

Binary Naive Bayes from scratch

African Master's in Machine Intelligence (AMMI)
Supervised by: Prof. Moustapha Cisse

April 22, 2022



- Introduction
- Methodology
- Results and discussions
- Conclusion
- References

What is the topic of this paper?

> [SSM Popul Health](#). 2021 Oct 5;16-100939. doi: 10.1016/j.ssmph.2021.100939.
eCollection 2021 Dec.

Spatial location, temperature and rainfall diversity affect the double burden of malnutrition among women in Kenya

Japheth Muema Kasomo ¹, Ezra Gayawan ²

Affiliations + expand

PMID: 34660880 PMCID: PMC8503666 DOI: 10.1016/j.ssmph.2021.100939

[Free PMC article](#)

Abstract

Studies have looked into how environmental and climate covariates affect under- and over-nutrition, but little is known about the spatial distribution of different forms of malnutrition in Kenya and whether there are locations that suffer from double-burden of malnutrition. This research quantifies spatial variations and estimates how climatic and environmental factors affect under- and over-nutrition among women in Kenya. This enables us to determine if the patterns in which these factors affect the malnutrition indicators are similar and whether there are overlaps in the spatial distributions. The study used data from the Demographic and Health Survey, which included cross-sectional data on malnutrition indicators as well as some climate and environmental variables. A multicategorical response variable that classified the women into one of four nutritional classes was generated from the body mass index (BMI) of the women, and a Bayesian geospatial regression model with an estimate based on the Markov chain Monte Carlo simulation technique was adopted. Findings show that women in Turkana, Samburu, Isiolo, Baringo, Garissa, and West Pokot counties are more likely to be underweight than women in other counties while being overweight is prevalent in Kirinyaga and Kitui counties. Obesity is prevalent in Kirinyaga, Lamu, Kiambu, Murang'a, and Taita Taveta counties. The study further shows that as mean temperature and precipitation increase, the likelihood of being underweight reduces. The chances of being underweight are lower among literate women [OR: 0.614; 95% CrI: 0.513,0.739], married women [OR: 0.702; 95% CrI: 0.608,0.819] and those from rich households [OR: 0.617; 95% CrI: 0.489,0.772], which is not the case for overweight and obesity. The generated spatial maps identify hot spots of the double burden of malnutrition that can assist the government and donor agencies in channeling resources efficiently.

Keywords: Climate; Food security; Geospatial regression model; Malnutrition; Spatial effects.

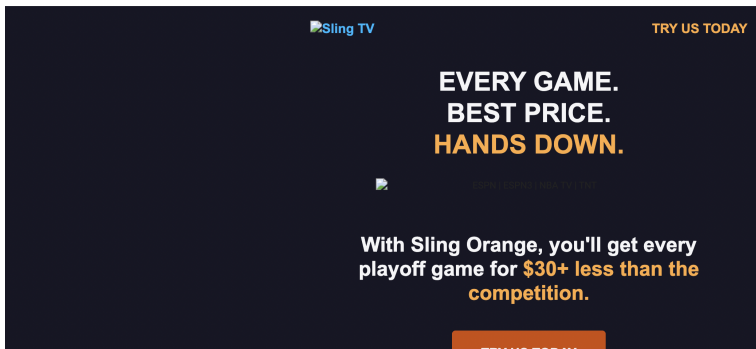
© 2021 The Authors. Published by Elsevier Ltd.

Figure: Source: *The National Center for Biotechnology Information*

Spam or ham?

Why is this message in spam? It is similar to messages that were identified as spam in the past.

Report not spam



Sling TV

TRY US TODAY

EVERY GAME.
BEST PRICE.
HANDS DOWN.

With Sling Orange, you'll get every
playoff game for **\$30+ less than the
competition.**

TRY US TODAY

How do rate this movie?

"I love this movie.
I've seen it many times
and it's still awesome."



"This movie is bad.
I don't like it it all.
It's terrible."



Figure: movie ratings

Source: CFML blog

Application of text classification

- Spam/ham detection
- Sentiment analysis
- Hate speech detection
- Automate CRM tasks
- Assigning subject topics
- ...

Machine learning (ML) text classification algorithms

Some of popular ML algorithms for text classification include:

- Multinormal Naive Bayes
- Binary Naive Bayes
- Support vector machines (SVM)
- Deep learning

The naives Bayes family of algorithms are some of most used algorithms in text classification and analysis. In this project we implement binary naive Bayes algorithm for sentiment analysis.

Sentiment analysis

- Sentiment analysis is a natural language processing (NLP) task where the objective is to predict a positive/negative orientation to a given task. The objective is to take an input, learn some useful features and classify the input into one of discrete classes.
- Sentiment analysis consists of finding opinion (negative, neutral or positive) from text documents such as movie reviews, as in our own case, or product reviews.
- Opinions about movies and products can be found in web blogs, social media, discussion forums e.t.c. Sentiment analysis can be a very important tool and companies can use it to improve customer experience, improve the services and products.

