A BAYESIAN ANALYSIS OF THE TITANIC DATA SET

ZI SHENG FENG

Abstract. We present a Bayesian analysis of the Titanic data set from the Kaggle competition. The aim is to work from first principles, using only the prior information given and the data set itself, so that each step in the analysis is justified mathematically. This is part of the author's ongoing project to practice the principles laid down in Jaynes [1] with respect to data analysis and inference.

1. Introduction

The motivation for this paper, and a series of papers that will follow, is the conviction that data analysis can be done from first principles, and there are priniciples that guide it, that allow one to do without ad hoc assumptions or unjustified intuitive shortcuts. Professor E.T. Jaynes is arguably the most vocal proponent of this point of view, and he lays down the principles in his treatise Jaynes [1] in an overwhelmingly convincing fashion. This is important because I believe it is the starting point to a scientific study of inference, and ultimately artificial intelligence.

The approach advocated by Jaynes is simply Bayesian. Before we debate the meaning of the word, let us go back to the beginning and ask ourselves for a given data set, how do we analyze it? First, let us be clear about what we know beforehand. For most data sets, we almost always know *something* about them, be they come from some controlled experiment and we are told about the design of the experiment; be it is designed for a health study; or the data set is collected post a catastrophic natural disaster. This is very important, and we must include it; we call it prior information. Second, the data should be at the center of the analysis, and we should use all the information in the data set, rather than formatting, manipulating, or changing it to suit some model we have in mind, which may not be justified. Next, we must be clear what we want to know from the data set, are we predicting the survival of passengers on board a ship; are we calculating the stock price the next day; are we testing the effectiveness of a drugs treatment? Combining these three points, an analysis of the data set boils down to given the data we see, and the prior information we have, what is the best estimate of the quantity of interest? The short, but rigorous, answer is the posterior probability for the quantity, and it is optimal.

Please note that we do not rule out the possibility that there is some model out there that could fit a given data set better than the Bayesian analysis. But if we want to analyze *any* data set in a consistent framework and under a consistent set of principles, without resorting to ad hoc assumptions or unjustified intuitive shortcuts, and that we want to be able to explain every detail of the analysis, rather than using a blackbox, the Bayesian analysis as described in Jaynes [1] is the best we can do and what we should strive for. I will illustrate this approach in this paper and the series of papers that will follow, working on different types of data sets.

2. Titanic Data Set

The Titanic data set is from the Kaggle competition, available at [2]. It is divided into two, one is called training set, and the other is test set. In both, we are given the passengers' attributes on board the Titanic, and they are

- Pclass (social class of the passenger either upper class 1, middle class 2, or lower class 3);
- Name;
- Sex;
- Age;
- SibSp (number of siblings and or spouses of the passenger on board);
- Parch (number of parents and or children of the passenger on board);
- Ticket (ticket number);
- Fare (fare price);
- Cabin (cabin assigned to the passenger);
- Embarked (point of embarkation, either Southampton S, Cherbourg C, or Queenstown Q).

The difference between the two is in the training set we are also given the survival status of the passenger, either survived or not survived, whereas in the test set we need to predict the survival status. So the data analysis problem can be formulated as, based on the information in the training set, and the prior information given, predict the survival of passengers in the test set.

Besides the data, the prior information is as important. In the data set description, we are told three relevant pieces of information

- 1502 out of 2224 passengers and crew were killed (note it is both passengers and crew, and the data set is only for passengers);
- there were not enough lifeboats (it means lifeboat for one made it one less lifeboat for the rest);
- Women, children and the upper-class were more likely to survive than men, adults, or the lower-class.

If we research the background of the Titanic accident, for example at [3], we can refine the first piece of prior information. Since we are concerned only with passengers, excluding crew members, the total number of passengers who survived was 500. Also, in the original data set on the Titanic, adults were defined to be aged 12 and over, children were between 1 and 12 years old, and infants less than 1 year old; whereas for the passenger attribute Pclass, first and second class were lumped together and called Cabin Class, and third class was called Non-Cabin Class.

