

Sentiment Analysis using Bernoulli NB

Catherine MONOUE

Junior Kaningini

Paul Sanyang

Sakayo Toadoum Sari

Adviser: Moustapha Cisse, Ph.D.

African Institute for Mathematical Sciences



AIMS

African Institute for
Mathematical Sciences

NEXT EINSTEIN INITIATIVE

April 21, 2022

Presentation Outline

1 Introduction

2 Method

3 example

4 Results

Introduction

Definition 1.1

Naive Bayes is a family of algorithm based on bayes' theorem and where the main hypothesis is conditional independence. This simple means that the presence of a particular feature in a class does not help to know if another feature will be present.

Among the Naive Bayes algorithm, the most popular are:

- 1 Gaussian NB
- 2 multinomial
- 3 Bernoulli

Here, we will focus on Bernoulli Naive bayes

Bernoulli Naive Bayes Framework

Definition 2.1

The Bernoulli NB is used when the features are binary. The algorithm is divided in 2 steps

1 training

In this part the algorithm learns for the training data the frequency of a word given the label.

② test or prediction

Here, we use the frequency (parameters) computed in training stage to compute the likelihood for each label and the predicted label is the one associated to the high likelihood.

Bernoulli Naive Bayes Framework

Definition 2.2

Below, we give the formalisation

Suppose a training set : $(x_i, y_i)_{i \leq N}$ where x_i is a test and y_i the associated label. We want to be able to predict the correct label given a new text. To achieve our goal, we will use maximum likelihood estimation

Bernoulli Naive Bayes Framework part 1

Definition 2.3

Since the number of classes can be greater than 2, we suppose $y_i = (y_i^1, \dots, y_i^K)$ where K is the number of classes. Then if the label of the class is l , it means that the $y_i^l = 1$ and the others components are zero.

Let be $\theta_k = \mathbf{P}(y = k)$ with $\sum_k \mathbf{P}(y = k) = 1$.

The distribution of y is $\mathbf{P}(y) = \prod_k \theta_k^{y_k}$

Bernoulli Naive Bayes Framework part 2

Definition 2.4

write the expression of the likelihood

$$\mathbf{P}(y|x_1, \dots, x_N) = \frac{\mathbf{P}(x_1, \dots, x_N, y)}{\mathbf{P}(x_1, \dots, x_N)} \quad (1)$$

$$= \frac{\prod_{i=1}^N \mathbf{P}(x_i, y_i)}{\mathbf{P}(x_1, \dots, x_N)} \quad \text{text are independent} \quad (2)$$

$$= \frac{\prod_{i=1}^N \mathbf{P}(x_i|y_i) \times \mathbf{P}(y_i)}{\mathbf{P}(x_1, \dots, x_N)} \quad (3)$$

$$= \frac{\prod_{i=1}^N \prod_{j=1}^d \mathbf{P}(x_i^j|y_i) \times \prod_k \theta_k^{y_i^k}}{\mathbf{P}(x_1, \dots, x_N)} \quad (4)$$

Bernoulli Naive Bayes Framework part 3

Definition 2.5

write the expression of the likelihood

Since the denominator is constant, maximizing the likelihood is maximizing the quantity below

$$Q = \prod_{i=1}^N [(\prod_{j=1}^d \mathbf{P}(x_i^j | y_i)) \times \prod_k \theta_k^{y_i^k}]$$

Also, we have $\mathbf{P}(x_i^j | y_i) = \prod_k (\phi_{j,k}^{x_i^j} (1 - \phi_{j,k})^{1-x_i^j})^{y_i^k}$

where $\phi_{j,k}$ is the probability that the word j appear knowing that the label is k

Thus the log likelihood became:

$$\log(\mathcal{L}(\theta_k, \dots, \phi_{j,k})) = \sum_{i=1}^N \log(\prod_k \theta_k^{y_i^k}) + \sum_{i=1}^N \sum_{j=1}^d \log \left(\prod_k (\phi_{j,k}^{x_i^j} (1 - \phi_{j,k})^{1-x_i^j})^{y_i^k} \right) \quad (5)$$

Bernoulli Naive Bayes Framework part 4

Definition 2.6

$$\log(\mathcal{L}(\theta_k, \dots, \phi_{j,k})) = \sum_{k=1}^K \sum_{i=1}^N y_i^k \log(\theta_k) + \quad (7)$$

$$\sum_{i=1}^N \sum_{j=1}^d \sum_{k=1}^K y_i^k \left(x_i^j \log(\phi_{j,k}) + (1 - x_i^j) \log(1 - \phi_{j,k}) \right)$$