The Naive Bayes Classifier

The naive Bayes Classifier is a probabilistic machine learning algorithm for classification task. The algorithm is based on Bayes theorem. The method makes a naive assumption on how the features interact. Naive Bayes is a probabilistic classifier, meaning for a given document, it returns the class with the maximum posterior probability given the document.

Bayes theorem : The simple version of the Bayes Theorem can be derived from basic probability concepts. The simple form of the original Bayes Theorem can be demonstrated as follow:

$$P(A | B) = \frac{P(B | A)P(A)}{P(B)} \quad (1)$$

Besides, the alternative form of Bayes Theorem is generally encountered when looking at two competing statements or hypotheses:

$$P(A | B) = \frac{P(B | A)P(A)}{P(B | A)P(A) + P(B | A')P(A')}$$

For some partition A_i of the sample space, the extended form of Bayes Theorem is:

$$P(A | B) = \frac{P(B | A)P(A)}{\sum_i P(B | A_i)P(A_i)}$$

Naive Bayes Classifier in sentimental Analysis

Naive Bayes Mode is one of the two most-used classification models. The basic concept of Naive Bayes Classifier (NBC) is applying Bayes Theorem where the objects or attributes have independence. The detailed process is show below.

- There are two possible classes or categories, denoted by symbol C and C' , to classify each review into in this application: positive and negative.
- Vector denoted by $\vec{X} = \langle X_1, X_2, X_3, X_4, \dots X_n \rangle$ is used to represent a series of common attributes of negative reviews. In this application, the attributes of the reviews are simply individual words or phrases.
- Every review is represented by a vector denoted by $\vec{x} = \langle x_1, x_2, x_3, x_4, \dots x_n \rangle$, where each x_i represents the value of the attribute X_i .
- The string will be examined by NBC, comparing with the threshold value and resulting the final decision.

We can represent this mathematically;

$$P(C \mid \vec{X} = \vec{x}) = \operatorname{argmax}_{C \in \mathcal{C}} P(C \mid X)$$

$$P(C \mid \vec{X} = \vec{x}) = \frac{P(\vec{X} = \vec{x} \mid C) \cdot P(C)}{P(\vec{X} = \vec{x} \mid C) \cdot P(C) + P(\vec{X} = \vec{x} \mid C') \cdot P(C')}$$

$$= \frac{P(\vec{X} = \vec{x} \mid C) \cdot P(C)}{\prod_{i=1}^n P(X_i = x_i \mid C) \cdot P(C) + \prod_{i=1}^n P(X_i = x_i \mid C') \cdot P(C')}$$

where

$$P(\vec{X} = \vec{x} \mid C) = \prod_{i=1}^n P(X_i = x_i \mid C).$$

Naive Bayes assumptions

- Bag of words assumption- position of words in the document does not matter.
- The naive Bayes assumption-conditional independence assumption. Feature probabilities $P(X_i | C_j)$ are independent given class C

$$P(X_1, \dots, X_n | c) = P(X_1 | c) \cdot P(X_2 | C) \cdot P(X_3 | C) \cdot \dots \cdot P(X_n | C) \quad (2)$$

- The problem of floating point underflow-a problem when multiplying small probabilities.

Solution: use logs. We sum log of probabilities instead of multiplying probabilities. Thus we have:

$$C_{NB} = \operatorname{argmax}_{c \in C} \log P(C) + \sum_{i \in \text{positions}} \log P(X_i | C) \quad (3)$$

Training naive Bayes

How do we learn $P(C)$ and $P(X_i | C)$?

- Use frequencies in the data

$$\hat{P}(C) = \frac{N_C}{N_{doc}}, \quad (4)$$

where N_C is the number of documents in the training data with class C and N_{doc} is the total number of documents.

- $P(X_i | C)$ is the frequency of X_i in all documents of class C .

$$\hat{P}(X_i | C) = \frac{\text{count}(X_i, C)}{\sum_{X \in V} \text{count}(X, C)} \quad (5)$$

V , is the vocabulary and is union of all words in the classes.

After computing the probability, further procedure need to be taken into consideration before making the final decision of classifying the review. One method is to compare the ratio of the probability of this review being negative to the probability of this review being positive, and set a threshold value. The equation of this method can be shown as follow, where α the selected threshold value is:

$$\frac{P(C \mid \vec{X} = \vec{x})}{P(C' \mid \vec{X} = \vec{x})} > \alpha$$

After formulating, the above inequality becomes:

$$P(C \mid \vec{X} = \vec{x}) > t$$

where $t = \frac{\alpha}{1+\alpha}$.

Problem with maximum likelihood

Suppose we are trying to compute likelihood of a word that was not in the training documents. In this case, the probability will be zero. Since naive Bayes multiplies probabilities, it means zeros probabilities will cause the class to have probability of zero, no matter the prior we have. A simple solution would be **Laplace smoothing**.