Let us also say a few words on the other prior information. The point on lifeboats means the survivals of the passengers were *not* independent because that a passenger survived would require a lifeboat; if he got on a lifeboat, then there would be one less space for the remaining passengers, decreasing their chances of survival. The last point on passengers with certain attributes were more likely to survive, we can reformulate it more precisely as follows: given everything else equal, women were more likely to survive than men; children were more likely to survive than adults; and upper-class (cabin class) were more likely to survive than lower-class (non-cabin class). The qualification of "given eveything else equal" means suppose we have passenger A who was a female adult, and passenger B who was a female child, then the prior information tells us passenger B was more likely to survive than passenger A.

3. Problem Setup

We are ready to set up the problem and define the key notations.

Let S_k be the survival status of the kth passenger, $S_k = 1$ denoting he survived, and $S_k = 0$ not survived. Let x_k be the attributes tuple of the kth passenger. For example, if the passenger was a female F, an adult A, and in cabin class YC, then we denote $x_k = (F, A, YC)$ for the attributes of this passenger. Since every passenger had 10 attributes, each x_k should be of length 10, but for the following calculations, we will use only attributes Sex, Age, and Pclass, and we will extend it to the full case in Section 5. The notations for the attributes are

- Sex = Female F or Male M;
- Age[1] = Adult A or Child C;
- Pclass[2] = Cabin-class YC or Non-Cabin-class NC.

Let D_0 be the data in the training set. As just mentioned, we will first work on a toy version of the problem, and assume only attributes Sex, Age and Pclass in the data set. Let I be the prior information. More precisely, it consists of

- a total of 500 passengers survived;
- there were not enough lifeboats, so the passengers were not independent;
- given everything else equal, women were more likely to survive than men; children were more likely to survive than adults; and cabin-class were more likely to survive than non-cabin-class.

The data analysis problem we want to solve is compute the posterior probability for the survival status of passengers in the test set. That is, given the data in the training set and the prior information, if the kth passenger in the test set was a female adult of cabin-class, for example, what is our prediction of her survival probability,

$$p(S_k = 1 \mid x_k = (F, A, YC), D_0, I)$$

[1] As in the original data set, we define Adult to be 12 years old and over, and Child otherwise.

[2] We also lump first and second class under Cabin-class, and third class as Non-Cabin class.

4. Computations

4.1. Step 1 - One attribute.

Let us start with a simpler calculation to see the main argument.

Consider $p(S_k = 1 \mid x_k = F, D_0, I)$. So we want to compute the posterior survival probability for a female passenger, with no information about the other attributes[3]. Note $p(S_k = 1 \mid x_k = F, D_0, I) = p(S_k = 1, x_k = F \mid D_0, I)/p(x_k = F \mid D_0, I)$, and we will do the calculations for $p(S_k = 1, x_k = F \mid D_0, I)$ instead, due to its leading to simpler expansion.

Let $N^{s|f}$ be the total number of female survivors in the whole data set (training and test sets combined), and $N_0^{s|f}$ be those in the training set; N^f be the total number of females in the whole data set, and N_0^f be those in the training set. Expanding by $N^{s|f}$ and using Bayes' Theorem on $p(N^{s|f} \mid D_0, I)$,

$$\begin{aligned} p(S_k = 1, x_k = F \mid D_0, I) &= \sum_{N^{s|f}} p(S_k = 1, x_k = F \mid N^{s|f}, D_0, I) * p(N^{s|f} \mid D_0, I) \\ &= \sum_{N^{s|f}} p(S_k = 1 \mid x_k = F, N^{s|f}, D_0, I) * \frac{p(N^{s|f} \mid I) * p(D_0 \mid N^{s|f}, I)}{p(D_0 \mid I)} \\ &\times \ p(x_k = F \mid D_0, I), \end{aligned}$$

where we note $p(x_k = F \mid N^{s|f}, D_0, I)$ is independent of $N^{s|f}$ since it is the proportion of females in the test set. We will go over each term in turn.