Bernoulli Naive Bayes Framework part 5

Definition 2.7

Hence

$$\log(\mathcal{L}) = \sum_{k=1}^K \sum_{i=1}^N y_i^k \log(\theta_k) + \sum_{j=1}^d \sum_{k=1}^K \left[\sum_{i=1}^N \left(y_i^k x_i^j \log(\phi_{j,k}) + y_i^k (1 - x_i^j) \right. \right. \\ \left. \left. \times \log(1 - \phi_{j,k}) \right) \right]$$

Bernoulli Naive Bayes Framework part 6

Definition 2.8

If we derive the $\log(\mathcal{L})$ with respect to $K \times d \times K$ parameters and set it to zero, we obtain:

$$\phi_{j,k} = \frac{\sum_{i=1}^N \mathbb{1}\{x_i^j = 1, y_i = k\}}{\sum_{i=1}^N \mathbb{1}\{y_i = k\}} \quad (8)$$

$$\theta_k = \frac{\sum_{i=1}^N \mathbb{1}\{y_i = k\}}{N} \quad (9)$$

Bernoulli Naive Bayes Framework part 7

Definition 2.9

use the parameters to predict for a new document. Let a be our new document, we compute for each label k $\mathbf{P}(y = k|a)$. The prediction will be the class that has the highest value for $\mathbf{P}(y = k|a)$

$$\mathbf{P}(y = k|a) = \frac{\mathbf{P}(a|y) \times \mathbf{P}(y = k)}{\mathbf{P}(a)} \quad (10)$$

$$\mathbf{P}(y = k|a) = \mathbf{P}(y = k) \times \prod_{j=1}^d \mathbf{P}(a_j|y = k) \quad (11)$$

$$\mathbf{P}(y = k|a) = \theta_k \times \prod_{j=1}^d (\phi_{j,k}^{a_j} (1 - \phi_{j,k})^{1-a_j}) \quad (12)$$

Laplace smoothing

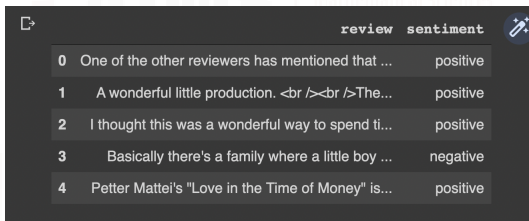
When the dictionary is big, some issues may arise we rare word present in there and not in the training set.

$$\phi_{j,k} = \frac{\alpha + \sum_{i=1}^N \mathbb{1}\{x_i^j = 1, y_i = k\}}{2\alpha + \sum_{i=1}^N \mathbb{1}\{y_i = k\}} \quad (13)$$

$$\theta_k = \frac{\sum_{i=1}^N \mathbb{1}\{y_i = k\}}{N} \quad (14)$$

Description of the Dataset

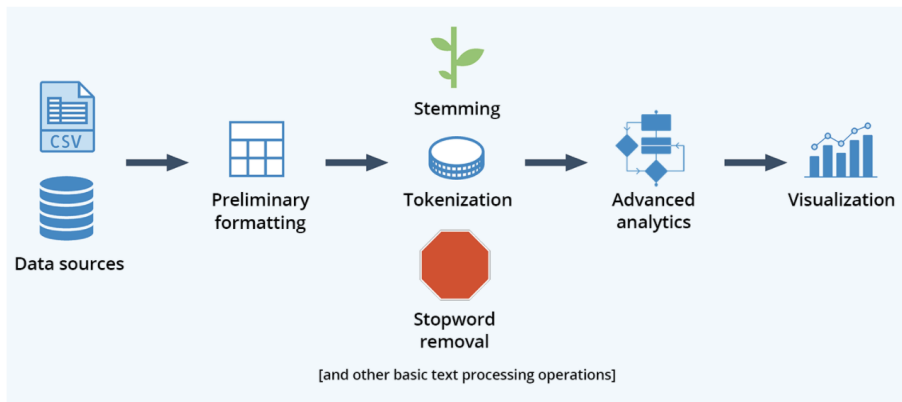
IMDB dataset have 50K movie reviews for natural language processing or Text analytics. This is a dataset for binary sentiment classification containing substantially more data than previous benchmark datasets. We provide a set of 25,000 highly polar movie reviews for training and 25,000 for testing.



	review	sentiment
0	One of the other reviewers has mentioned that ...	positive
1	A wonderful little production. The...	positive
2	I thought this was a wonderful way to spend ti...	positive
3	Basically there's a family where a little boy ...	negative
4	Petter Mattei's "Love in the Time of Money" is...	positive

Figure 1: head of dataset

Data Preprocessing



8/11

Example

D	REVIEW	C
1	The rooms were good and i like the location since it was good	+
2	The hotel was very bad and the stay was unpleasant	-
3	Liked the huge play area and the food was nice	+
4	The stay was good and pleasant	+
5	location was good but was bad overall because the staff were rude	-

Test set The rooms where good and the staff were nice ?

