A Realistic Approach to Letter Classification

Introduction

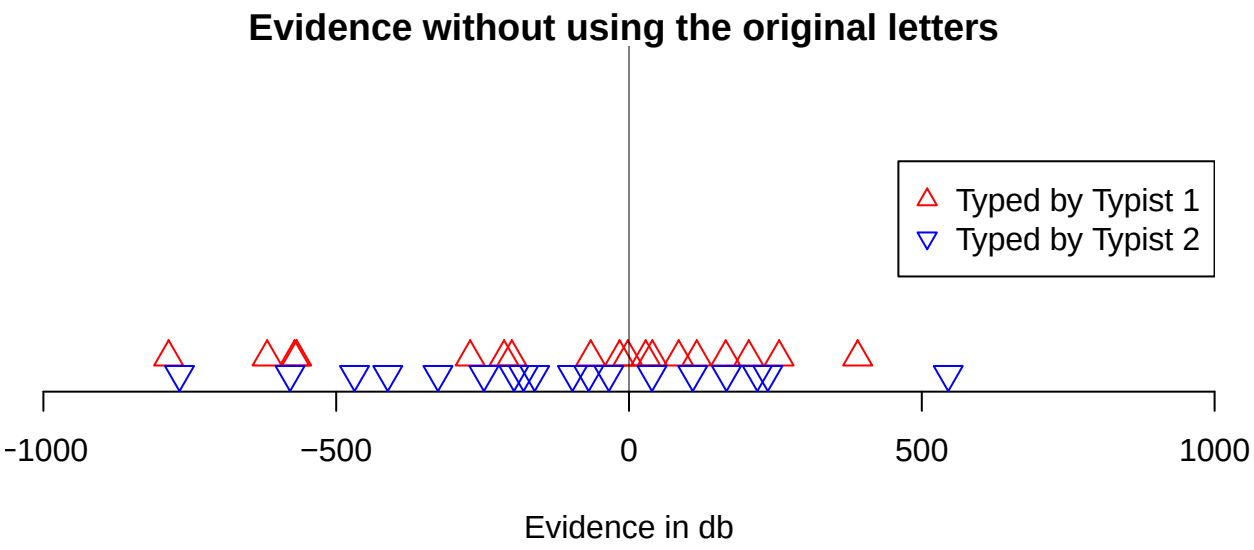
The goal of this study is to attribute the remaining incorrectly transcribed versions of the letters that John Mills wrote to his son to either Typist-1 or Typist-2. Unlike the previous lab, the new letter that we want to classify do not come with their orginal version. We now have access to a total of twenty-four corrupted letters. Among those letters, only 6 of them have the corresponding original version: Typist-1 typed letters 1, 8, 16 and Typist-2 typed letters 4, 9, and 18. Furthermore, we know that Typist-1 typed letters 2, 14, and 17, and Typist-2 typed letters 12, 13, and 19. However, these 6 newly attributed letters do not come with their corresponding original version.

The primary objective of this lab is to classify six letters (letter 5, 6, 20, 21, 23, and 24) that both lack attribution and their corresponding original versions. We will attempt to attribute these letters to either Typist-1 or Typist-2. We will begin by training on a single letter from lab 7 that is attributed to a specific typist to create a sensor model. We then calculate Jaynes evidence for the typist of the other two attributed letters without using their original versions. We will then repeat the same process by training these letters using the model we built in the last lab. This model takes into account the original versions of each letter. Doing so allows us to compare the values of Jaynes evidence in these two cases and assess how the unavailability of the original text effects our predictions. In order to understand how reliable our predictions are, we will train models for both typists on the six letters that are attributed and have original versions, and use this information as evidence to confidently say how often our predictions are correct. Finally, we hope to give attribution to these new six letters that lack both attribution and originals by testing them on our reliable model.