For $p(S_k = 1 \mid x_k = F, N^{s|f}, D_0, I)$, it asks given $N^{s|f}$ out of N^f females in the data set who survived, and $N_0^{s|f}$ out of N_0^f females in the training set who survived, what is the survival probability we shoud assign to a female passenger in the test set. Clearly, it should be the proportion of female survivors in the test set, so

$$p(S_k = 1 \mid x_k = F, N^{s|f}, D_0, I) = \frac{N^{s|f} - N_0^{s|f}}{N^f - N_0^f}.$$

For $p(D_0 \mid N^{s|f}, I)$, it asks given $N^{s|f}$ out of N^f females survived, what is the probability of having $N_0^{s|f}$ out of N_0^f females in the training set who survived[4]. This is a hypergeometric distribution, and the probability is

$$p(D_0 \mid N^{s|f}, I) = \begin{pmatrix} N^f \\ N_0^f \end{pmatrix}^{-1} * \begin{pmatrix} N^{s|f} \\ N_0^{s|f} \end{pmatrix} * \begin{pmatrix} N^f - N^{s|f} \\ N_0^f - N_0^{s|f} \end{pmatrix}.$$

Finally, we need to compute the prior $p(N^{s|f} \mid I)$. Let $N^{s|m}, N^m$ be the number of male survivors and the number of males in the data set respectively. Note we can decompose the

[3] It implies that in this case, we should interpret the data D_0 as the only information given for each passenger in the training set is their sex and survival status.

[4] Recall the earlier remark that in this toy problem, since the only attribute given is Sex, we should interpret the training data D_0 as such.

proportion of survivors into the proportions of female and male survivors, weighted by the proportions of females and males, i.e. $\left(\frac{N^{s|f}}{N^f}\right)\left(\frac{N^f}{N}\right)+\left(\frac{N^{s|m}}{N^m}\right)\left(\frac{N^m}{N}\right)=\frac{N^s}{N}$. From the prior, we know females were more likely to survive than males, i.e.

$$\frac{N^{s|f}}{N^f} > \frac{N^{s|m}}{N^m}.$$

But this condition is equivalent to $\frac{N^{s|m}}{N^m} < \frac{N^s}{N}$, or $N^{s|m} < \frac{N^s}{N}N^m$, as can be checked from the proportion of survivors equation we described earlier. It says the number of male survivors is at most $\frac{N^s}{N}N^m$. Since any value less than $\frac{N^s}{N}N^m$ is possible, and this is the only information we have about $N^{s|m}$, by the principle of maximum entropy[5], the prior distribution we should assign[6] for $N^{s|m}$ is

$$p(N^{s|m} \mid I) = \frac{1}{\frac{N^s}{N}N^m} * \mathbf{1}(N^{s|m} \leq \frac{N^s}{N}N^m).$$

And that for $N^{s|f}$ is obvious as $N^{s|f} = N^s - N^{s|m}$.

4.2. **Step 2 - Several Attributes.**

Now we compute $p(S_k = 1 \mid x_k = (F, A, YC), D_0, I)$. As in Step 1, we work with $p(S_k = 1, x_k = (F, A, YC) \mid D_0, I)$ instead.

Let $N^{s|f,a,yc}$ be the number of female adults of cabin-class in the data set who survived, and $N_0^{s|f,a,yc}$ be those in the training set; $N^{f,a,yc}$ be the number of female adults of cabin-class in the data set, and $N_0^{f,a,nc}$ be those in the training set. And any other notations of the form $N^{s|\cdot}, N_0^{s|\cdot}$, and $N^{\cdot}, N_0^{\cdot}$ will have similar meanings.

