A Simple approach to Naive Bayes Classifier (NBC)

By Shoaib Rain
Date: 11/03/2022

Introduction

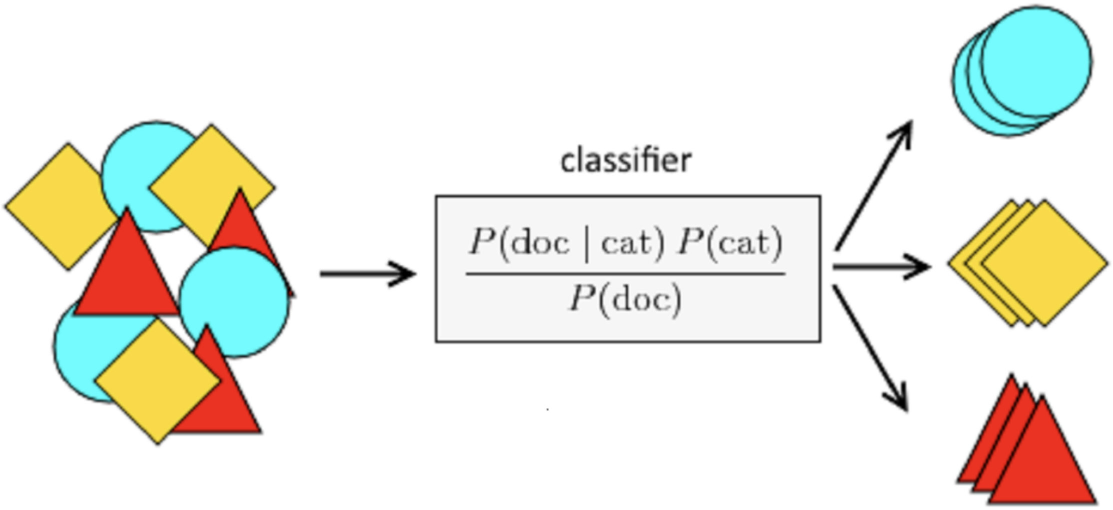

classifier
$\frac{P(\text{doc} \mid \text{cat})\, P(\text{cat})}{P(\text{doc})}$

Image taken from source

Naive Bayes Classifiers are simple yet powerful Machine Learning algorithms that are based on conditional probability and Bayes's theorem. . There're many use cases and requirements, where the computer algorithm has to label images to different categories/ classes. Identifying images and objects that we see on a regular basis is a pretty easy task for us, but it has been difficult to identify images and objects and therefore an image classification has been an important task within the field of computer vision. Image classification basically refers to labeling an image to one of the number of predefined classes. And Naive Bayes Classifiers (NBC) allows us to make that complex boolean decision to map an image instance to its most probabilistic category class.

Math Background

Conditional Probability

As stated in the Paper [source], a conditional probability of an event B is the probability that the event will occur given the knowledge that an event A has already occurred. This probability is written P(B|A), notation for the probability of B given A. In the case where events A and B are independent (where event A has no effect on the probability of event B), the conditional probability of event B given event A is simply the probability of event B, that is P(B).

