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 occurrences of each word in a headline (past the first occurrence). 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 occurrence 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 significant reduction in the dimensionality of the sentences. We are effectively reducing a multidimensional object - where the features mentioned above are but a few of the dimensions, to a one-dimensional object: occurrence or non-occurrence 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 constraint 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,

$$
\begin{aligned}
p_i(c)(1 + e^{h_i(c)}) &= e^{h_i(c)} \\
p_i(c) + p_i(c)e^{h_i(c)} &= e^{h_i(c)} \\
p_i(c) &= e^{h_i(c)}(1 - p_i(c)) \\
e^{h_i(c)} &= \frac{p_i(c)}{1 - p_i(c)} \\
h_i(c) &= \log\left(\frac{p_i(c)}{1 - p_i(c)}\right)
\end{aligned}
$$

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:

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

namely,

$$
h_i(0) = \log\left(\frac{\frac{1}{M_0+1}\left(\sum_{\vec{\sigma}\in\mathcal{D}_0} \sigma_i + 1\right)}{1 - \frac{1}{M_0+1}\left(\sum_{\vec{\sigma}\in\mathcal{D}_0} \sigma_i + 1\right)}\right) \tag{2}
$$

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

Similarly, for the sarcastic class, we have:

$$
h_i(1) = \log\left(\frac{\frac{1}{M_1+1}\left(\sum_{\vec{\sigma}\in\mathcal{D}_1} \sigma_i + 1\right)}{1 - \frac{1}{M_1+1}\left(\sum_{\vec{\sigma}\in\mathcal{D}_1} \sigma_i + 1\right)}\right) \tag{3}
$$

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

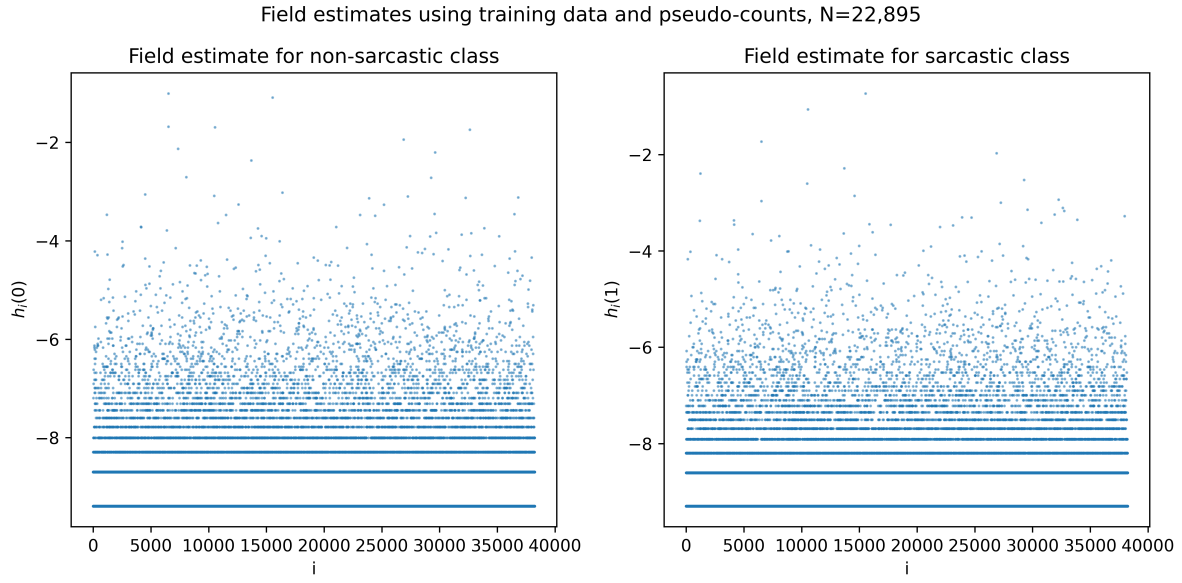

Figure 1: Field estimates for the vocabulary given each class. The x-axis represents the word index in the vocabulary, while the y-axis represents the corresponding field values.

