Lab 7: Probability

3/14/2018

Introduction

Given an assortment of poorly transcribed letters of the instructions of John Mills to his son regarding radio operation, we attempt to discover who has transcribed each letter. Mr. Mills had two typists under his employ, both who did a miserable job leading to a myriad of typographical errors. We know that letters 1, 8, and 16 are attributed to typist 1 while letters 4, 9, and 18 are attributed to typist 2. We seek to find who transcribed letters 3, 7, 10, 11, 15, and 22 by looking at the pattern of errors established by each of the typists in the letters we know they transcribed.

We start with a set of assumptions. The first of these assumptions is that the there are only two typists that we can choose from. A second is that these typists are mutually exclusive, meaning that each letter must be ascribed to either but not both of the typists. Another assumption would be that the code we are using works correctly. The range of our characters is also an assumption we take, that we consider only the 26 letters of the alphabet and do not regard capitalization, as well as space and newline. We must also assume that typographical errors are only conditioned on the letter that was intended to be typed, no outside conditions that could cause errors are considered such as the probability of a cup falling on the keyboard and mistyping letters. A final assumption is that the original letters will have no errors.

To find the probability that a certain letter belongs to a typist we will train a model for each typist on a letter we know they wrote, then calculate the likelihood for each model given a

different letter. The training will grow more extensive as we go on to hopefully gain a more accurate prediction.

Analysis

We will need to calculate a likelihood and a prior to calculate the Jayne's evidence which we will be using.

$$e(H|DX) = e(H|X) + 10 \log_{10} \left[\frac{P(D|HX)}{P(D|\overline{H}X)} \right]$$

Which for this problem would be:

e(T = t | O = o, C = c, D = d, *I*)

= e(T = t | C = c, D = d, *I*)

+ 10 log10 [P(O = o | C = c, D = d, T = t, *I*) /

P(O = o | C = c, D =d, T = t', *I*)]

The prior e(T = t | C = c, D = d, *I*) will be zero. In accordance to the Jayne's equation the prior would have the form 10log(p/q) which is equivalent to 10(logp - logq). p and q will be calculated with the rule of succession:

$$P(X_{n+1} = 1 \mid X_1 + \cdots + X_n = s) = \frac{s+1}{n+2}.$$

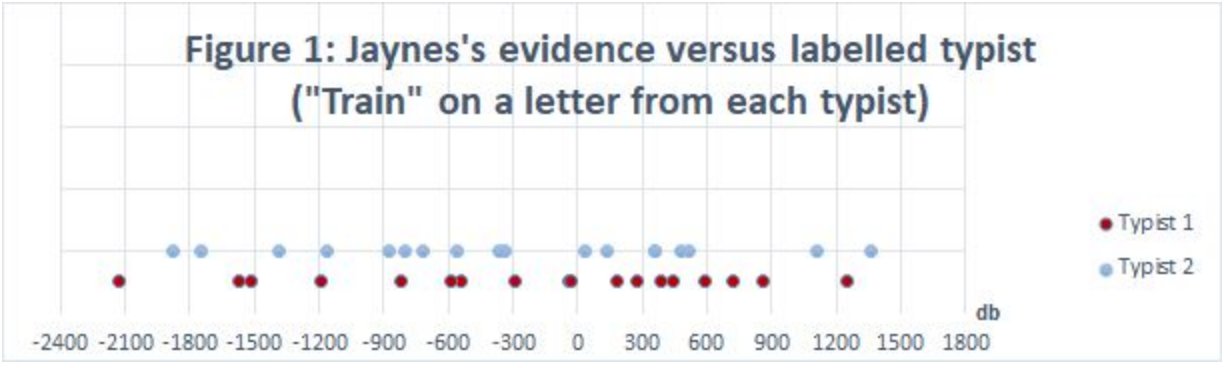
As we know there 3 letters of the confirmed 6 are by typist 1, Laplace's rule of succession states that the probability that the next letter will be by typist 1 is :

$$3 + 1 / 6 + 2 = 4 / 8 = ½$$

The same applies to q where 3 of the confirmed 6 letters are by typist 2. This leaves us with 10(log (½) - log(½)), which reduces to 0.

The training occurs in three iterations. For the first round each typist model trains on a single letter then is measured on one of the other letters that is already assigned. For example, " train typist 1's model on letter 16 and typist 2's model on letter 18, then calculate the evidence on letter 4". This leads to 36 possible combinations.