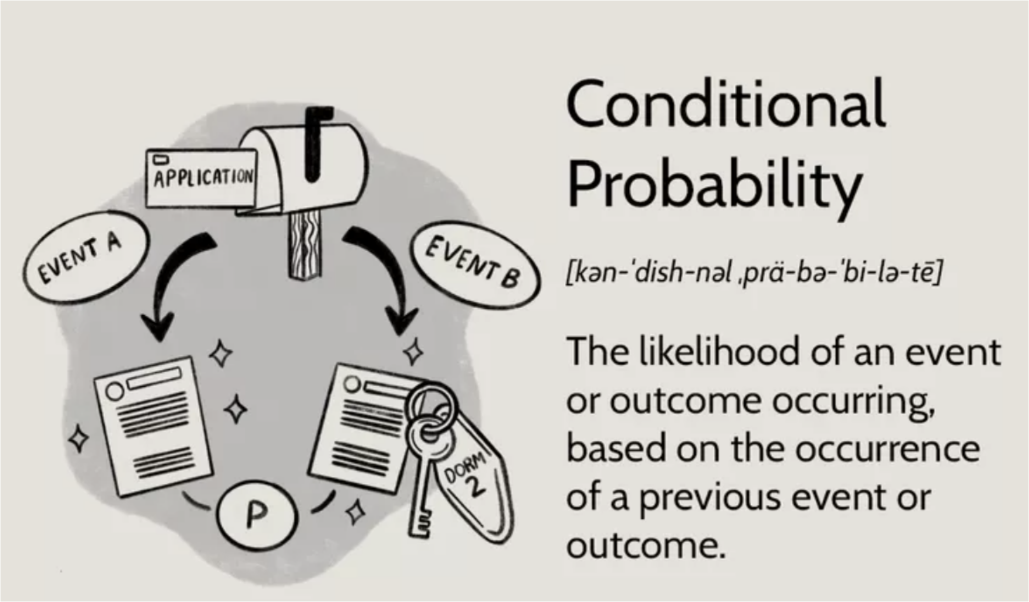
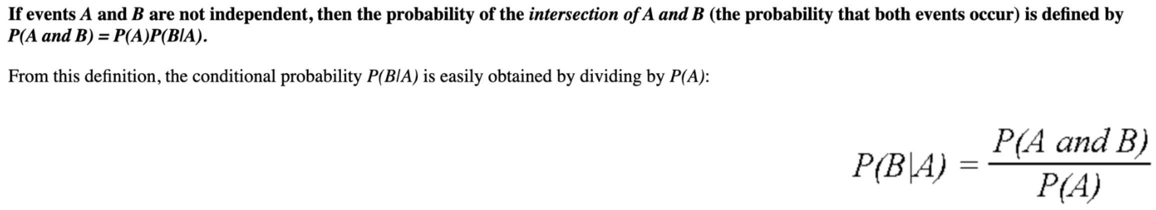

Conditional Probability
[kən-ˈdish-nəl ˌprä-bə-ˈbi-lə-tē]
The likelihood of an event or outcome occurring, based on the occurrence of a previous event or outcome.
APPLICATION
EVENT A
EVENT B
P
DORM 2

Image taken from source



If events *A* and *B* are not independent, then the probability of the *intersection of A and B* (the probability that both events occur) is defined by *P(A and B) = P(A)P(B|A)*.

From this definition, the conditional probability *P(B|A)* is easily obtained by dividing by *P(A)*:

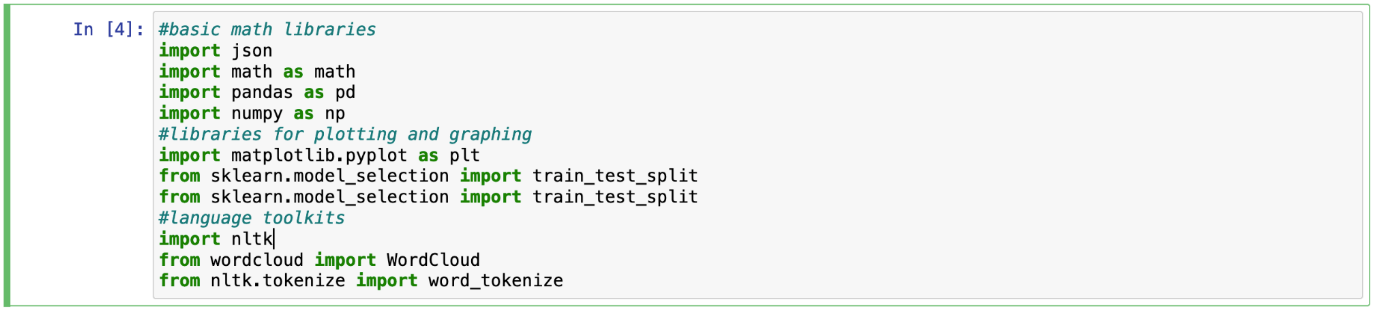
$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$$

conditional probs and Bayes's Theorem

Bayes's Theorem is a mathematical notation for determining the conditional probability of any event’s occurrence given some condition/s. This theorem provides an elegant way for updating the probability of any heuristic agent as condition changes over time/iteration. This theorem is widely used in machine learning applications like risk of lending money to a borrower.