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 $\log P(\vec{\sigma}|c = 1)$, for both sarcastic and non-sarcastic headlines in the test set.

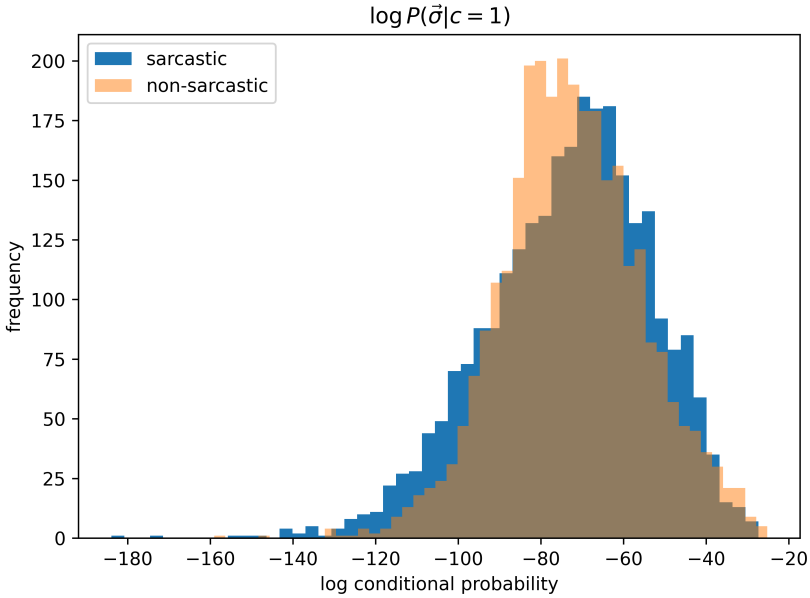

Figure 2: log conditional probability of a headline's being in the sarcastic class, given the data in the test set.

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 $\log P(\vec{\sigma}|c = 0)$, for both sarcastic and non-sarcastic headlines in the test set.

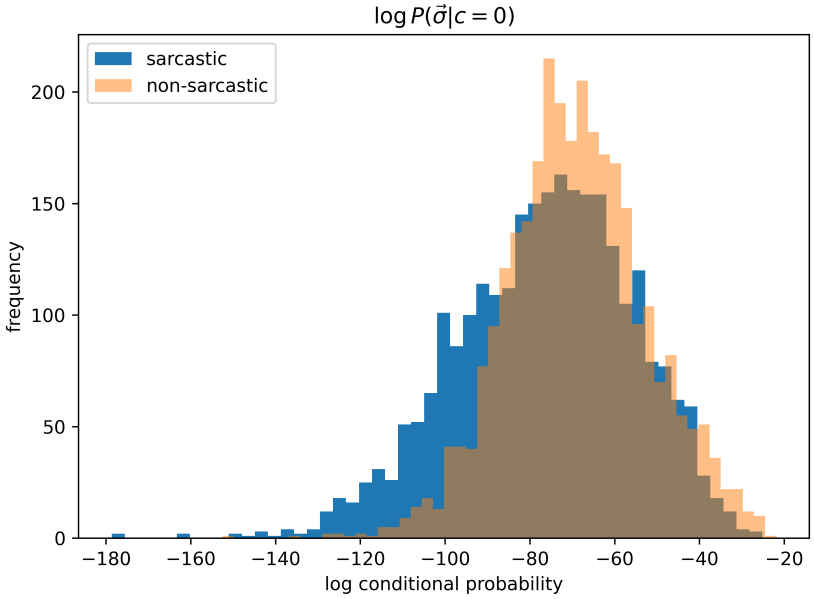

Figure 3: log conditional probability of a headline's being in the non-sarcastic class, given the data in the test set.

From the figures, we see that there is significant overlap between the two distributions. This means

that the model will have a difficult time distinguishing between the two classes, with slightly less overlap in the latter case. This suggests that the model will have a difficult time accurately classifying the headlines as sarcastic or non-sarcastic based on the log conditional probabilities alone.

