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). Thus, "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, 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.

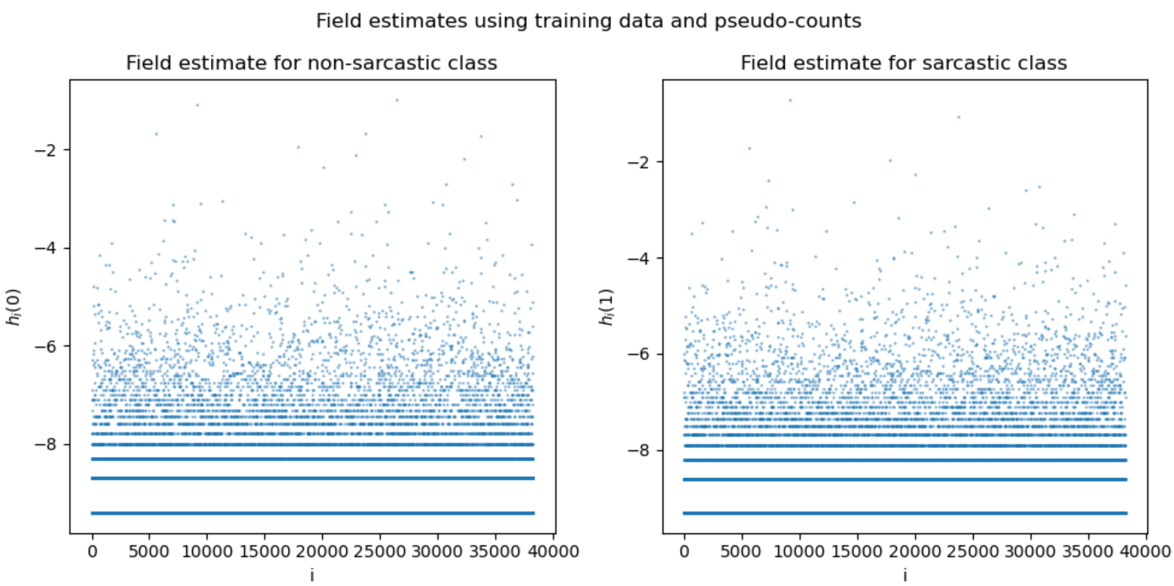

Figure 1: Field estimates for the vocabulary given each class. The x-axis represents the word index in the vocabulary, while the y-axis represents the corresponding field values.

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.

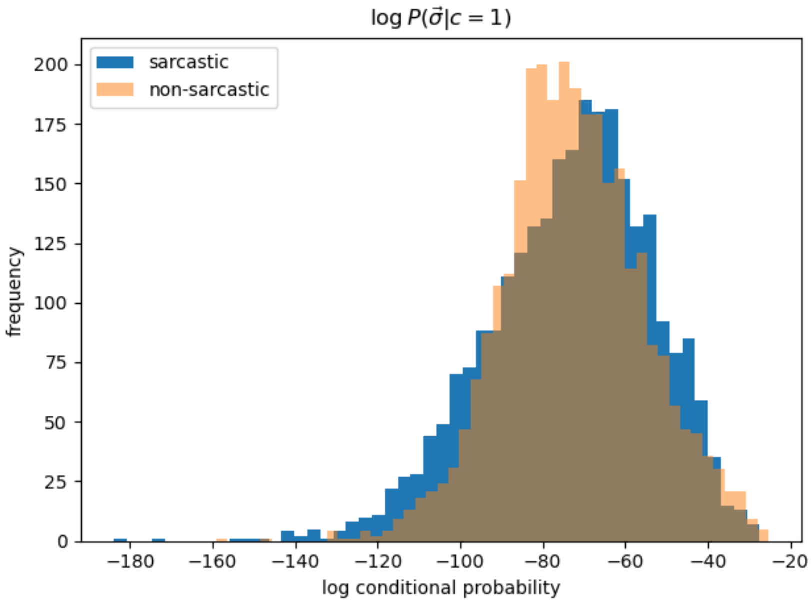

Figure 2: Example Caption

Explain: previous explanation is below but i'm not sure if it's correct

Figure 3 shows the histogram of the log conditional probability of a headline's being in the non-sarcastic class. That is, the histogram of $\log P(\vec{\sigma}|c=0)$, seperated by whether the headline was categorized as sarcastic or not-sarcastic.

