**Assignment – Bayesian Classification**

The **titanic dataset[[1]](#footnote-1)** describes the survival status of individual passengers on the Titanic. The principal source for data about Titanic passengers is the [*Encyclopedia Titanica*](http://www.encyclopedia-titanica.org/). The dataset used here were begun by a variety of researchers. One of the original sources is Eaton & Haas (1994) *Titanic: Triumph and Tragedy*, Patrick Stephens Ltd, which includes a passenger list created by many researchers and edited by Michael A. Findlay.



The variables on the dataset are all nominal: pclass, survived, age, and gender.

* pclass refers to passenger class (1st, 2nd, 3rd, crew), and is a proxy for socio-economic class.
* age is dichotomized at adult vs. child.
* gender is male or female.
* survived is yes or no

The dataset has 2201 instances with no missing data (quite convenient ). It is called titanic.xlsx

PROBLEM 1

I have randomly sampled 32 records out of the original dataset (called titanic\_reduced.xlsx)

a) The task is to manually build a Naive Bayes classifier that, by learning from previously collected data, is able to produce predictions on new demographic data. You must also smooth the model to deal with zero-frequency issues

b) Suppose that you look at a given individual record: (crew, adult, male). What would your Naïve Bayes model predict about the fate of this individual[[2]](#footnote-2)?

Note1: as a recommendation, use Excel, it makes your life easier with calculations.

Note 2: do not partition the data in training and validation sets, use all the data for training (this is just a toy exercise for you to compute the probabilities, and we only have 33 records in this case).

Note 3: You don’t have to be an experienced data miner to predict the fate of a male adult crew member ☺.

PROBLEM 2

In this case you are going to use the full dataset (2201 recs)

Using Modeler, you must:

1. Create a TAN classifier, with zero frequency considerations, trained with 70% random data and tested with the other 30%
2. You must report the model parameters (conditional probabilities and prior probabilities for each class) after training the model.
3. How accurate is the procedure on the training and the test datasets?
4. Once again, suppose that you look at a given individual record: (crew, adult, male). What would your TAN model predict about the fate of this individual[[3]](#footnote-3)?

1. This is an all-nominal-features, no-missing-data dataset typically used in machine learning to assess the performance of classifiers. [↑](#footnote-ref-1)
2. This is a trivial question on a trivial problem: you don’t have to be Sherlock Holmes to figure this out :)) [↑](#footnote-ref-2)
3. This is a trivial question on a trivial problem: you don’t have to be Sherlock Holmes to figure this out :)) [↑](#footnote-ref-3)