LIAR Dataset Prediction Model: Approach and Mathematics

Satwik (2022CS51150) Aneeket Yadav (2022CS11116)

October 2024

1 Introduction

This report details our approach to making predictions on the LIAR dataset, which classifies political statements into six categories: `pants-fire`, `false`, `barely-true`, `half-true`, `mostly-true`, and `true`. Our method utilizes a modified Bernoulli Naive Bayes algorithm, focusing on the historical accuracy of speakers rather than the content of their statements.

2 Data Selection

Despite the availability of various input features such as the statement text, speaker identity, and job description, our experimental results showed that these did not lead to effective predictions. Instead, we focused on the following key information:

- The number of times each speaker's statements were classified as `pants-fire`, `false`, `barely-true`, `half-true`, or `mostly-true` in the past.
- The overall distribution of `true` statements in the dataset.

3 Approach

Our approach can be summarized in the following steps:

1. Estimate the probability of a `true` statement for each speaker.
2. Convert historical counts of other labels into probabilities.
3. Predict the most likely class based on these probabilities.

3.1 Estimating `True` Statement Probability

We estimated the probability of a `true` statement by calculating the fraction of `true` labels in the entire training dataset. This global estimate is used for all speakers due to the lack of speaker-specific `true` counts.

3.2 Converting Counts to Probabilities

For the other five classes, we converted the historical counts into probabilities by normalizing them. This normalization accounts for the varying number of statements made by different speakers.

3.3 Making Predictions

To make a prediction, we compare the probabilities for each class and select the one with the highest probability.

4 Mathematical Details

Our implementation uses a modified Bernoulli Naive Bayes model. Here's a breakdown of the key mathematical components:

4.1 Class Probability Estimation

For each class c, we estimate the log probability as:

$$\log P(c) = \log \frac{N_c + \alpha}{N + K\alpha}$$

Where:

- N_c is the count of class c in the training set,
- N is the total number of samples,
- K is the number of classes (6 in our case),
- α is the smoothing parameter (set to 1 in our implementation).

4.2 Feature Probability Estimation

For each feature (historical count) f_i and class c, we estimate the probability as:

$$P(f_i|c) = \frac{count_i + \alpha}{total\_count + 6\alpha}$$

Where:

- $count_i$ is the historical count for the current class,
- $total\_count$ is the sum of historical counts for all classes except `true`.

4.3 Prediction

For a given sample, we calculate the score for each class c as:

$$score(c) = \begin{cases} \log P(c), & \text{if } c \text{ is } \texttt{true} \\ \log(1 - P(\texttt{true})) + \log P(f_c|c), & \text{otherwise} \end{cases}$$

The predicted class is the one with the highest score:

$$\text{predicted\_class} = \arg\max_c score(c)$$

5 Conclusion

This approach leverages the historical accuracy of speakers to make predictions about the truthfulness of their statements. By focusing on these historical statistics rather than the content of the statements, we aim to capture patterns in speaker reliability that may be indicative of statement truthfulness.

The effectiveness of this method suggests that a speaker's past record is a strong predictor of the truthfulness of their future statements, at least within the context of the LIAR dataset. However, it's important to note that this approach does not consider the content of the statements themselves, which could potentially provide additional valuable information for prediction.