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

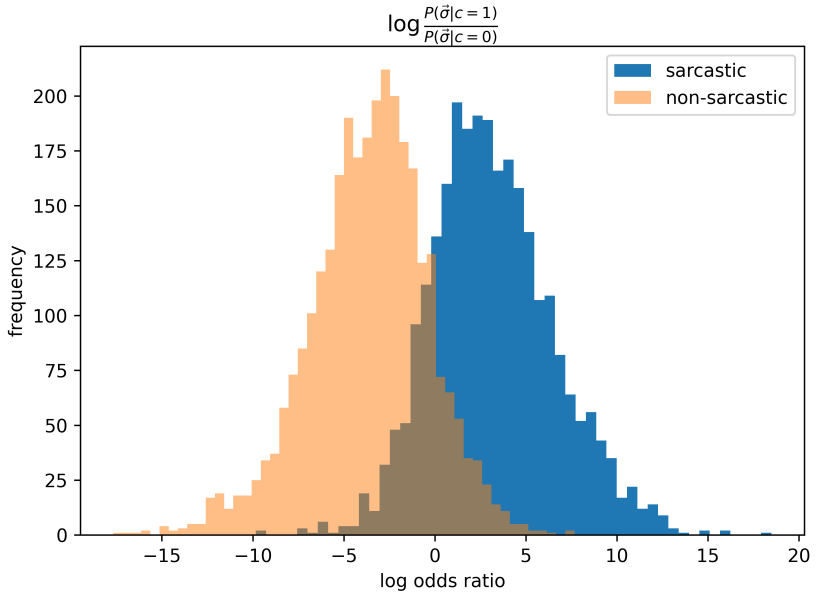

Figure 4: log odds ratio of a headline being sarcastic, separated by sarcastic and non-sarcastic headlines.

Fig. 4 shows a much clearer separation between the distribution of the two classes. Thus, given the log odds ratio, the model should be able to accurately classify the headlines.

c) The intuition we formed from Figs 2 - 4 can be expressed more precisely if we compute the separation between the histograms of the two classes. We can do this by computing the difference in means between their histograms, for the conditions above. Namely:

$$\frac{|\langle x \rangle - \langle y \rangle|}{\sqrt{\sigma_x \sigma_y}} \tag{4}$$

where x an y are the means of the log likelihoods, and σ_x and σ_y are the standard deviations of the log likelihoods 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 separation between the two classes is not very distinct when considering only the log conditional probabilities - 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 reflected in the distinct separation between the histograms seen in Figure 4.

Question 3

a) Bayes theorem tells us that:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{5}$$

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

$$P(c = 1|\vec{\sigma}) = \frac{P(\vec{\sigma}|c = 1)P(c = 1)}{P(\vec{\sigma})} \tag{6}$$

Likewise,

$$P(c = 0|\vec{\sigma}) = \frac{P(\vec{\sigma}|c = 0)P(c = 0)}{P(\vec{\sigma})} \tag{7}$$

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

$$\frac{P(c = 1|\vec{\sigma})}{P(c = 0|\vec{\sigma})} = \frac{P(\vec{\sigma}|c = 1)P(c = 1)}{P(\vec{\sigma}|c = 0)P(c = 0)} \tag{8}$$