Term document matrix

features	D1 +	D2 -	D3 +	D4 +	D5 -
good	1	0	0	1	1
liked	1	0	1	0	0
bad	0	1	0	0	1
unpleasant	0	1	0	0	0
nice	0	0	1	0	0
pleasant	0	0	0	1	0
rude	0	0	0	0	1

Number of documents in "+" = 3 i.e D1, D3 and D4

$$\text{Prob}(\text{Document} = +) = P(Y = +) = \frac{3}{5}$$

Number of documents in "-" = 2 i.e D2 and D5

$$\text{Prob}(\text{Document} = -) = P(Y = -) = \frac{2}{5}$$

CLASSIFICATION MODEL

features	P(Feature +)	P(Feature -)
good	$\frac{2}{3}$	$\frac{1}{2}$
liked	$\frac{2}{3}$	$\frac{0}{2}$
bad	$\frac{0}{3}$	$\frac{2}{2}$
unpleasant	$\frac{0}{3}$	$\frac{1}{2}$
nice	$\frac{1}{3}$	$\frac{0}{2}$
pleasant	$\frac{1}{3}$	$\frac{0}{2}$
rude	$\frac{0}{3}$	$\frac{1}{2}$

Laplace Smoothing

Definition 3.1

$$P(\text{Feature} | \text{class} = c) = \frac{\text{Number of documents of class 'c' with feature } x_i + 1}{\text{Total number of documents of class 'c' + 2}}$$

Classify: "The rooms were good and the staff were nice"

Definition 3.2

$P(+| \text{good, liked, bad, unpleasant, nice, pleasant, rude})$

$\rightarrow P(\text{good, liked, bad, unpleasant, nice, pleasant, rude}|+) \times P(+)$

$$= \frac{3}{5} \times \left(1 - \frac{3}{5}\right) \times \left(1 - \frac{1}{5}\right) \times \left(1 - \frac{1}{5}\right) \times \frac{2}{5} \times \left(1 - \frac{2}{5}\right) \times \left(1 - \frac{2}{5}\right) \times \frac{3}{5} \quad (15)$$

$$= 0.013271 \quad (16)$$

Classify: "The rooms were good and the staff were nice"

Definition 3.3

$P(-| \text{good, liked, bad, unpleasant, nice, pleasant, rude})$

$\rightarrow P(\text{good, liked, bad, unpleasant, nice, pleasant, rude}|-) \times P(-)$

$$= \frac{2}{4} \times (1 - \frac{1}{4}) * (1 - \frac{3}{4}) \times (1 - \frac{2}{4}) \times \frac{1}{4} \times (1 - \frac{1}{4}) \times (1 - \frac{2}{4}) \times \frac{2}{5} \quad (17)$$

$$= 0.001758 \quad (18)$$

$$(19)$$

Since $0.013271 > 0.001758$ we conclude that the sentiment is positive

Advantages of Bernoulli Naive Bayes:

1. They are extremely fast as compared to other classification models
2. As in Bernoulli Naive Bayes each feature is treated independently with binary values only, it explicitly gives penalty to the model for non-occurrence of any of the features which are necessary for predicting the output y . And the other multinomial variant of Naive Bayes ignores this features instead of penalizing.
3. In case of small amount of data or small documents(for example in text classification), Bernoulli Naive Bayes gives more accurate and precise results as compared to other models.
4. It is fast and are able to make to make real-time predictions.

Disadvantages of Bernoulli Naive Bayes:

1. Being a naive(showing a lack of experience) classifier, it sometimes makes a strong assumption based on the shape of data
2. If at times the features are dependent on each other then Naive Bayes assumptions can affect the prediction and accuracy of the model and is sensitive to the given input data

Conclusion

Bernoulli Naive Bayes is one of the variants of the Naive Bayes algorithm in machine learning. It is very useful to be used when the dataset is in a binary distribution where the output label is either present or absent.

Results



Recommendations



List of References

- ① Andrew Ng, *CS229 Lecture Notes*
- ② Hyeong In Choi, *Lectures on Machine Learning (Fall 2017)*, Seoul National University, 2017