Next, we expand by $N^{s|f,a,yc}$ and use Bayes' Theorem, then get

$$\begin{aligned} p(S_k = 1, x_k = (F, A, YC) \mid D_0, I) &= \sum_{N^{s|f,a,yc}} p(S_k = 1 \mid x_k = (F, A, YC), N^{s|f,a,yc}, D_0, I) \\ &\times \frac{p(N^{s|f,a,yc} \mid I) * p(D_0 \mid N^{s|f,a,yc}, I)}{p(D_0 \mid I)} \\ &\times p(x_k = (F, A, YC) \mid D_0, I), \end{aligned}$$

[5] The principle of maximum entropy is one of the ways to assign prior probability. It shows how to compute a probability distribution for a quantity of interest subject to the limited amount of prior information we have about the quantity itself, for example the quantity of interest may be the measurement error in an experiment and the only thing we know is its first and second moment. The principle says we choose the distribution that maximizes entropy subject to what we know, say the first and second moment; the intuition of maximizing entropy is we want the distribution to be as uniform as possible so that it includes all the possibilities that satisfy the constraints - think of the uniform distribution but skew it in such a way that it satisfies some pre-known conditions. For details, see Chapter 11 in Jaynes [1]. In our case, the principle of maximum entropy is also called the principle of indifference, and we just assign a uniform distribution over an appropriate interval, or see Section 6 - Appendix for the computations.

[6] The upper bound $\frac{N^s}{N}N^m$ might not be an integer; more precisely $N^{s|m} \leq \lfloor \frac{N^s}{N}N^m \rfloor$, and the normalization constant is $\lfloor \frac{N^s}{N}N^m \rfloor + 1 = \lceil \frac{N^s}{N}N^m \rceil$ since $N^{s|m}$ could be zero.

For $p(S_k = 1 \mid x_k = (F, A, YC), N^{s|f,a,yc}, D_0, I)$, it is the survival probability of a passenger in the test set given $N^{s|f,a,yc}$ female adults of cabin-class in the data set who survived, so it is equal to $\frac{N^{s|f,a,yc} - N_0^{s|f,a,yc}}{N^{f,a,yc} - N_0^{f,a,yc}}$.

For $p(D_0 \mid N^{s|f,a,yc}, I)$, it is the probability of having $N_0^{s|f,a,yc}$ in the training set given $N^{s|f,a,yc}$ in the data set, and it is equal to $\begin{pmatrix} N^{f,a,yc} \\ N_0^{f,a,yc} \end{pmatrix}^{-1} \begin{pmatrix} N^{s|f,a,yc} \\ N_0^{s|f,a,yc} \end{pmatrix} \begin{pmatrix} N^{f,a,yc} - N^{s|f,a,yc} \\ N_0^{f,a,yc} - N_0^{s|f,a,yc} \end{pmatrix}$.

Finally, for the prior probability $p(N^{s|f,a,yc} \mid I)$, we decompose the proportion of female adult survivors into those who were cabin-class and those not as $\left(\frac{N^{s|f,a,yc}}{N^{f,a,yc}}\right)\left(\frac{N^{f,a,yc}}{N^{f,a}}\right) + \left(\frac{N^{s|f,a,nc}}{N^{f,a,nc}}\right)\left(\frac{N^{f,a,nc}}{N^{f,a}}\right) = \frac{N^{s|f,a}}{N^{f,a}}$. Using the prior information that given everything else equal, cabin-class were more likely to survive than non-cabin-class in the sense that

$$\frac{N^{s|f,a,yc}}{N^{f,a,yc}} > \frac{N^{s|f,a,nc}}{N^{f,a,nc}},$$

which is equivalent to $\frac{N^{s|f,a,nc}}{N^{f,a,nc}} < \frac{N^{s|f,a}}{N^{f,a}}$, the prior distribution we assign for $N^{s|f,a,nc}$ is

$$p(N^{s|f,a,nc} \mid N^{s|f,a}, I) = \frac{1}{\frac{N^{s|f,a}}{N^{f,a}} N^{f,a,nc}} * \mathbf{1}(N^{s|f,a,nc} \leq \frac{N^{s|f,a}}{N^{f,a}} N^{f,a,nc}),$$

and $p(N^{s|f,a,nc} \mid I) = \sum_{N^{s|f,a}} p(N^{s|f,a,nc} \mid N^{s|f,a}, I) p(N^{s|f,a} \mid I)$, and that for $N^{s|f,a,yc}$ follows from $N^{s|f,a,yc} = N^{s|f,a} - N^{s|f,a,nc}$.