Attribution

Sensor Model Using 1 Letter Each

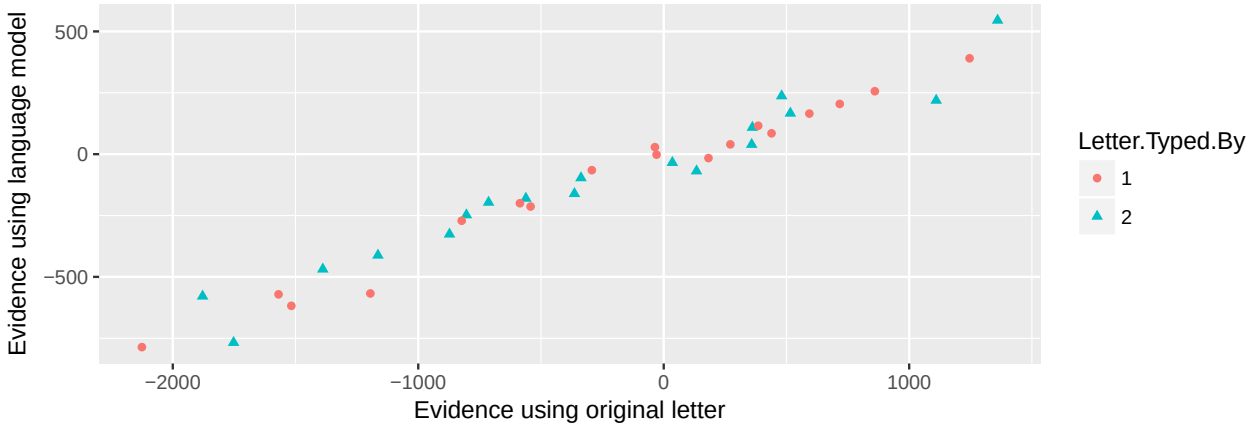
Recall that we need both the corrupted and the original version to build a sensor model. There are three ways to choose a letter by Typist-1, three ways to choose a letter by Typist-2, and 2 * 2 = 4 ways to choose the third letter, giving us a total of 36 different configurations. We used both the original copies of both the chosen letters to build a language model for each of the 36 tests. We trained the language model by adjusting our transition model according to how many times each character was preceded by another in the original letter. That is, we used the six attributed letters (three by each typist) that have their original versions to train on a single letter at a time for each typist and calculate Jaynes's evidence for the other four labeled letters. **The major difference in this part of the lab and the previous one was that we did not use the original versions of these four letters while calculating the evidence.**



Two things were noticably different comparing the evidence obtained from this test to the evidence from the previous lab. #### Difference 1: Magnitude of Evidence Comparing the one-dimensional scatterplot with that of the previous Lab, we noticed that the scale of evidence has shrunk significantly. We are less certain of all our classifications. The reduction in the magnitude of Jaynes Evidence is understandable. Since we do not know the actual sequence of characters in the original letters, we have to maintain a degree of belief of each possible character and take their marginal. This creates uncertainty that was not present in the previous study where we had the original letters. This extra uncertainty is reflected in the magnitude of Jaynes Evidence.

Difference 2: Percent of Correct Predictions

Out of the 36 predictions that we made, 20 of them were correct. In the previous lab, we got only 18 of the 36 predictions correct. A superficial reading of these numbers will makes us beleive that the method of using a language model and sensor model for classification is better than having the original letters. This is clearly a strange result. However, upon further investigation, we noticed that the evidence for each classification is much closer to 0db than in was for the previous test. This means that just by random chance, we could have had a better percentage of prediction. We did not take too much of joy in getting a better prediction percentage rate than the previous Lab.



As the scatter plot shows clearly, the magnitude of Jaynes Evidence using the orignal letter, on average, is much higher compared to the magnitude of Jaynes Evidence using a language model.

Attribution of the 6 new attributed letters

Letter	Typed.By	Evidence.db.	Correct
2	1	-8.4881318	FALSE
14	1	18.1574865	TRUE
17	1	-0.3978632	FLASE
12	2	-125.8807925	TRUE
13	2	-83.3990921	TRUE
19	2	-116.7380986	TRUE

The attribution of the six new attributed letters lacking accompanying text were found to be correct 4 out of 6 times. While this percentage is hardly flattering, it is still an indication that this method of classification can perform better than random chance. Considering that we do not even have the orignial, this method of training a language model and taking the marginal of all posssible values is still a good method for classification.

This method does have a weakness, however, Notice that the magnitude of evidence can be worryingly small. Notice, for example, that the Jaynes Evidence for the classification of Letter 17 is approximately -0.40. The low magnitude of evidence is of some concern. We cannot be as certain in our classification as compared to the previous lab.

Attribution for the 6 new unattributed letters

Letter	Evidence	Attribution
5	9.154493	Typist-1
6	-10.913768	Typist-2
20	-124.989125	Typist-2
21	-50.151795	Typist-2
23	1.991904	Typist-1
24	-30.168813	Typist-2

