COGSCI 316: Machine Learning
Homework 1

Ciprian Bangu

November 6, 2024

Question 1

(a) In the context of our problem, we are using a kind of one-hot encoding to represent the linguistic content of the headlines. That is, we represent each headline as a vector $\vec{\sigma}$ of length L, where the elements of $\vec{\sigma}$ denote the presence of a word in the headline (1), or its absence (0), and L is the total number of unique words in the dataset.

In using this kind of representation, we lose many important aspects of language. Some of these are structural in nature. For example, we lose the ability to account for the number of occurances of each word in a headline (past the first occurance). For example, "The hungry hungry hippo" and "The hungry hippo" would have the same representation in our model, while being plainly different in natural language.

However, this limited ability to capture sentence structure also effects the level of semantic information we can capture with our model. For example, the order of the occurance of the words in a headline is not captured by our vector representation. Therefore, "Man bites dog" and "Dog bites man" would have the same representation given our model, while the natural language versions of the sentences have very different semantic interpretations. So, natural language will distinguish these as two very different sentences - they mean different things, but our model will treat them as identical.

Moreover, we also lose the ability to distinguish between homographs. That is, "bat" (the animal), and "bat" (as in baseball bat), have the same representation in our model - a 1 in the position corresponding to the letters 'bat'. However, a natural language interpretation would categorize these as two different words with different meanings.

These shortcomings are due to the fact that our binary vectorial representation is a signficant reduction in the demensionality of the sentences. We are effectivly reducing a multidimensional object - where the features mentioned above are but a few of the dimensions, to a one-dimensional object: occurance or non-occurance of a word. And, as the homograph example shows, we are even reducing the dimensionality of some of the words (or perhaps better put, 'lingusitic symbols') themselves.

(b) Given that each element of $\vec{\sigma}_i$ is a binary variable, the marginal probability of $\vec{\sigma}_i = 1$ as: $P(\sigma_i = 1|c)$. Since the MaxEnt defines $P(\vec{\sigma}|c) = \frac{e^{\sum_i h_i(c)\sigma_i}}{Z(c)}$, where $Z(c) = \Pi(1 + e^{h_i(c)})$ we can write the marginal probability as: $P(\vec{\sigma}_i = 1|c) = \frac{e^{h_i(c)\sigma_i}}{1+e^{h_i(c)}}$.

Since the $\vec{\sigma}_i$ is a binary variable, the expected value, $\langle\sigma_i\rangle_c = P(\sigma_i = 1|c)$. Therefore, we can write the expected value as: $\langle\sigma_i\rangle_c = \frac{e^{h_i(c)}}{1+e^{h_i(c)}}$.

To express $h_i(c)$ while satisfying the constraing that $\langle \sigma_i \rangle_c = p_i(c)$, we can write: $\frac{e^{h_i(c)}}{1+e^{h_i(c)}} = p_i(c)$. Therefore,

$$p_i(c)(1 + e^{h_i(c)}) = e^{h_i(c)} \tag{1}$$
$$p_i(c) + p_i(c)e^{h_i(c)} = e^{h_i(c)} \tag{2}$$
$$p_i(c) = e^{h_i(c)}(1 - p_i(c)) \tag{3}$$
$$e^{h_i(c)} = \frac{p_i(c)}{1 - p_i(c)} \tag{4}$$
$$h_i(c) = \log\left(\frac{p_i(c)}{1 - p_i(c)}\right) \tag{5}$$

Thus, the fields $h_i(c)$ that satisfy the constraint $\langle \sigma_i \rangle_c = p_i(c)$ are given by $h_i(c) = \log\left(\frac{p_i(c)}{1-p_i(c)}\right)$.

(c) Given the equation for the field above, we can provide an estimate for the value of the fields of each word, specifically using only the training data to estimate the empirical frequency of $p_i(c)$, as well as the pseudo-count of words. Using the pseudo count allows us to avoid issues with zero probabilities. That is, some words might appear in the test data, but not in the training data. This would result in a zero probability for the word, which would make the log ratio undefined.

For the non-sarcastic class, this would be:

$$h_i(0) = \log\left(\frac{p_i(0)}{1 - p_i(0)}\right) \tag{6}$$

namely,

$$h_i(0) = \log\left(\frac{\frac{1}{M_0+1}\left(\sum_{\vec{\sigma}\in\mathcal{D}_0} \sigma_i + 1\right)}{1 - \frac{1}{M_0+1}\left(\sum_{\vec{\sigma}\in\mathcal{D}_0} \sigma_i + 1\right)}\right) \tag{7}$$

where M_0 is the number of headlines in the non-saracstic class, and Z_0 is defined as above.

Similarly, for the sarcastic class, we have:

$$h_i(1) = \log\left(\frac{\frac{1}{M_1+1}\left(\sum_{\vec{\sigma}\in\mathcal{D}_1} \sigma_i + 1\right)}{1 - \frac{1}{M_1+1}\left(\sum_{\vec{\sigma}\in\mathcal{D}_1} \sigma_i + 1\right)}\right) \tag{8}$$

Figure 1 below shows the resulting field estimates for the vocabulary given each class.