In summary,

$$\begin{aligned} p(S_k = 1 \mid x_k = (F, A, YC), D_0, I) &= \sum_{N^{s|f,a,yc}} p(S_k = 1 \mid x_k = (F, A, YC), N^{s|f,a,yc}, D_0, I) \\ &\times \frac{p(N^{s|f,a,yc} \mid I) * p(D_0 \mid N^{s|f,a,yc}, I)}{p(D_0 \mid I)} \\ &= \frac{1}{p(D_0 \mid I)} * \sum_{N^{s|f,a,yc}} \frac{N^{s|f,a,yc} - N_0^{s|f,a,yc}}{N^{f,a,yc} - N_0^{f,a,yc}} \\ &\times \begin{pmatrix} N^{f,a,yc} \\ N_0^{f,a,yc} \end{pmatrix}^{-1} * \begin{pmatrix} N^{s|f,a,yc} \\ N_0^{s|f,a,yc} \end{pmatrix} * \begin{pmatrix} N^{f,a,yc} - N^{s|f,a,yc} \\ N_0^{f,a,yc} - N_0^{s|f,a,yc} \end{pmatrix} \\ &\times \sum_{N^{s|f,a}} p(N^{s|f,a} - N^{s|f,a,yc} \mid N^{s|f,a}, I) * \sum_{N^{s|f}} p(N^{s|f,a} \mid N^{s|f}, I) * p(N^s - N^{s|f} \mid I), \end{aligned}$$

where $N^s = 500$ is the total number of survivors, and the three probabilities in the last line are respectively the prior distribution for $N^{s|f,a,nc}$, $N^{s|f,a}$ and $N^{s|m}$. More precisely,

$$\begin{cases} p(N^{s|f,a} - N^{s|f,a,yc} \mid N^{s|f,a}, I) &= \frac{1}{\frac{N^{s|f,a}}{N^{f,a}} N^{f,a,nc}} * \mathbf{1}(\frac{N^{s|f,a}}{N^{f,a}} N^{f,a,yc} \leq N^{s|f,a,yc}); \\ p(N^{s|f,a} \mid N^{s|f}, I) &= \frac{1}{\frac{N^{s|f}}{N^f} N^{f,a}} * \mathbf{1}(N^{s|f,a} \leq \frac{N^{s|f}}{N^f} N^{f,a}); \\ p(N^s - N^{s|f} \mid I) &= \frac{1}{\frac{N^s}{N} N^m} * \mathbf{1}(\frac{N^s}{N} N^f \leq N^{s|f}). \end{cases}$$

4.3. Step 3 - Bounds on the Ranges.

In the posterior probability, we have three sums, and we need to specify the ranges. Since the sums are finite, we can reorder the summation. We will first sum over $N^{s|f,a,yc}$, then $N^{s|f,a}$, and finally $N^{s|f}$.

Given the data in the training set D_0, the number of female adult survivors of cabin-class in the whole data set must be at least that in D_0, and it is at most the number of female adults of cabin-class, i.e. $N_0^{s|f,a,yc} \leq N^{s|f,a,yc} \leq N^{f,a,yc}$. On the other hand, from the prior distribution, $\frac{N^{s|f,a}}{N^{f,a}} N^{f,a,yc} \leq N^{s|f,a,yc}$. Thus

$$\max(N_0^{s|f,a,yc}, \frac{N^{s|f,a}}{N^{f,a}} N^{f,a,yc}) \leq N^{s|f,a,yc} \leq N^{f,a,yc}.$$

Note these bounds always make sense, because $N_0^{s|f,a,yc} \leq N^{f,a,yc}$, and $\frac{N^{s|f,a}}{N^{f,a}} N^{f,a,yc} \leq N^{f,a,yc}$.