Figure 1: Jaynes's evidence versus labelled typist ("Train" on one letter from each typist)

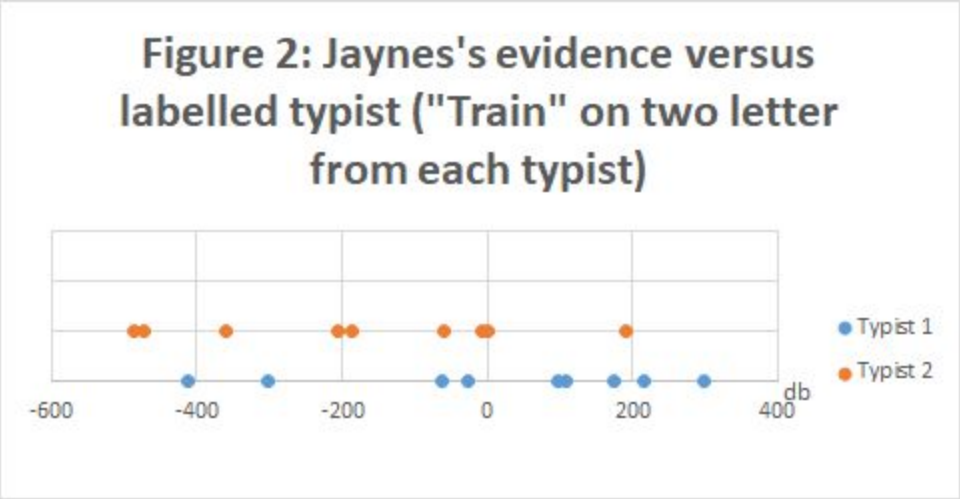


In the Figure 1, there are no clear distinction between the letters which from typist 1 and the letters from typist 2. There are 18 incorrect resulting predictions, almost ½ of all predictions.

Because of the vague distribution and the incorrectness of resulting predictions, it is not sufficient to predict the typist with data from one letter of each typist.

The second round trains each model with two of the letters from each typist, then measures on one of the unassigned letters. For example , train typist 1's model on letters 1 and 16 and train typists 2's model on letters 4 and 18, then calculate the evidence on letter 7. This leads to 18 possible combinations.

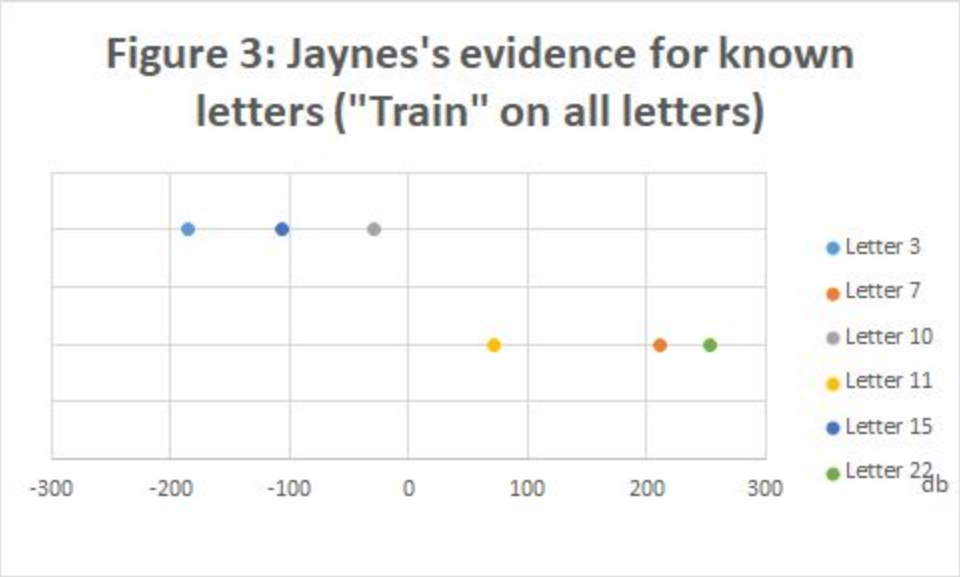
Figure 2: Jaynes's evidence versus labelled typist ("Train" on two letters from each typist)



In the Figure 2, there are slight distinction between the letters from typist 1 and the letters from typist 2. There are 6 incorrect resulting predictions, almost 1/3 of all predictions: 4 of them happen when we predict a letter which the actual typist is typist 1. We notice that the evidence of letters from typist 1 is at the right side of the graph, and that from typist 2 is at the left side of the graph. Since we find the general pattern of the distribution for each typist, we can predict the typist with high uncertainty.

