Binary Naive Bayes

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

April 21, 2022

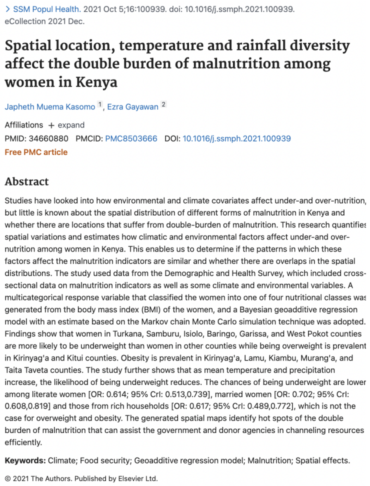
AIMS | African Institute for Mathematical Sciences SENEGAL

Outline

- Introduction
- Methods
- Implementation
- Results and evaluation
- Conclusion
- References

Motivation

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

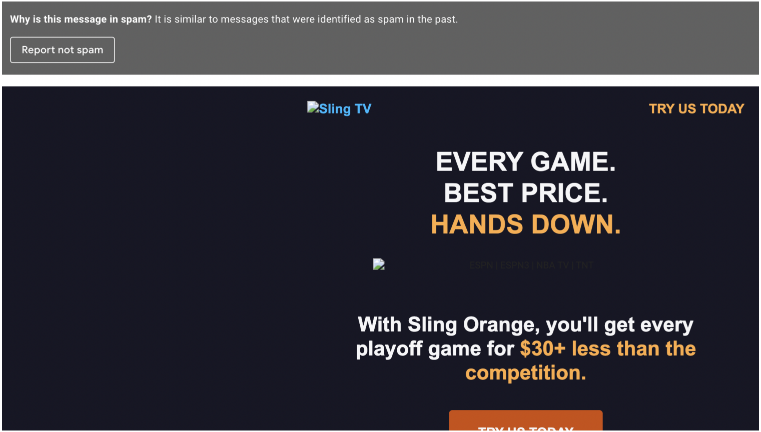
Abstract

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 geoadditive 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 Kirinyag'a and Kitui counties. Obesity is prevalent in Kirinyag'a, 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; Geoadditive regression model; Malnutrition; Spatial effects.

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

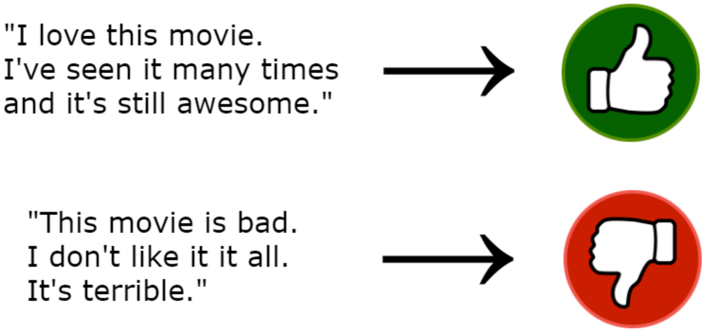
EVERY GAME.
BEST PRICE.
HANDS DOWN.

ESPN | ESPN3 | NBA TV | TNT

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 concentrate on sentiment analysis and we use binary Naive Bayes algorithm.

Introduction

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 clases.
- 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. The classifier is one of most used method for sentiment classification and some of its cons include:

- It is simple and easy to implement
- It requires much less training data
- it can be used for continuous and discrete data
- It is fast
- If the assumption of conditional independence holds, it can give good results

However, the classifier has its own disadvantages:

- The conditional independence assumptions does not always hold in real life
- The problem of zero probability. This might occur when we have some words of a certain class that were not in the training data.
- it can be a bad estimators, and the output probabilities should not be taken too seriously.

Methods

- For document d and class c, the Naive Bayes is probabilistic classifier, and returns the class $\hat{c}$ that has the highest posterior probability

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} P(c \mid d) \tag{1}$$

- Use Bayes rule

$$P(x \mid y) = \frac{P(y \mid x)P(x)}{P(y)} \tag{2}$$

- From Equation 1 and Equation 2

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} P(c \mid d) = \underset{c \in C}{\operatorname{argmax}} \frac{P(d \mid c)P(c)}{P(d)} \tag{3}$$

- Dropping the denominator we obtain

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} P(c \mid d) = \underset{c \in C}{\operatorname{argmax}} P(d \mid c) P(c) \tag{4}$$

  the prior probability $P(c)$ and the likelihood $P(d|c)$.

- $$\hat{c} = \underset{c \in C}{\operatorname{argmax}} \overbrace{P(d \mid c)}^{\text{likelihood}} \overbrace{P(c)}^{\text{prior}} \tag{5}$$

- Documents represented as features

$$\underset{c \in C}{\operatorname{argmax}} P\left(x_1, x_2, \ldots, x_n \mid c\right) P(c) \tag{6}$$

Naive Bayes assumptions

- Bag of words assumption- position of words in the document does not matter.
- The naives Bayes assumption-conditional independence assumption. Feature probabilities $P(x_i \mid c_j)$ are independent given class c.

$$P(x_1, \ldots, x_n \mid c) = P(x_1 \mid c) \cdot P(x_2 \mid c) \cdot P(x_3 \mid c) \cdot \ldots \cdot P(x_n \mid c) \quad (7)$$

- The final naive Bayes class can be given as

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c) \prod_{x \in X} P(x \mid c) \quad (8)$$

- When we use naive Bayes for text analysis, we need to consider the position of the words in the document.

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c) \prod_{i \in \text{ positions}} P\left(w_i \mid c\right) \tag{9}$$

- 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:

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}} \log P(c) + \sum_{i \in \text{ positions}} \log P\left(w_i \mid c\right) \tag{10}$$

Note: Log is strictly increasing functions and taking logs does not change the output classes. Log is used for convenience.

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.