Next, for $N^{s|f,a}$, given D_0, $N_0^{s|f,a} \leq N^{s|f,a}$ and $N^{s|f,a} \leq N^{f,a}$. From the prior distribution, $N^{s|f,a} \leq \frac{N^{s|f}}{N^f} N^{f,a}$, since $\frac{N^{s|f}}{N^f} N^{f,a} \leq N^{f,a}$, so

$$N_0^{s|f,a} \leq N^{s|f,a} \leq \frac{N^{s|f}}{N^f} N^{f,a}.$$

Before we proceed to $N^{s|f}$, the bounds on $N^{s|f,a}$ give rise to the condition that they should make sense, i.e. $N_0^{s|f,a} \leq \frac{N^{s|f}}{N^f} N^{f,a}$, which becomes an condition on the lower bound on $N^{s|f}$ as $N_0^{s|f,a}/(N^{f,a}/N^f) \leq N^{s|f}$

Finally, for $N^{s|f}$, given D_0, we have $N_0^{s|f} \leq N^{s|f} \leq N^f$; from the prior, $\frac{N^s}{N} N^f \leq N^{s|f}$, and from the condition on $N^{s|f,a}$ just saw, the bounds on $N^{s|f}$ are

$$\max(N_0^{s|f}, \frac{N^s}{N} N^f, N_0^{s|f,a}/(N^{f,a}/N^f)) \leq N^{s|f} \leq N^f.$$

We can simplify the lower bound by substituting in the given values in the data set; we have $N_0^{s|f} = 233, N^s = 500, N = 1309, N_0^{s|f,a} = 213, N^{f,a} = 419, N^f = 466$, so $\max(N_0^{s|f}, \frac{N^s}{N} N^f, N_0^{s|f,a}/(N^{f,a}/N^f)) = \max(233, 178, 237) = 237$.

In summary, the procedure to compute the ranges is use the bounds given in D_0, and combine them with the bounds from the prior distributioin. Also, there may be an extra condition that the bounds should make sense. Sometimes, we might be able to simplify the bounds by substituting in the given values in the data set.

4.4. **Step 4 - Summary of Results.**

It may not be possible to compute the posterior probability by hands, though the formula is explicit. We have coded the formulas in Python to do the computations, and the files are in Section 6.2 - Attachments.

After running the codes, we get

- $p(S_k = 1 \mid x_k = (F, A, YC), D_0, I) = 0.9448579$;
- $p(S_k = 1 \mid x_k = (F, A, NC), D_0, I) = 0.5082974$;
- $p(S_k = 1 \mid x_k = (F, C, YC), D_0, I) = 0.8555349$;
- $p(S_k = 1 \mid x_k = (F, C, NC), D_0, I) = 0.4322401$;
- $p(S_k = 1 \mid x_k = (M, A, YC), D_0, I) = 0.2344567$;
- $p(S_k = 1 \mid x_k = (M, A, NC), D_0, I) = 0.1119066$;
- $p(S_k = 1 \mid x_k = (M, C, YC), D_0, I) = 0.6220168$;
- $p(S_k = 1 \mid x_k = (M, C, NC), D_0, I) = 0.3848169$.

