Department of Statistics

STATS 760: A Survey of Modern Applied Statistics

Assignment 3 2016

As you will no doubt be aware, one of the most terrible maritime disasters of the last century occurred when the White Star liner Titanic struck an iceberg in the North Atlantic on 14th April 1912 and sank with the loss of 1502 out of the 2224 persons aboard.

Primary reasons for the loss of life were the insufficient number of lifeboats, and the disorganized evacuation of the ship that saw many lifeboats launched only partially full.

Survival was to some extent by chance, but was influenced by age, gender and social class.

The data set titanic.csv on the web page contains data on 891 passengers. The variables 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)

You task is to use both random forests and support vector machines to predict the survival of passengers from the other variables. Then figure out the most important variables influencing survival and how they relate to the probability of survival.

PTO

Some variables have missing values. Rather than delete cases with missing values, try to fill them in using some suitable method. You may wish to create some new variables: In particular, the variable “cabin” may have some useful information – it is possible those having cabins on the lower decks had lower survival probabilities, and the variable Name contains information in the title part.

There is a new test set on the webpage. Use this to assess the error rate of your prediction rule. Report the test error calculated from this test set.
You may find that the test error calculated from the test set is rather more than that calculated (using CV etc) from the training set. Can you offer any explanation for this?

Email answers to me by Friday May 27th.