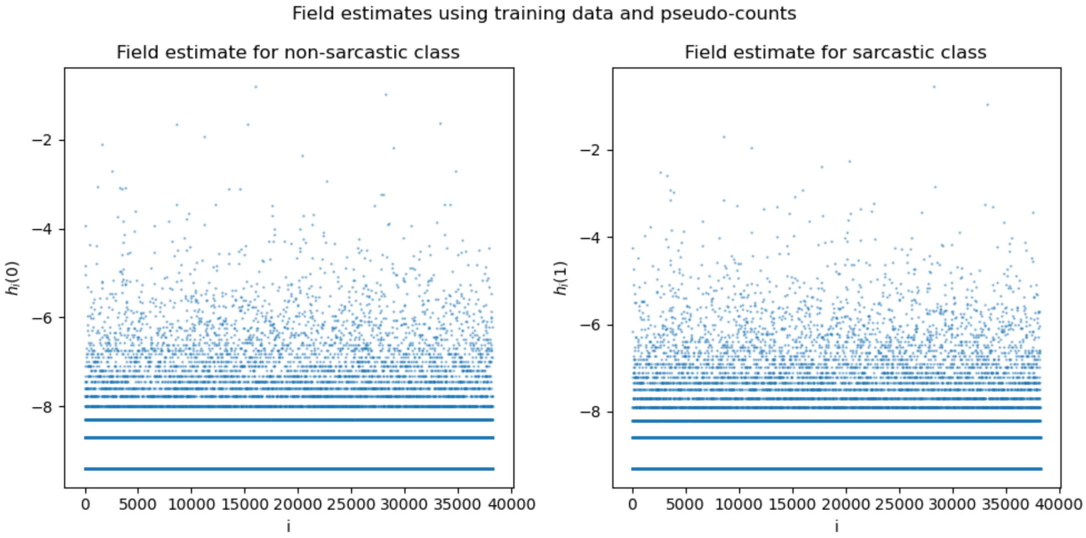

Figure 1: Field estimates for the vocabulary given each class. The x-axis represents the word index in the vocabulary, while the y-axis represents the corresponding field values.

Question 2

a) Figure 2 below shows the histogram of the log conditional probability of a headline's being in the sarcastic class. That is, the histogram of $\log P(\vec{\sigma}|c = 1)$, seperated by whether the headline was categorized as sarcastic or not-sarcastic.

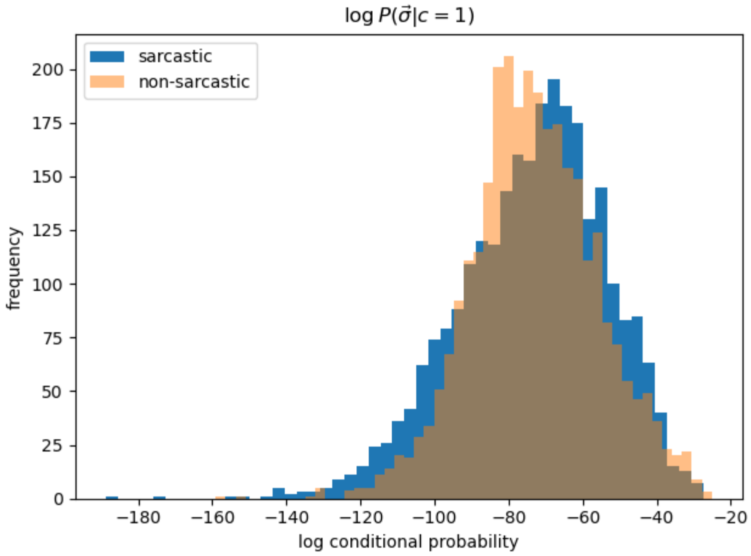

Figure 2: Example Caption

From the figure, we can see that the histograms for the two categories are slightly different. Specifically, the histogram of the sarcastic class is shifted to the right relative to the non-sarcastic class. Since it is a log scale, the closer the value to 0, the higher the probability of the outcome. Therefore, the most common log probability in the sarcastic class is higher than the most common log probability in the non-sarcastic class, given that a headline is sarcastic. This is to be expected if we had any hope of our model being able to classify the headlines correctly: if the distributions were identical, the model would have no way of distinguishing between the two classes via the vectorial representation of the headlines.

b) Figure 3 below shows the histogram of the log odds ratio of a headline based on the class it belongs to. **EXPLAIN**

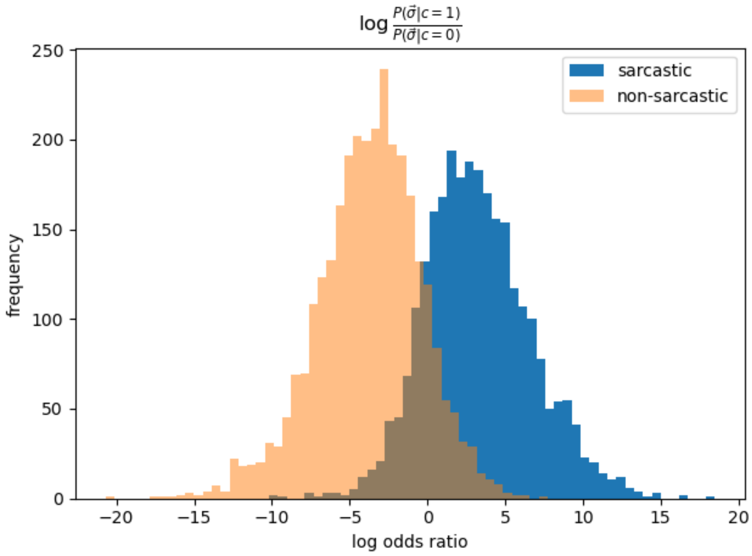

Figure 3: Example Caption

c) **Difference In Means - double check results**

Question 3

a)