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

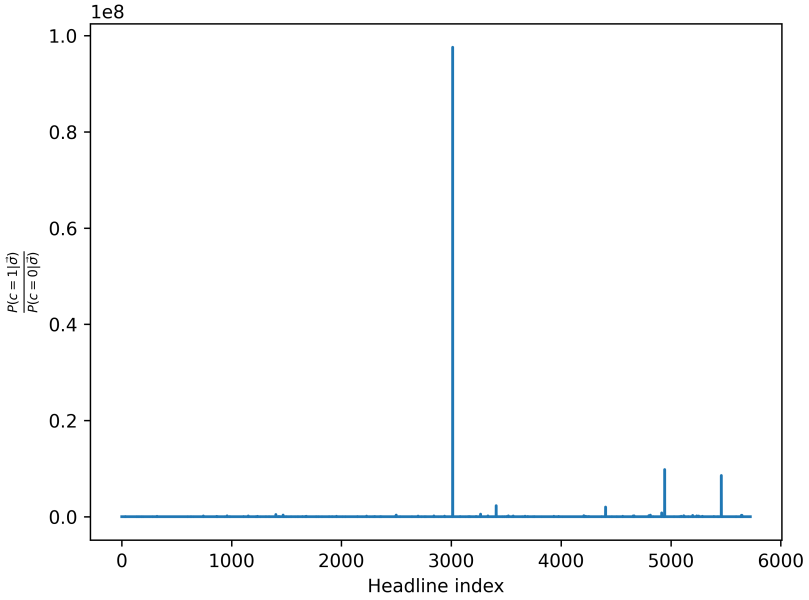

Figure 5: Odds ratio comparing sarcastic to non-sarcastic classifications for each headline in the test set.

Inspecting the data, we see that these scores pass the eye test as well. Namely, the highest score corresponds to headline number 3013 : *nation wishes area man were a creep, but, ugh, he's actually really fucking nice.* One immediately 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 accurately classify headlines as sarcastic or non-sarcastic. However, to assess the true accuracy of the model, we can plot the Receiver-Operating-Characteristic Curve (ROC), and compute the Area under the Curve (AUC).

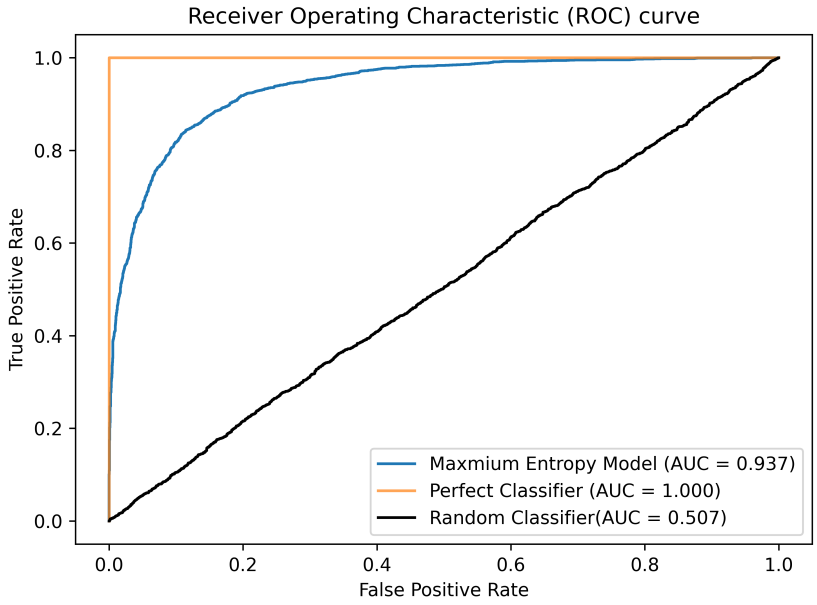

Figure 6: Reciever Operating Characteristic curve for the Maximum Entropy Model, as well as a Random Classifier and a Perfect Classifier. The AUC is also reported for each model, indicating that the Max Entropy Model is significantly better than chance, and slightly worse than perfect.

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 distributions had little overlap (as seen in the difference in means), meaning there was little room for confusion between the two classes. This separation 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.

Specifically, suppose we have some threshold t where if the log-odds ratio is greater than t, the sample is classified as sarcastic; non-sarcastic otherwise. This allows us to define the probability the log-odds of a sarcastic sample exceeds the threshold, the True Positive Rate, as:

$$TPR(t) = P(log - odds > t | c = 1) \tag{9}$$

and similarly, the probability the log-odds of a non sarcastic sample exceeds the threshold, the False Positive Rate, as:

$$FPR(t) = P(log - odds > t | c = 0) \tag{10}$$

Per Figure 4. we can confidently approximate these as distributions, so we can write the TPR and FPR as:

$$TPR(t) = \int_t^\infty f_{log-odds|c=1}(x)dx \tag{11}$$

$$FPR(t) = \int_t^\infty f_{log-odds|c=0}(x)dx \tag{12}$$

Namely, the area under the PDF of each distribution, which is approximated by our histograms, between t and ∞. Fig 4. shows that the overlap between the two distributions is small. So for values of t greater than 7 (but less than 20), the TPR is small but the FPR is 0 - we should expect

few true positives, but also no false positives. For t smaller than 5, the TPR will increase (the area swept under the PDF(c=1) increases), but so will the FPR (the area swept under the PDF(c=0) also increases). Eventually, the area under both distributions will be swept, so the ROC tends to 1.

