CSCI 5822 Homework 0

Evan Sidrow

January 23, 2018

0 Joint Probability Table

The following table shows the raw proportion of any combination of gender, age, class, and survival of individuals in the titanic disaster. It is mainly used as a reference in future exercises.

Table 1: Joint Probability Table of Passenger Identity and Survival in the Titanic Disaster

		Male		Female	
		Child	Adult	Child	Adult
1st	Live	0.00227	0.02590	0.00045	0.06361
	Die	0.0	0.05361	0.0	0.00182
2nd	Live	0.00500	0.00636	0.00591	0.03635
	Die	0.0	0.06997	0.0	0.00591
3rd	Live	0.00591	0.03408	0.00636	0.03453
	Die	0.01590	0.17583	0.00772	0.04044
Crew	Live	0.0	0.08723	0.0	0.00909
	Die	0.0	0.30441	0.0	0.00136

1 Simple Prediction of Death Given Passenger Identity

Using Table 1 above, it is possible to find the probability that a passenger survived given their class, gender, and age. These conditional probabilities are shown below:

Table 2: Conditional Probability of Death Given Passenger Identity

	Male		Female	
	Child	Adult	Child	Adult
1st	0.0	0.67429	0.0	0.02778
2nd	0.0	0.91667	0.0	0.13978
3rd	0.72917	0.83766	0.54839	0.53939
Crew	NA	0.77726	NA	0.13043

In addition, I used this information to predict whether a random individual from this group survived the disaster. There are many ways to do this, but I want my rule to minimize the probability that I will guess wrong. Therefore, I will predict that an individual survived if the probability above is less than or equal to 0.5. Otherwise, I will predict that they died. The resulting classification table is shown below:

Table 3: Prediction of Death Given Passenger Identity

	Male		Female	
	Child	Adult	Child	Adult
1st	Survive	Die	Survive	Survive
2nd	Survive	Die	Survive	Survive
3rd	Die	Die	Die	Die
Crew	NA	Die	NA	Survive

2 Naive Bayes Prediction of Death Given Passenger Identity

In addition to using raw proportions to predict if an individual died in the titanic disaster, it is also possible use a Naive Bayes classifier. In particular:

$$P(die|class, age, gender) = \frac{P(class, age, gender|die) * P(die)}{P(class, age, gender)}$$

$$P(die|class, age, gender) = \frac{P(class, age, gender|die) * P(die)}{P(class, age, gender|die) * P(die) + P(class, age, gender|live) * P(live)}$$

Using the Naive Bayes assumption:

$$P(class, age, gender|death) \approx P(class|death) * P(age|death) * P(gender|death)$$

$$P(class, age, gender|survival) \approx P(class|survival) * P(age|survival) * P(gender|survival)$$

so, combining these equations:

$$P(die|class, age, gender) \approx \frac{P(class|die) * P(age|die) * P(gender|die) * P(die)}{\sum_{D \in \{die, live\}} P(class|D) * P(age|D) * P(gender|D) * P(D)} \tag{1}$$

First, I calculated the conditional probabilities listed above:

Table 4: Probabilities of a Passenger's Class Given that They Lived or Died

	Pr(Class | Die)	Pr(Class | Live)
1st	0.08188	0.285513
2nd	0.11208	0.16596
3rd	0.35436	0.25035
Crew	0.45168	0.29817

Table 5: Probabilities of a Passenger's Age Given that They Lived or Died

	Pr(Age | Die)	Pr(Age | Live)
Child	0.03490	0.08018
Adult	0.96510	0.91831

Table 6: Probabilities of a Passenger's Gender Given that They Lived or Died

	Pr(Gender | Die)	Pr(Gender | Live)
Male	0.91544	0.51617
Female	0.08456	0.48383

Then, I found the raw probabilities of survival and death to complete the Naive Bayes classifier:

Table 7: Total Death and Survival Probabilities

Pr(Death)	Pr(Survival)
0.67697	0.32303

Then I used equation (1) to create the final table shown below:

Table 8: Naive Bayes Probabilities of Death Given Passenger Identity

	Male		Female	
	Child	Adult	Child	Adult
1st	0.31693	0.52792	0.04373	0.09927
2nd	0.52214	0.72478	0.09721	0.20606
3rd	0.69606	0.84662	0.18414	0.35232
Crew	0.71022	0.85522	0.19455	0.36795

As above, I used the rule that $Pr(death) \leq 0.5$ means that I predict an individual from that group will survive. Otherwise, I predict that they will die. Using this rule with Naive Bayes, I get:

Table 9: Prediction of Death Given Passenger Identity using Naive Bayes

	Male		Female	
	Child	Adult	Child	Adult
1st	Survive	Die	Survive	Survive
2nd	Die	Die	Survive	Survive
3rd	Die	Die	Survive	Survive
Crew	Die	Die	Survive	Survive

3 Comparison of Methods

3.1 Advantages and Disadvantages of Each Method

One major advantage to the empirical method is that it does not make any assumptions about the independence of random variables. As a result, it can capture interactions between random variables much better than Naive Bayes. For example, one could argue that the empirical method above was able to capture that male children in 2^{nd} class were prioritized to escape. However, even though all these children survived, Naive Bayes predicted that these children would die, since they were males in 2^{nd} class (both of which are deadly on their own).

However, a major advantage of the Naive Bayes model is that it uses much more information than the empirical method. In particular, The empirical method only used information from the *intersection* of categories an individual was in to make a prediction, while the Naive Bayes used information from the *union* of those categories. If there is hardly any data in the intersection of some categories, then the prediction made by the empirical method may not have enough data behind it to be valid. For example, the empirical method was unable to make a prediction about child crew members since there were none. The Naive Bayes method, however, was able to take a stab at the probability that any hypothetical illegally employed child on the titanic would survive. Whether or not this estimate is valid to be made at all, however, is up to the researcher.

3.2 When to Use Each Method

If it is clear that many of the categories used to identify individuals are not independent (i.e. if a child in 2^{nd} class is *much* more likely to survive than an adult in 2^{nd} class, while a child in general is only *moderately* more likely to survive than an adult), and if there is plenty of data in each category, than the empirical method would be preferred, since it can capture a lack of independence and the abundance of data will make up for its shortfalls.

If the categories appear to be independent and there is not enough data to justify the empirical method, then Naive Bayes would be the preferred method. This is because Naive Bayes has access to more data for each classifier and the Naive Bayes assumption not a terrible one if the data is independent.

If there is not enough data to justify the empirical method, and the categories do not appear to be independent, then it is possible to use a method described below:

3.3 Compromise between the methods

For Naive Bayes, one way to fix the problem of correlation is to combine two highly correlated attributes into one attribute. This fixes the issue of correlation, but it creates many more values that an attribute can take, which makes less data available to Naive Bayes. This new issue is similar to the problem with the empirical approach. In fact, the empirical approach is just a specific case of Naive Bayes where every identifying attribute is combined into a single attribute. Therefore, the optimal strategy is for the researcher to combine attributes carefully such that the attributes are as independent as possible while providing as much data to the Naive Bayes classifier as possible.