The simplest form of Laplace smoothing is the add-one Laplace smoothing given by:

$$\hat{P}(X_i | C) = \frac{\text{count}(X_i, C) + 1}{\sum_{X \in V} (\text{count}(X, C) + 1)} = \frac{\text{count}(X_i, C) + 1}{(\sum_{X \in V} \text{count}(X, C)) + |V|} \quad (6)$$

Naive Bayes algorithm

The Figure below naive Bayes algorithm with laplace smoothing.

```

function TRAIN NAIVE BAYES(D, C) returns  $\log P(c)$  and  $\log P(w|c)$ 

for each class  $c \in C$            # Calculate  $P(c)$  terms
     $N_{doc}$  = number of documents in D
     $N_c$  = number of documents from D in class c
     $\text{logprior}[c] \leftarrow \log \frac{N_c}{N_{doc}}$ 
     $V \leftarrow$  vocabulary of D
     $\text{bigdoc}[c] \leftarrow$  append(d) for d  $\in D$  with class c
    for each word w in V           # Calculate  $P(w|c)$  terms
         $\text{count}(w, c) \leftarrow$  # of occurrences of w in  $\text{bigdoc}[c]$ 
         $\text{loglikelihood}[w, c] \leftarrow \log \frac{\text{count}(w, c) + 1}{\sum_{w' \in V} (\text{count}(w', c) + 1)}$ 
    return  $\text{logprior}, \text{loglikelihood}, V$ 

function TEST NAIVE BAYES( $\text{testdoc}, \text{logprior}, \text{loglikelihood}, C, V$ ) returns best c

for each class  $c \in C$ 
     $\text{sum}[c] \leftarrow \text{logprior}[c]$ 
    for each position i in  $\text{testdoc}$ 
         $\text{word} \leftarrow \text{testdoc}[i]$ 
        if  $\text{word} \in V$ 
             $\text{sum}[c] \leftarrow \text{sum}[c] + \text{loglikelihood}[\text{word}, c]$ 
    return  $\text{argmax}_c \text{sum}[c]$ 

```

Figure: naives Bayes algorithm

Source: Speech and Language Processing, An Introduction ...

Data preprocessing

Data preprocessing is the process of preparing the raw data and making it suitable for a machine or deep learning model and it is also the first and crucial step while creating a model. There are many preprocessing data techniques depending on the work one is working and the goal one wants to achieve. The following were the data preprocessing steps that were undertaken:

- lower case- the text is being converted to lower case letters, such that there won't be any difference in meaning between the capital letters and the lower letters.
- Remove stop words- stops words are common words that do not add much information to the text, e.g “the”, “a” etc. Removing stop words gives much focus information to the text, this improves training and reduces the size of the data as well.

- Word tokenisation- This is the process of splitting the text into smaller units called tokens.
- Stemming and Lemmatization- stemming and Lemmatization are methods to normalize a text document. Stemming usually refers to chopping off a few characters, it operates on a single word without knowledge of the context, it is not a well-defined process and often suffers from incorrect spelling and meaning. Whereas lemmatization is a method that switches any kind of word to its base root form. This is similar to stemming, however, it group different inflected form of words into a root base having the same meaning.
- One hot encoding- this can be defined as the essential process of converting the categorical data variables to be provided to machine and deep learning algorithms. One hot encoding is a common way of preprocessing categorical features for machine learning models. This type of encoding creates a new binary feature for each possible category and assigns a value of 1 to the feature of each sample that corresponds to its original category.

- TF-IDF- It is a numerical statistic that measures how important a word is in a document. Basically it helps us to associate each word in a document with a number that represents how relevant each word is in the document. Mathematically it can be written as;

$$TF - IDF = TF \times IDF$$

where;

$$TF = \frac{\text{number of representation of words in the document}}{\text{total number of words in the document}}$$

$$IDF = \log \left(\frac{\text{number of the documents in the corpus}}{\text{number of documents in the corpus contain the term}} \right)$$

Results and discussions

In this project we implemented one hot encoding, TF-IDF and count frequency. We split the data into test and train and train our model using 80% of the data . The overall accuracy on the test set was 70%.

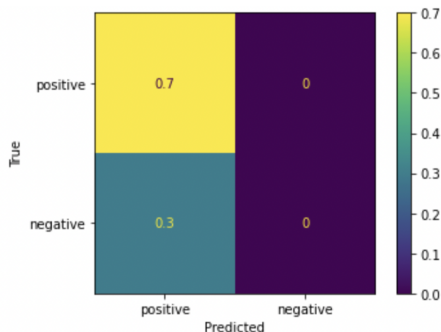


Figure: Confusion matrix

Conclusion

Naive Bayes classifier is one the machine learning algorithms for text classification. In this project, we implemented naive Bayes from scratch for sentiment analysis. The algorithm makes naive assumption on the features, which may not hold in real life. The algorithm is fast, simple and it does not require a large training data set. Other methods that can be used for sentiment analysis include multinomial naive Bayes, support vector machine and Gaussian mixture models.

Dieureudieuf ci dèglou bi !

References

- [1] Speech and Language Processing, An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Dan Jurafsky and James H. Martin, 2020.
- [2] Sentiment analysis on movie review data using machine learning approach, Rahman, Atiqur and Hossen, Md Sharif, 2019 International Conference on Bangla Speech and Language Processing (ICBSLP), 2019.
- [3] Understanding of the naive Bayes classifier in spam filtering, Wei, Qijia, AIP Conference Proceedings, 2018.