describes the interaction between the features Adult and Class for the survival rate.

	Training set	Test set
# passengers	891	418
Adults	84.4 %	87.1 %
Children	15.6 %	12.9 %
Male	64.8 %	63.6 %
Female	35.2 %	36.4 %
1st class	24.2 %	25.6 %
2nd class	20.7 %	22.2 %
3rd class	55.1 %	52.2 %

Table 1: Comparison Training set / Test set.

Total	0.384 %
Adults	0.362 %
Children	0.504 %
Male	0.189 %
Female	0.742 %
1st class	0.63 %
2nd class	0.473 %
3rd class	0.242 %

Table 2: Survival rates for each feature individually.

	Male	Female	Total
Adults	16.8 %	76.0 %	36.2 %
Children	33.8 %	67.6 %	50.4 %
Total	18.9 %	74.2 %	38.4 %

Table 3: Survival rates : Interaction Gender / Adult.

	Male	Female	Total
1st class	36.9 %	96.8 %	63.0 %
2nd class	15.7 %	92.1 %	47.3 %
3rd class	13.5 %	50.0 %	24.2 %
Total	18.9 %	74.2 %	38.4 %

Table 4: Survival rates : Interaction Gender / Class.

	Adult	Child	Total
1st class	61.0 %	87.5 %	63.0 %
2nd class	41.3 %	79.3 %	47.3 %
3rd class	21.7 %	35.1 %	24.2 %
Total	36.2 %	50.4 %	38.4 %

Table 5: Survival rates : Interaction Adult / Class.

1.3 First conclusions

From those statistics, one can draw some first raw conclusions:

- From Table 2, one can predict that the importance of the features will be in descending order Gender, then Class and finally Adult. The moto "Ladies and Children first" seems to have worked extremely well for female passengers, not so well for children.
- The low number of children that appears in Table 1 may weaken any conclusion for the feature Adult.
- From Table 4, one can see that being a woman in first or second class was nearly the insurance to survive, while being a man was bad news in any class (although 1st class were not that desperate).

Given those first conclusions, we focus on the model that should be the most relevant, that is the Gender/Class model.

2 SVM for the Gender/Class model

Figure 1 shows the result of the LinearSVC algorithm from the library svm of sklearn (`http://scikit-learn.org/stable/modules/svm.html`). The Python code that generates this figure is available on Github at `https://github.com/alexandrecaze/kaggle-titanic` (file svm.py).

In Fig. 1, we have represented the 6 possible cases corresponding to the Gender-Class model (Male and 1st class, Male and 2nd class, ...) and the associated prediction according to the LinearSVC algorithm trained on the training set described in Table 1. Survivors are represented in red and non-survivors in blue. Note that this problem is not linearly separable, which implies non-trivial calculations in the black-box LinearSVC (for details, see e.g. Ref. [1]) in order to find the line that minimises the error, represented in blue. As a conclusion, one can see that the best model in that framework ignores the feature Class and only predicts that women survive and men die. This leads to a prediction efficiency of 76.555% in the Kaggle leaderboard, calculated on the test set. Although it might sound like a trivial result, it is elegant to obtain a mathematical justification for it (it is a close call for the women of 3rd class to be predicted as survivors, another training set using the same algorithm could have led to a different model).

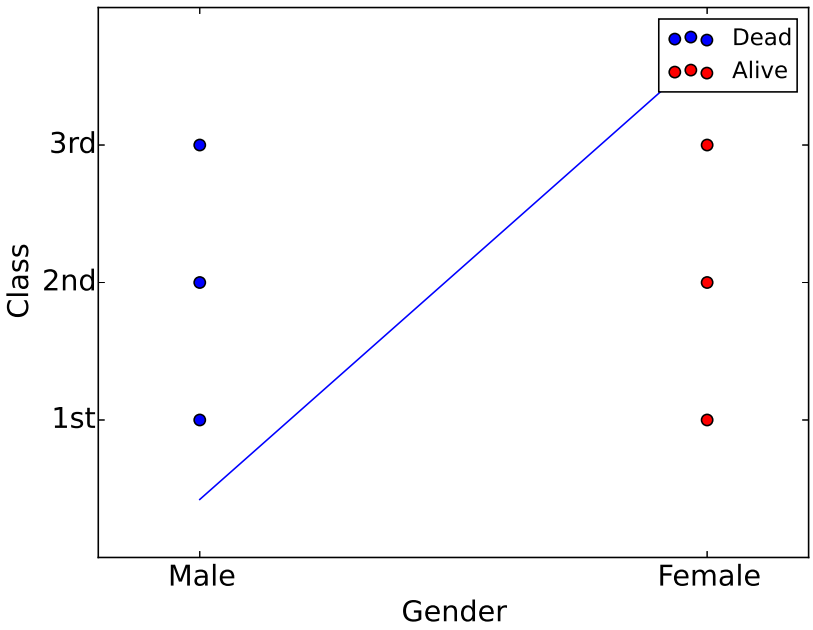

Figure 1: Separating line between survivors and non-survivors according to the SVM model based on the Gender/Class model.

References

[1] T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition* (Springer Series in Statistics).