The AUC is the area under the ROC curve. Plugging in the functional form of the TPR and FPR we established above, we can write the AUC as:

$$AUC = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{x} f_{\text{log-odds}|c=0}(y)\, dy \right) f_{\text{log-odds}|c=1}(x)\, dx \tag{13}$$

$$\int_{-\infty}^{\infty} F_{\text{log-odds}|c=0}(x)\, f_{\text{log-odds}|c=1}(x)\, dx \tag{14}$$

Where $F_{\text{log-odds}|c=0}(x)$ is the cumulative distribution function of the log-odds of the non-sarcastic class (the integral of the PDF).

Thus, the AUC is the PDF of the sarcastic class multiplied by the CDF of the non-sarcastic class, as shown in the equation above.

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

$$MI(\vec{\sigma}, c) = \sum_{\vec{\sigma}, c} P(\vec{\sigma}, c) \log \left(\frac{P(\vec{\sigma}, c)}{P(\vec{\sigma})P(c)} \right) \tag{15}$$

which can be approximated by:

$$MI(\vec{\sigma}, c) \approx \frac{1}{M} \sum_{d} \log \left(\frac{P(\vec{\sigma}_d, c_d)}{P(\vec{\sigma}_d)P(c_d)} \right) \tag{16}$$

Where M is the number of headlines in the data, and d is the index of the headline. For the training set, the estimated Mutual Information is 0.553; for the test set it is 0.571.

Question 4

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

$$D_{KL}(P||Q) = \sum_{x} P(x) \log \left(\frac{P(x)}{Q(x)} \right) \tag{17}$$

Where P is the 'ground-truth' distribution, and Q is the approximating distribution. The KL divergence measures the difference between the two distributions - i.e., how well Q approximates P.

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

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

or,

$$D_{KL}(P(\vec{\sigma}|c = 0)||P(\vec{\sigma}|c = 1)) = \sum_{\vec{\sigma}} P(\vec{\sigma}|c = 0) \log \left(\frac{P(\vec{\sigma}|c = 0)}{P(\vec{\sigma}|c = 1)} \right) \tag{19}$$

In the first case, we have that

$$D_{KL}(P(\vec{\sigma}|c = 0)||P(\vec{\sigma}|c = 1)) = 4.748 \tag{20}$$

and in the second case, we have that

$$D_{KL}(P(\vec{\sigma}|c=1)||P(\vec{\sigma}|c=0)) = 4.965 \tag{21}$$

These two KL Divergences are different because they measure different things. In the first case, we measure the ability of $P(\vec{\sigma}|c=1)$ to approximate $P(\vec{\sigma}|c=0)$, while in the second case, we measure the ability of $P(\vec{\sigma}|c=0)$ to approximate $P(\vec{\sigma}|c=1)$. And there is no reason for these values to be the same. Suppose P is the ground-truth distribution. If P and Q disagree about the probability of a certain event that is common in P, the KL divergence will be large because Q will not be able to capture this property of P. And vice-versa if Q is used as the ground-truth distribution.