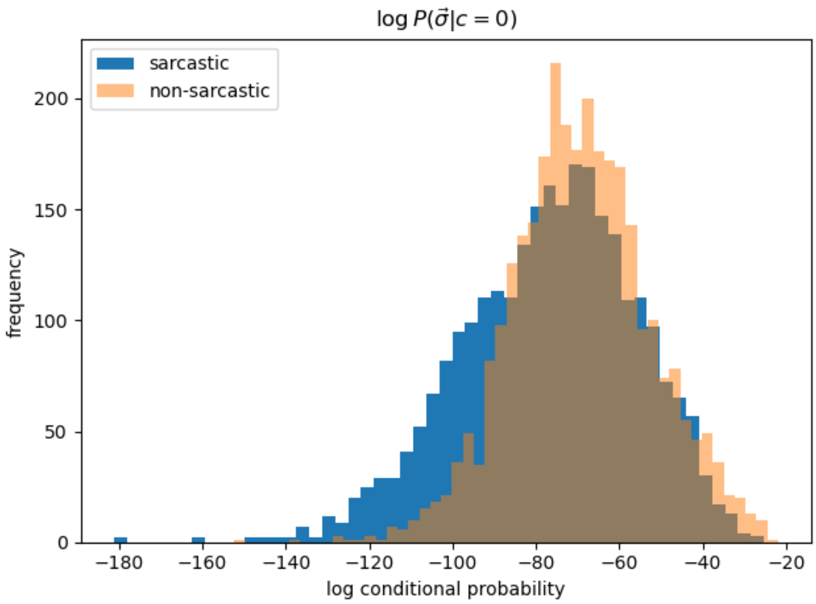

Figure 3: log conditional probability of a headline's being in the non-sarcastic class, given the data in the test set.

b) Figure 4 below shows the histogram of the log odds ratio of a headline based on the class it belongs to.

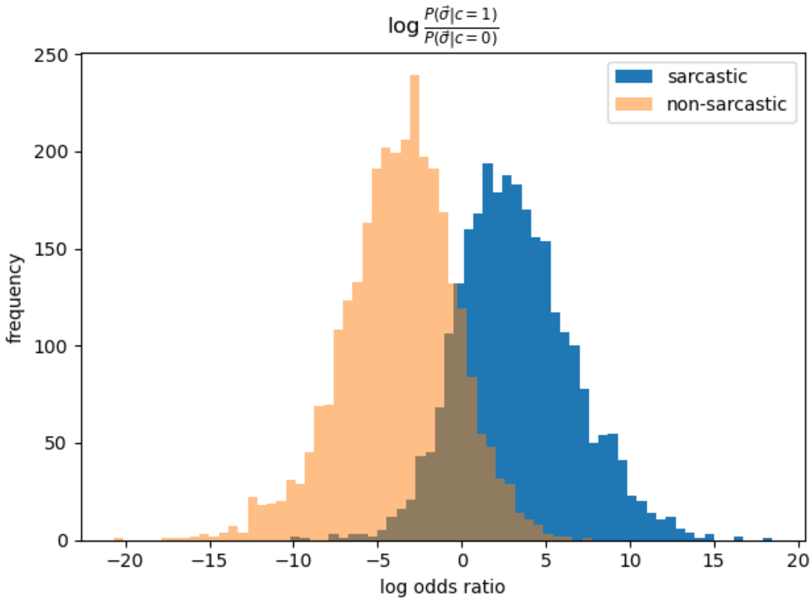

Figure 4: Example Caption

c) We can compute the seperation between the histograms of the two classes by computing the difference in means between their histograms, for the conditions above. Namely:

$$\frac{|\langle x \rangle - \langle y \rangle|}{\sqrt{\sigma_x \sigma_y}} \tag{9}$$

where x an y are the means of the log likelyhoods, and σ_x and σ_y are the standard deviations of the log likelyhoods for the two classes.