The final round trains each model with all three of the letters known to be from each typist and calculates the evidence for each unknown letter being attributed to either typist. For example, train typist 1's model on letters 1, 8, and 16 then train typist 2's model on letters 4, 9, and 18, then calculate the evidence on letter 22. We hypothesize that this will lead to more accurate predictions than the second round because the patterns of errors by each typist should be more completely defined with more evidence.

Figure 3: Jaynes's evidence versus labelled typist ("Train" on all letters)



In the Figure 3, there are obvious distinction between the letters from typist 1 and the letters from typist 2 without overlapping. Since the accuracy increases as the increasing of training letters, the prediction is much more accurate than what we have in the Figure 2. Thus, we believe it is suffice to say letter 3, 15, 10 are typed by typist 2, and letter 7, 11, 22 are typed by typist 1.

Conclusion

What we are looking for is a change in the distribution of evidence. As we become more certain that a certain letter belongs to a certain typist, there should be a more notable grouping of letters thought to be typed by typist one or two and less overlap between the distributions. Essentially, as the evidence grows stronger, the evidence also becomes more polarized. In Figure 2 we can see that Typist 1 tends to group slightly to the right while Typist 2 groups slightly to the left. The difference is not particularly pronounced but it is noticeable. Figure 3 shows a distinct divide between the letters expected to belong to typist 1 and typist 2 with no overlap between the groups in evidence. To sum up, it is reasonable to state that provided with more effective data, the Jaynes's evidence could create more accurate prediction.

Appendix

Table 1: Configurations Jaynes's evidence ("Train" with one letter from each typist)

Configuration	Jaynes's evidence	P(T = typist 1)	Resulting Prediction	Actual Typist
1	-542.4094	<1/2	2	1
2	-822.9633	<1/2	2	1
3	-1195.0854	<1/2	2	1
4	-1569.1003	<1/2	2	1
5	-1516.928	<1/2	2	1
6	-2125.8082	<1/2	2	1
7	384.9113	>1/2	1	1
8	717.0082	>1/2	1	1
9	-35.8485	<1/2	2	1
10	-29.1288	<1/2	2	1
11	-292.8413	<1/2	2	1
12	-585.8367	<1/2	2	1
13	860.0157	>1/2	1	1
14	1245.9359	>1/2	1	1
15	439.256	>1/2	1	1
16	593.2599	>1/2	1	1
17	182.2631	>1/2	1	1
18	271.4173	>1/2	1	1

19	-803.7612	<1/2	2	2
20	-872.6695	<1/2	2	2
21	-1164.1632	<1/2	2	2
22	-1752.3077	<1/2	2	2
23	-1388.7877	<1/2	2	2
24	-1878.5114	<1/2	2	2
25	361.186	>1/2	1	2
26	515.8728	>1/2	1	2
27	-336.961	<1/2	2	2
28	-363.7653	<1/2	2	2
29	-561.5856	<1/2	2	2
30	-713.5643	<1/2	2	2
31	1109.9431	>1/2	1	2
32	1360.0884	>1/2	1	2
33	358.34	>1/2	1	2
34	480.4503	>1/2	1	2
35	133.7155	>1/2	1	2
36	35.1928	>1/2	1	2

Table 2: Configurations Jaynes's evidence ("Train" with two letters from each typist)

Configuration	Jaynes's evidence	P(T = typist 1)	Resulting Prediction	Actual Typist

1	-63.3784	<1/2	2	1
2	-301.7836	<1/2	2	1
3	-412.9772	<1/2	2	1
4	215.4976	>1/2	1	1
5	96.1361	>1/2	1	1
6	-25.5225	<1/2	2	1
7	296.7688	>1/2	1	1
8	175.4074	>1/2	1	1
9	108.0111	>1/2	1	1
10	-358.5059	<1/2	2	2
11	-472.5578	<1/2	2	2
12	-485.2422	<1/2	2	2
13	1.2077	>1/2	1	2
14	-185.9933	<1/2	2	2
15	-204.7125	<1/2	2	2
16	190.9332	>1/2	1	2
17	-7.1153	<1/2	2	2
18	-60.5826	<1/2	2	2

Table 3: Configurations Jaynes's evidence ("Train" with all letters)

Letter	Jaynes's evidence	P(T = typist 1)	Resulting Prediction
Letter 3	-185.147	<1/2	2

Letter 7	211.0077	>1/2	1
Letter 10	-28.6914	<1/2	2
Letter 11	72.3054	>1/2	1
Letter 15	-105.6103	<1/2	2
Letter 22	253.5092	>1/2	1