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:

$$D_{KL}(P(\vec{\sigma}|c)||P(\vec{\sigma}|c')) = \sum_{\vec{\sigma}} P(\vec{\sigma}|c) \log\left(\frac{P(\vec{\sigma}|c)}{P(\vec{\sigma}|c')}\right) \tag{22}$$

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:

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

Distributing, we get:

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

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

$$D_{KL}(P(\vec{\sigma}|c)||P(\vec{\sigma}|c')) = \sum_i \left[\frac{e^{h_i(c)}}{1+e^{h_i(c)}}(h_i(c) - h_i(c')) + \log \frac{1+e^{h_i(c')}}{1+e^{h_i(c)}}\right] \tag{23}$$

From equation 22, 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:

$$P_{error} \leq e^{-ND_{KL}(P(\vec{\sigma}|c)||P(\vec{\sigma}|c'))} \tag{24}$$

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

$$10^{-10} \leq e^{-ND_{KL}(P(\vec{\sigma}|c)||P(\vec{\sigma}|c'))}$$
$$\log 10^{-10} \leq -ND_{KL}(P(\vec{\sigma}|c)||P(\vec{\sigma}|c'))$$
$$N \geq \frac{-\log 10^{-10}}{D_{KL}(P(\vec{\sigma}|c)||P(\vec{\sigma}|c'))}$$

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

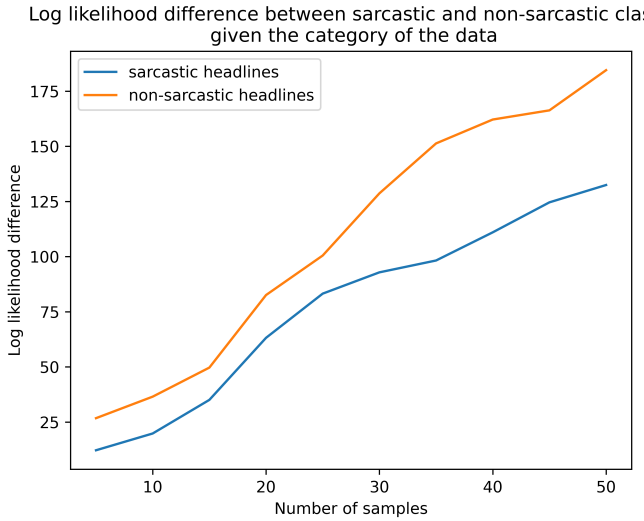

Figure 7: Log Likelihood differences between the sarcastic model and non-sarcastic model, given the dataset seen is either sarcastic or non-sarcastic. The x axis represents the number of samples seen, while the y axis represents the log likelihood difference.

d) Figure 7 above shows log likelihood differences between the two models, based on the number of samples seen, separated by the sarcastic and non-sarcastic datasets. The maximum number of samples in the plot is 50, so the data are limited. Thus, the trend is only approximate. Extrapolating, however, we should expect that we should see a positive linear trend in the log likelihood difference as the number of samples increases. i.e., as more samples are added, the log-likelihood of the ground truth model goes up ($P \to 1 \therefore \log(P) \to 0$), while the log-likelihood of the approximating model goes down ($Q \to 0 \therefore log(Q) \to -\infty$). That is, each additional sample should increase the log likelihood difference by the same amount. More concretely, the slope of each curve should match the KL divergence between the distributions where the sarcastic model is the ground truth for the sarcastic datasets and vice versa.

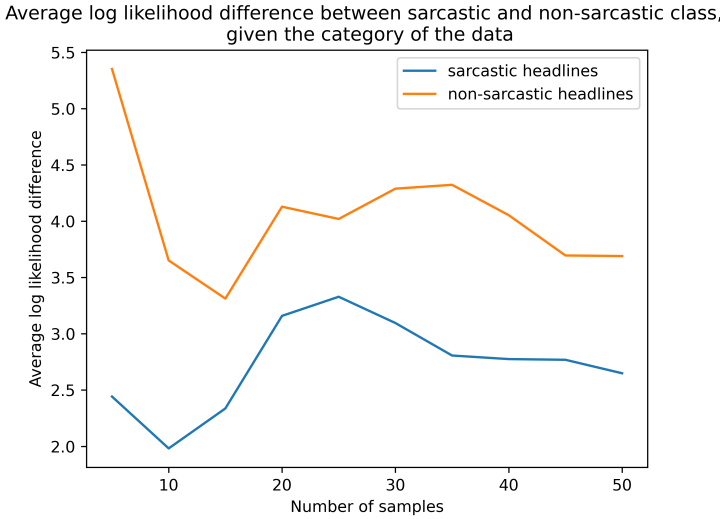

Figure 8: Average Log Likelihood Differences between the sarcastic model and non-sarcastic model, given the dataset seen is either sarcastic or non-sarcastic. The x axis represents the number of samples seen, while the y axis represents the average log likelihood difference.

e) Figure 8 above shows the average log-likelihood differences of the models, given class-homogenous datasets of different sizes. As we can see from the plot, the asymptotic trend is that of reaching the same value as the slopes in Figure 9, namely the KL Divergence between the models.