	$\log P(\vec{\sigma}\|c=1)$	$\log P(\vec{\sigma}\|c=0)$	$\log \frac{P(\vec{\sigma}\|c=1)}{P(\vec{\sigma}\|c=0)}$
Diff. in means	0.04209	0.40834	2.0577

Table 1: Difference in means of the histograms of the two classes, given the different conditions.

The numerical results in table 1 reflect the visual results of the histograms. Namely, we see that the seperation between the two classes is not very distinct when considering only the log conditional probabilitie - most of the histograms are overlapping. However, when considering the log odds ratio, the difference in means is 1 to 2 orders of magnitude greater, which is reflectd in the distinct seperation between the histograms seen in Figure 4.

Question 3

a) Bayes theorem tells us that:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{10}$$

Thus, we can write the posterior probability of a headline being sarcastic given the data as:

$$P(c=1|\vec{\sigma}) = \frac{P(\vec{\sigma}|c=1)P(c=1)}{P(\vec{\sigma})} \tag{11}$$

Likewise,

$$P(c=0|\vec{\sigma}) = \frac{P(\vec{\sigma}|c=0)P(c=0)}{P(\vec{\sigma})} \tag{12}$$

Thus, the ratio between the two can be written as:

$$\frac{P(c=1|\vec{\sigma})}{P(c=0|\vec{\sigma})} = \frac{P(\vec{\sigma}|c=1)P(c=1)}{P(\vec{\sigma}|c=0)P(c=0)} \tag{13}$$

Figure 5 shows the results of the above calculation for each of the headlines in the test set.

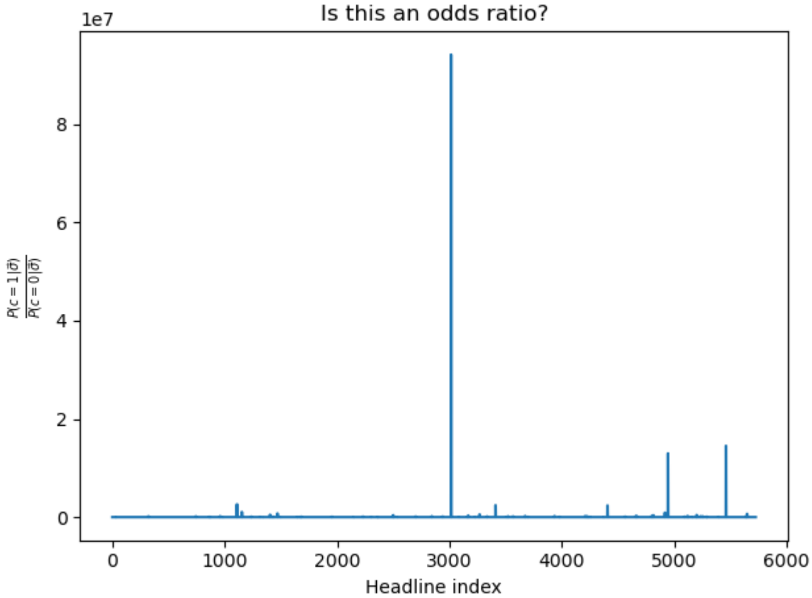

Figure 5: Likelyhood **Is this the right word?** of each headline in the test set being sarcastic.

Inspecting the data, we see that these scores pass the eye test as well. Namely, the highest score corresponds to headling number 3013 : *nation wishes area man were a creep, but, ugh, he's actually really fucking nice.* One imediately recognizes this headline as sarcastic, without needing to check the category in the data (although it is confirmed by doing so). Moreover, the lowest score is given to headline number 4346: *here's what cops and their supporters are saying about the sandra bland arrest video.* Again, on it's face, this is a non-sarcastic headline.

b) The above results suggest that our classifier, at least in the extreme cases, can accuartely classify headlines as sarcastic or non-sarcastic. However, to assess the true accuracy of the model, we can plot the Reciever-Operating-Characteristic Curve (ROC), and compute the Area under the Curve (AUC).