5. Comments

5.1. **Extensions to the Full Case.**

To complete the analysis, we should include all attributes given in computing the posterior probabilities, for example, $p(S_k = 1 \mid x_k = (F, A, YC, \ldots), D_0, I)$, where the attributes tuple x_k also indicates how many spouses/siblings/parents/children the passenger had accompanied on board, how much they paid for the voyage, where they got on board, etc. And this is *all* the information in the data set. But for these attributes, we do not have any relevant prior information, for example, we do not know if a passenger embarked at Cherbourg was more likely to survive than one who embarked at Southampton, or if a passenger with many children on board had higher survival probability than one with none. In order to include these other attributes in the analysis, we need to assign a prior distribution on each. In situations where we do not have any, the principle of maximum entropy (or the principle of indifference in this case) also applies, namely, we assign a uniform distribution subject to bounds on its possible values. Let us outline briefly how to deal with each of the remaining attributes.

5.1.1. *Embark.* Attribute Embark has 3 discrete values, S, C, Q. Let $N^{s|S}, N^{s|C}, N^{s|Q}$ be the number of survivors who embarked at the three cities respectively. The prior information given in the data set description does not tell us how they are related, or if they are correlated at all, though we know their sum is the total number of survivors N^s. Based on this, and by the principle of maximum entropy, we assign the uniform distribution; since the three quantities are related by $N^{s|S} + N^{s|C} + N^{s|Q} = N^s$, they are *not* independent. To derive the uniform distribution, we consider it in two steps. First, we find the distribution for $(N^{s|S}, N^{s|C})$; second, from the equation $N^{s|S} + N^{s|C} + N^{s|Q} = N^s$, we can readily compute the distribution for $N^{s|Q}$. So it is enough to complete step 1. Note $0 \leq N^{s|S} + N^{s|C} \leq N^s$, and this is the only constraint, so the maximum entropy distribution is uniform on $0 \leq k \leq N^s, 0 \leq l \leq N^s - k$. See Section 6 - Appendix for the computations.

5.1.2. *Fare.* Attribute Fare is a continuous variable. The reason it might be relevant to predicting the survival status is we expect passengers who paid a high fare price might be more likely to survive, but note that high-paying passengers were not necessarily in the upper class, as it could be due to where they embarked. In any case, if what we are interested in is to distinguish between high-paying and low-paying passengers, we could choose some threshold of the fare price so that if above it, we call those HighFare, and LowFare otherwise - for example, we can pick the 75th quantile, which is \$30. And the prior distribution we assign to HighFare and LowFare survivors is again uniform, because we do not have relevant prior information on this attribute.

5.1.3. *SibSp, Parch, Ticket.* If we think for a second how these three attributes are related, we will realize they tell us whether or not the passenger came on board in a family, and the size of the group the passenger belonged, be it a family or a group of friends. To see this, we first sort the data set by Ticket. Then passengers with the same ticket number are grouped together. For passengers with the same ticket number, if we count the number of spouses/siblings/children/parents each of these passenger had accompanied on board, i.e. SibSp + Parch+1, in most cases, it should equal the number of passengers with the same ticket number. If it is not equal, it is either due to some of the passengers in the group were Cross-Channel passengers, who embarked at Southampton, but got off at Cherbourg, for example, and they did not take the full voyage to New York City; or some passengers in the group were not related to the family of the group, for example a friend of the family, and these passengers had SibSp=Parch=0. However, the main point is by grouping passengers with the same ticket number, we see how many families and non-family groups there were, and the size of each group. If we think this is the information most relevant to predicting the survival status of the passengers, we can define two alternative attributes, one is called GroupType, 1 indicating the passenger had family members on board, 0 otherwise; and another is called GroupSize, L indicating the passenger belonged to a group of at least 4 members, and S indicating the passenger belonged to a group of at most 3 members.