Question 5

Finally, we can observe how these results change as we change the size of our training set. Recall, our field estimates throughout the this analysis were based only on the training data, $N = 22,895$ samples. This represented 80% of the total data, so the model could learn the distribution of the data well.

First, let us reduce the training set drastically to $N = 100$ samples. From the field estimates, we can see $h_i(c)$ is much more uniform than in figure 1. This reflects the fact that the model has seen much less data, and thus has imbibed information about fewer words in the vocabulary. As we increase the number of samples, the fields begin to be less uniform. Figure 7 demonstrates this trend.

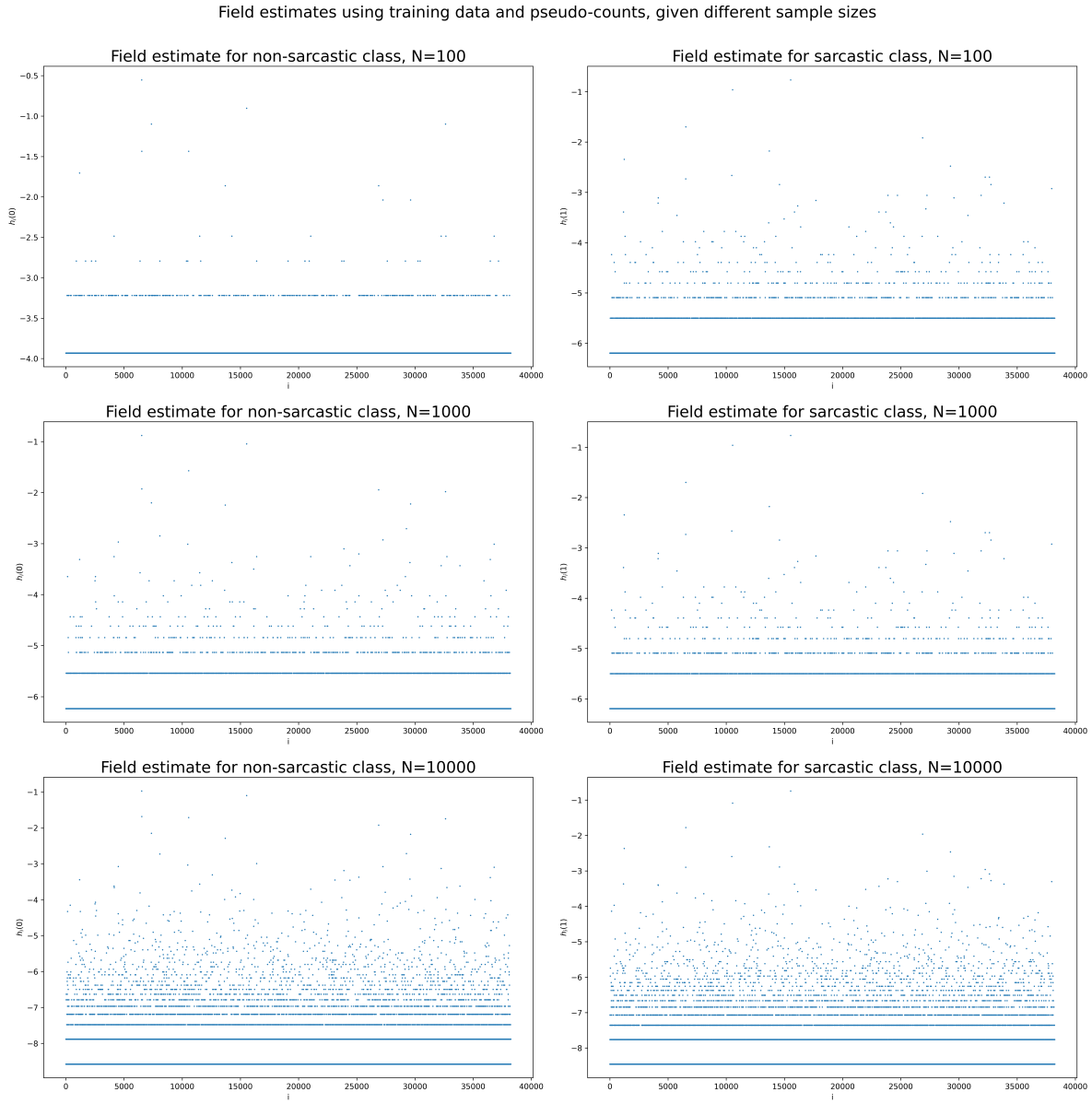

Figure 9: Field estimates for the vocabulary given each class, for different training sample sizes. The x-axis represents the word index in the vocabulary, while the y-axis represents the corresponding field values.