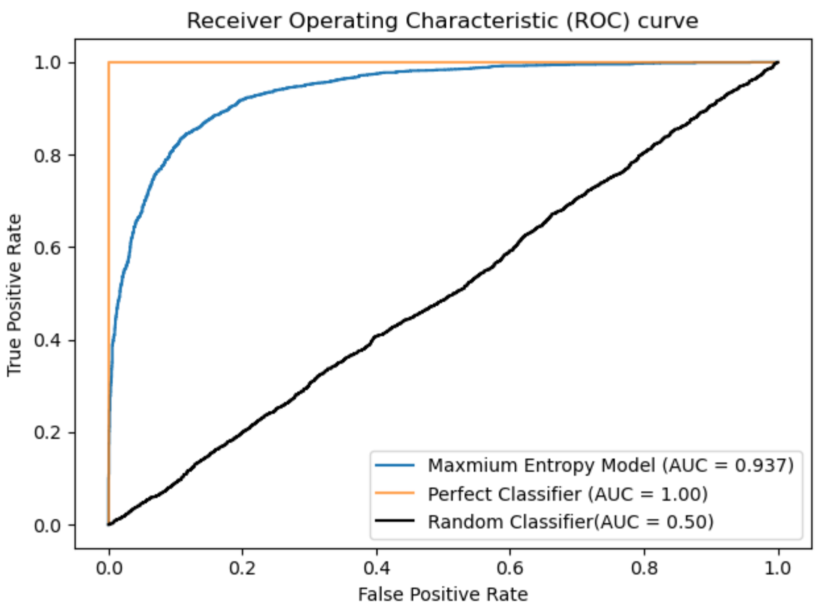

Figure 6: Reciever Operating Characteristic curve for the Maximum Entropy Model, as well as a Random Classifier and a Perfect Classifier. The AUC is also reported for each model, indicating that the Max Entropy Model is signficantly better than chance, and slightly worse than perfect.

Per Figure 6 above, we see that the Max Entropy Model results in an AUC of 0.94. This means that given a random sarcastic headline and a random non-sarcastic headline, the model will correctly classify the sarcastic headline as sarcastic 94% of the time. This is much better than chance (AUC = 0.51), and slightly worse than a perfect classifier (AUC = 1).

c) The ROC and AUC of the model reflect the relationship between the log odds ratio of the two categories we saw in Figure 4. Namely, the distribtions had little overlap (as seen in the difference in means), meaning there was little room for confusion between the two classes. This seperation between the two distributions allows the model to be very accurate in classifying the headlines as sarcastic or non-sarcastic, as shown by the AUC of 0.94.

d) The mutual information between the encoded headline and the class label is given by:

$$MI(\vec{\sigma}, c) = \sum_{\vec{\sigma}, c} P(\vec{\sigma}, c) \log \left(\frac{P(\vec{\sigma}, c)}{P(\vec{\sigma})P(c)} \right) \tag{14}$$

which can be approximated by:

$$MI(\vec{\sigma}, c) \approx \frac{1}{M} \sum_{d} \log \left(\frac{P(\vec{\sigma}_d, c_d)}{P(\vec{\sigma}_d)P(c_d)} \right) \tag{15}$$

Where M is the number of headlines in the data, and d is the index of the headline.

FINISH THIS LATER

Question 4

a) The Kullback-Leiber (KL) divergence between two distributions P and Q is given by:

$$D_{KL}(P||Q) = \sum_{x} P(x) \log \left(\frac{P(x)}{Q(x)} \right) \tag{16}$$

Here, we can compute the KL divergence between two models, $P(\vec{\sigma}|c = 1)$ and $P(\vec{\sigma}|c = 0)$, as either:

$$D_{KL}(P(\vec{\sigma}|c=1)||P(\vec{\sigma}|c=0)) = \sum_{\vec{\sigma}} P(\vec{\sigma}|c=1)\log\left(\frac{P(\vec{\sigma}|c=1)}{P(\vec{\sigma}|c=0)}\right) \tag{17}$$

or,

$$D_{KL}(P(\vec{\sigma}|c=0)||P(\vec{\sigma}|c=1)) = \sum_{\vec{\sigma}} P(\vec{\sigma}|c=0)\log\left(\frac{P(\vec{\sigma}|c=0)}{P(\vec{\sigma}|c=1)}\right) \tag{18}$$

