Assignment 0
Probabilistic Models of Human and Machine Intelligence
CSCI 7222
Fall 2015

Assigned: Aug 27
Due: Sep 3

Goal

The goal of this assignment is to give you a bit of practice manipulating data, using Bayes' rule, and constructing a naive Bayes classifier. Naive Bayes is described in 10.1 of Barber and understanding examples 10.1 and 10.2 of the text should help you do this assignment.

Data Set

The titanic data set gives the values of four categorical attributes for each of the 2201 people on board the Titanic when it struck an iceberg and sank. The attributes are social class (first class, second class, third class, crew member), age (adult or child), gender, and whether or not the person survived. The titanic data set is available here.

Task 0

Build a joint probability table, like the ones we discussed in class notes, that represents the joint distribution over all varibles, i.e., P(Gender, Age, Class, Outcome). This table should have 32 entries because Gender $\in$ {male, female}, Age $\in$ {child, adult}, Class $\in$ {1st, 2nd, 3d, crew}, and Outcome $\in$ {death, survival}. You will use the data in this table for the following tasks. There is nothing to hand in for Task 0.

Task 1

Build a probability table indicating P(death | Gender, Age, Class) for each combination of class, age, and gender. Display this table in the following way:

	Male		Female	
	Child	Adult	Child	Adult
First				
Second				
Third				
Crew				

The rows of each table represent the different classes and the columns the different ages

and genders. In each cell of the table, insert the conditional probability. Warning: Be alert to the possibility of a cell containing no data.

After you've built the probability table, make a second table, a classification table, which predicts death or survival for each feature combination. If P(death | Gender, Age, Class) > .5, then label that cell as **death**; otherwise label that cell as **survival**.

Task 2

Build a Naive Bayes classifier. To build the classifier, you must first construct six one-dimensional tables: p(Class | death), p(Age | death), p(Gender | death), p(Class | survival), p(Age | survival), p(Gender | survival). To be clear on this notation, for p(Age | death), your table should have two rows, one for adult and one for child, and you should compute, for each age group, the probability of the deceased being in that age group. Also compute the unconditional probabilities, P(death) and P(survival), with P(death)+P(survival)=1. From this information, compute P(death | Gender, Age, Class) using the Naive Bayes assumption. In addition to the probability table, build the classification table as well.

Task 3

The classification tables you built in Tasks 1 and 2 are not identical. Discuss the advantages/disadvantages of each table for making predictions in case of another disaster like the Titanic (assuming it occurred at the same time in history). Under what circumstances would you expect an empirical table to provide better predictions? Under what circumstances would you expect the naive Bayes table to provide better predictions?