	$\log P(\vec{\sigma}\|c=1)$	$\log P(\vec{\sigma}\|c=0)$	$\log \frac{P(\vec{\sigma}\|c=1)}{P(\vec{\sigma}\|c=0)}$
N = 100	0.111	0.281	1.02
N = 1000	0.115	0.334	1.37
N = 10000	0.086	0.399	1.92

Table 2: Difference in means of the histograms of the two classes, given the different conditions and different sample sizes.

Figure 10 below shows the effect the size of the training set has on the model's performance. As we can see from the ROC curves, and corresponding AUCs, at only 100 samples, the model's performance is random. As expected, as the sample size increases, the difference between the two distributions increases, and so does the model's performance. This is supported numerically in Table 2 above, which shows $\log \frac{P(\vec{\sigma}|c=1)}{P(\vec{\sigma}|c=0)}$; as the distributions become more different, as the sample size increases.

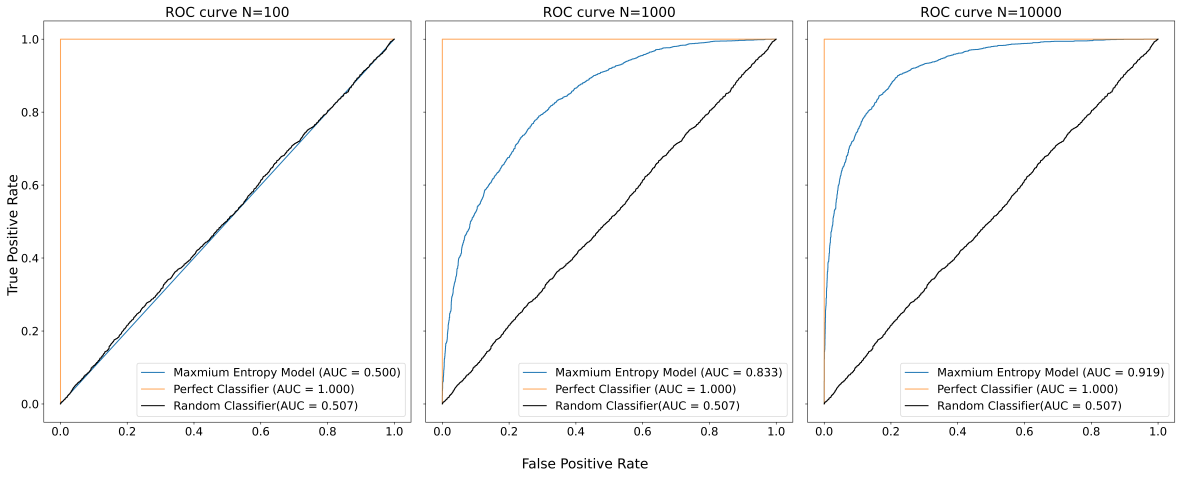

Figure 10: ROC curve and corresponding AUC for the Maximum Entropy Model given training sample sizes. As N increases, the model's performance increases, as noted by the increase in AUC.