5.2. **Is it not just counting?**

In Section 4.4, we summarized the posterior probabilities for passengers with attributes Sex, Age, and Pclass. If we predict the survival status of passengers in the test set based on whether the posterior probability is bigger than 0.5, we will predict all female adult passengers survived, and all cabin-class children survived. Interestingly, if we also look at the corresponding proportions of survivors with these attributes in the training set, we come to the same conclusion, specifically,

- $N_0^{s|f,a,yc}/N_0^{f,a,yc} = 0.9503106$;
- $N_0^{s|f,a,nc}/N_0^{f,a,nc} = 0.5172414$;
- $N_0^{s|f,c,yc}/N_0^{f,c,yc} = 0.8888889$;
- $N_0^{s|f,c,nc}/N_0^{f,c,nc} = 0.4285714$;
- $N_0^{s|m,a,yc}/N_0^{m,a,yc} = 0.2293578$;
- $N_0^{s|m,a,nc}/N_0^{m,a,nc} = 0.1128527$;
- $N_0^{s|m,c,yc}/N_0^{m,c,yc} = 1.0000000$[7];
- $N_0^{s|m,c,nc}/N_0^{m,c,nc} = 0.3928571$.

So is not our Bayesian analysis just counting? This is not surprising because as argued in the beginning of this paper, any reasonable method of data analysis should put the data at the center. If the data happens to support the prior information, then the analysis should

[7] This differs markedly from the corresponding posterior probability, which is 0.6220168; the reason is it reflects the fact that the data assigns a large weight to females passengers, and relatively small weight to male passengers.

still support the data; but if the data happens to differ from the prior information, then the information in the data should override the prior information. And Bayesian analysis is a formal and rigorous way to carry out the argument. In the next few papers of the series, we will gain a better understanding as to how this analysis is carried out on different types of data sets.

6. Appendix

6.1. Maximum Entropy Computations.

Claim 6.1. *For a quantity A such that $0 \leq A \leq N$, the maximum entropy distribution we assign to A is uniform on $(0, N)$.*

Proof. Let $p(k)$ be a probability distribution for A, where $k = 0, 1, \dots, N$, and $H = -\sum_{k=0}^{N} p(k) \log p(k)$ be the entropy of p. The only constraint on p is that it be a probability distribution, i.e. $\sum_{k=0}^{N} p(k) = 1$. Using Lagrange multiplier, we want to find p that maximizes

$$H - \mu \sum_{k=0}^{N} p(k).$$

Differentiating each $p(k)$, we get $-(\log p(k) + 1) - \mu = 0$, so $p(k) = e^{-(\mu+1)}$. Summing both sides over k and using the constraint that the sum is 1, we deduce $p(k) = e^{-(\mu+1)} = \frac{1}{N+1}$. □

Claim 6.2. *For a pair of quantities (A, B) such that $0 \leq A+B \leq N$, the maximum entropy distribution we assign to (A, B) is uniform on $0 \leq A + B \leq N$.*

Proof. The proof is almost the same as before. Let $p(k, l)$ be a probability distribution for (A, B), and $H = -\sum_{k=0}^{N} \sum_{l=0}^{N-k} p(k, l) \log p(k, l)$ be its entropy. Again, the only constraint is its sum over the region $0 \leq k \leq N, 0 \leq l \leq N - k$ be equal to 1, and using Lagrange multipliers, we want to maximize $H - \mu \sum_{k=0}^{N} \sum_{l=0}^{N-k} p(k, l)$. The rest is the same, and the normalization constant is $\frac{2}{(N+1)(N+2)}$, whch is 1 over the area of $0 \leq k \leq N, 0 \leq l \leq N - k$. □

6.2. Attachments.

We attach here a file that lists the formulas for the other 7 posterior probabilities, and a python file for implementing the formulas.

- titanic_posterior_probabilities.pdf;
- titanic_python.py.

References

[1] E.T. Jaynes: *Probability Theory the Logic of Science*, edited by G. Larry Bretthorst. Cambridge University Press (2003)

[2] Kaggle: *Titanic: Machine Learning from Disaster.* https://www.kaggle.com/c/titanic, accessed October 15 2017

[3] Titanic: *Encyclopedia Titanica.* https://www.encyclopedia-titanica.org/, accessed October 15 2017

FENGZISHENG@HOTMAIL.COM