In the first case, we have that **CHECK**

$$D_{KL}(P(\vec{\sigma}|c=0)||P(\vec{\sigma}|c=1)) = 4.9648 \tag{19}$$

and in the second case, we have that **CHECK**

$$D_{KL}(P(\vec{\sigma}|c=1)||P(\vec{\sigma}|c=0)) = 4.7481 \tag{20}$$

These are different **Because?**

b) Since our classes are binary, we can denote them as c and c'. Thus the KL divergence between the two classes is given by:

$$D_{KL}(P(\vec{\sigma}|c)||P(\vec{\sigma}|c')) = \sum_{\vec{\sigma}} P(\vec{\sigma}|c)\log\left(\frac{P(\vec{\sigma}|c)}{P(\vec{\sigma}|c')}\right) \tag{21}$$

Recall that: $P(\vec{\sigma}|c) = \frac{e^{\sum_i h_i(c)\sigma_i}}{Z(c)}$. So, we can rewrite the D_{KL} as:

$$D_{KL}(P(\vec{\sigma}|c)||P(\vec{\sigma}|c')) = \sum_{\vec{\sigma}} P(\vec{\sigma}|c)\left(\sum_i h_i(c)\sigma_i - \log Z(c) - \sum_i h_i(c')\sigma_i + \log Z(c')\right) \tag{22}$$

$$D_{KL}(P(\vec{\sigma}|c)||P(\vec{\sigma}|c')) = \sum_{\vec{\sigma}} P(\vec{\sigma}|c)\left(\sum_i (h_i(c) - h_i(c'))\sigma_i + \log\frac{Z(c')}{Z(c)}\right) \tag{23}$$

Distributing, we get:

$$D_{KL}(P(\vec{\sigma}|c)||P(\vec{\sigma}|c')) = \sum_{\vec{\sigma}}\sum_i P(\vec{\sigma}|c)(h_i(c) - h_i(c'))\sigma_i + \sum_{\vec{\sigma}} P(\vec{\sigma}|c)\log\frac{Z(c')}{Z(c)} \tag{24}$$

$$D_{KL}(P(\vec{\sigma}|c)||P(\vec{\sigma}|c')) = \sum_i\sum_{\vec{\sigma}} P(\vec{\sigma}|c)(h_i(c) - h_i(c'))\sigma_i + \log\frac{Z(c')}{Z(c)} \tag{25}$$

$$D_{KL}(P(\vec{\sigma}|c)||P(\vec{\sigma}|c')) = \sum_i \langle\sigma_i\rangle_c(h_i(c) - h_i(c')) + \log\frac{Z(c')}{Z(c)} \tag{26}$$

Finally, substituting the values of the fields, we get:

$$D_{KL}(P(\vec{\sigma}|c)||P(\vec{\sigma}|c')) = \sum_i\left[\frac{e^{h_i(c)}}{1+e^{h_i(c)}}(h_i(c) - h_i(c')) + \log\frac{1+e^{h_i(c')}}{1+e^{h_i(c)}}\right] \tag{27}$$

From equation 27, we can see that the KL divergence can be expressed in terms of the fields of the two classes, and a sum over the individual words (i)

c) The Chernoff bound for the classification error is given by:

$$P_{error} \leq e^{-ND_{KL}(P(\vec{\sigma}|c)||P(\vec{\sigma}|c'))} \tag{28}$$

Thus, to compute how many articles our model needs to see to confidently guess the class of a headline within a certain error bound, we will have to compute N from equation 28.

Specifically, if we take our maximum error to be 10^{-10}, then

$$\begin{aligned} 10^{-10} &> e^{-ND_{KL}(P(\vec{\sigma}|c)||P(\vec{\sigma}|c'))} \\ \log 10^{-10} &> -ND_{KL}(P(\vec{\sigma}|c)||P(\vec{\sigma}|c')) \\ N &< \frac{-\log 10^{-10}}{D_{KL}(P(\vec{\sigma}|c)||P(\vec{\sigma}|c'))} \end{aligned}$$

So, if the newspaper only produces sarcastic headlines (c = 1), then the model will need to see less than 4.64 headlines. If the newspaper only produces non-sarcastic headlines (c = 0), then the model will need to see less than 4.85 headlines.