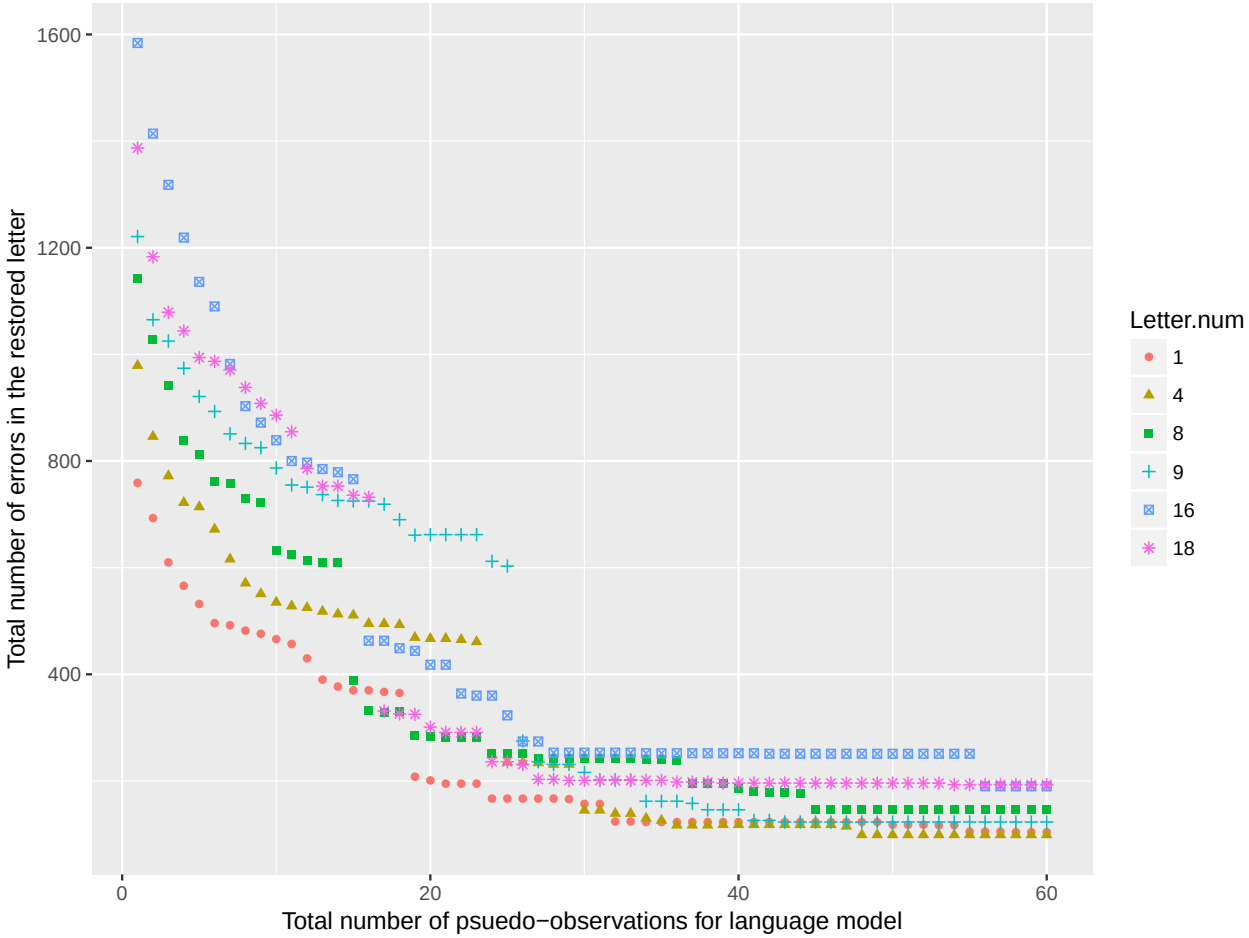
As shown in the results above, we have attributed letters 5 and 23 to Typist-1, and letters 6, 20, 21, and 24 to Typist-2. Since the accuracy of using our language-model is around 55%, we can confidently say that this classification will work better than random chance. However, this is only slightly better than randomly attributing the letters to one of the two typists. We also notice that the magnitude of the evidence of letter 23 at around 2 decibels is a lot smaller than that of the other letters, making us the least confident about this attribution. We believe that having a larger set of uncorrupted, attributed letters to train our model on will result in greater accuracy of our language and sensor models.

3.2 Restoration [Extra Credit]

Errors in corrupted and restored letters

letter.num	errors.corrupted	errors.restored
1	104	759
8	147	1143
16	190	1584
4	99	979
9	123	1221
18	193	1387

Errors in restored letters using different psuedo-observations



The error decreases exponentially as we increase the number of psuedo-observations. This is because the distribution of the language model is flattened by adding extra psuedo-observations. As anticipated, as the number of psuedo-observations exceeds a particular point (the value of this stationary point depends on the leter itself), the restored letter is the same as the corrupted letter.

Unfortunately, we could not decrease the number of errors. However, we did observe that initiliazling the language model by 20 psuedo-observations significantly reduced the number of errors. After this point onwards, the number of errors asympotically converges to the actual number of errors in the corrupted file.

We initialized the language model by 20 psuedo-observations for the restorations of the six attributed letters having no corresponding original text. We are not confident if we have managed to reduce the total number of errors.