Python NoteBook

Getting started with imports



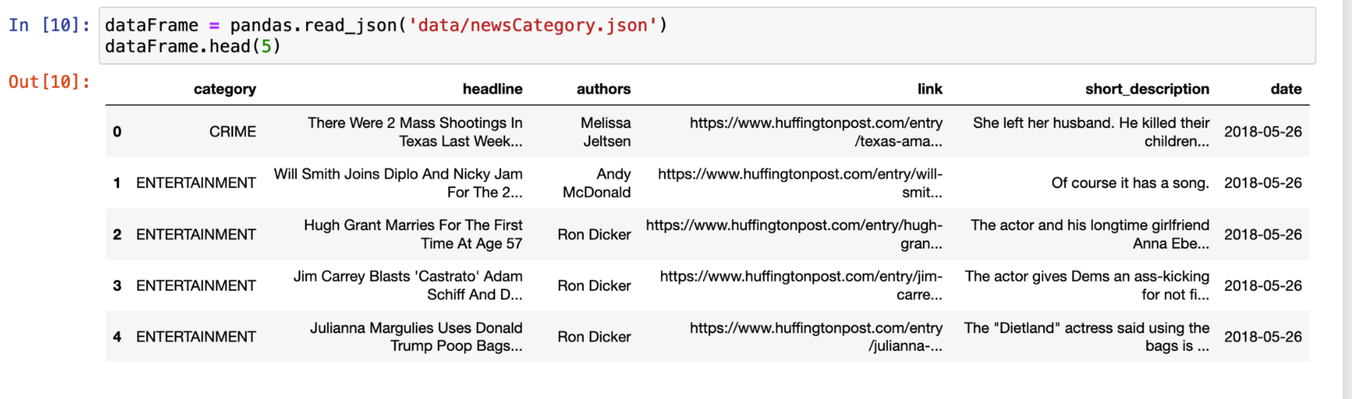
In [4]:

```python
#basic math libraries
import json
import math as math
import pandas as pd
import numpy as np
#libraries for plotting and graphing
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.model_selection import train_test_split
#language toolkits
import nltk
from wordcloud import WordCloud
from nltk.tokenize import word_tokenize
```

Full code here: [github]
Here we've a preprocessed data in csv format that we will upload in the memor

Exploring dataSet

We will perform a classification of news headlines into different predefined catefgories using NBC. This News Category Dataset [data set source] which has more than 200 thousands news headlines from different sources across different time starting from 2012 to 2018 and it was archived from HuffPost. In this datas set, each news headlines is mapped with its corresponding headline category which is labbeled manually.
The data set has six identifying features for each instance of news records. As we can see the dataSet is pretty well modulated given each columns serves a meaning correlations with each other and makes the obvious use of Naive Bayes Classifier (NBC) based on conditional probability theory.



In [10]:

```python
dataFrame = pandas.read_json('data/newsCategory.json')
dataFrame.head(5)
```

Out[10]:

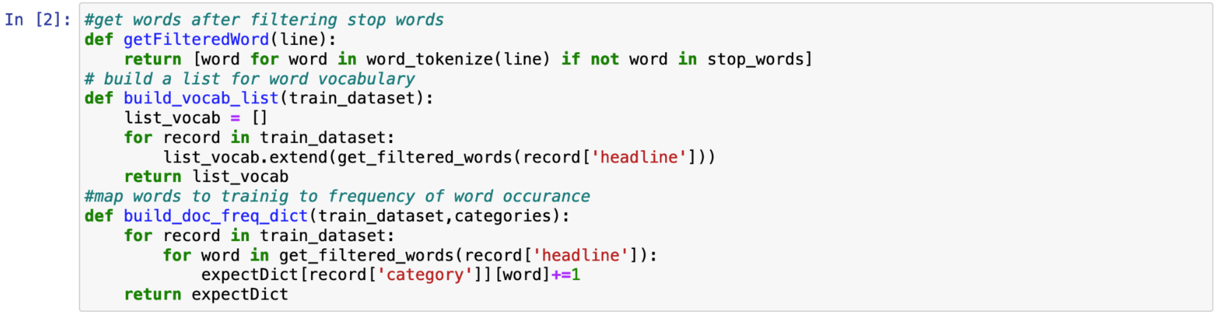
	category	headline	authors	link	short_description	date
0	CRIME	There Were 2 Mass Shootings In Texas Last Week...	Melissa Jeltsen	https://www.huffingtonpost.com/entry/texas-ama...	She left her husband. He killed their children...	2018-05-26
1	ENTERTAINMENT	Will Smith Joins Diplo And Nicky Jam For The 2...	Andy McDonald	https://www.huffingtonpost.com/entry/will-smit...	Of course it has a song.	2018-05-26
2	ENTERTAINMENT	Hugh Grant Marries For The First Time At Age 57	Ron Dicker	https://www.huffingtonpost.com/entry/hugh-gran...	The actor and his longtime girlfriend Anna Ebe...	2018-05-26
3	ENTERTAINMENT	Jim Carrey Blasts 'Castrato' Adam Schiff And D...	Ron Dicker	https://www.huffingtonpost.com/entry/jim-carre...	The actor gives Dems an ass-kicking for not fi...	2018-05-26
4	ENTERTAINMENT	Julianna Margulies Uses Donald Trump Poop Bags...	Ron Dicker	https://www.huffingtonpost.com/entry/julianna-...	The "Dietland" actress said using the bags is ...	2018-05-26

Full code here: [github]

Building the Naive Bayes Classifiers

There are many implementations out there for Naive Bayes Classifier. Some of them are more or less efficient than others. But they all come with their own tradeoffs. In this exercise, We will be implementing a Naive Bayes Classifier from scratch. There're many libraries and packages available which is optimized for efficiency and accuracy, but in this exercise, we will be implementing it from scratch using basics of conditional probability and Bayes's Theroem.

We will define some atomic functions that will help us to accomplish the goal of our custom Naive Bayes Classifier. Here, we will get the filtered words from a bag of words. And using that result, we will build vocabulary list out of all the words that was seen in the training data sets. Also, we have to do smoothing for words.



```
In [2]: #get words after filtering stop words
        def getFilteredWord(line):
            return [word for word in word_tokenize(line) if not word in stop_words]
        # build a list for word vocabulary
        def build_vocab_list(train_dataset):
            list_vocab = []
            for record in train_dataset:
                list_vocab.extend(get_filtered_words(record['headline']))
            return list_vocab
        #map words to trainig to frequency of word occurance
        def build_doc_freq_dict(train_dataset,categories):
            for record in train_dataset:
                for word in get_filtered_words(record['headline']):
                    expectDict[record['category']][word]+=1
            return expectDict
```

Full code on **github**

Self contribution

In this exercise, I've contributed a lot regarding data analysis and finding correlations between dataset features. I used plotting tools to visualize and compare different zones of the dataset and implemented atomic functions that make up the whole Naive Bayes Classifiers executable. I've also used various techniques for dealing with overfitting and underfitting. I trained and tested the model multiple times with slight changes in params and came to the conclusion that it's the best I can do given time. Along the way, I also learned a lot of new concepts and skills like Laplace smoothing and I successfully managed to implement it in NBC. I've made great use of basic libraries and dataFrame packages and improved the accuracy of the model by cleaning the data properly, and making an optimal split between training and testing data. I've also faced many challenges while completing this task . Handling the dataset with such complexity and without use of any external library was difficult. This increases a bug in code and I would still consider choosing a standard library instead of creating my own. Probability calculation based on informed search and making a conditional probability prediction was quite challenging.

Conclusion

As you can see, our model makes predictions among three different object classes very well.

Using this simple technique, we can build an efficient news category classifier for any custom dataset. In this blog, we have successfully learned how to implement Naive Bayes Classifiers from scratch.
You can find the entire code from **Github**

References

Conditional probability. (n.d.). Retrieved November 14, 2022, from http://www.stat.yale.edu/Courses/1997-98/101/condprob.htm

Chaudhuri, K. D. (2022, March 25). *Building naive Bayes classifier from scratch to perform sentiment analysis*. Analytics Vidhya. Retrieved November 14, 2022, from https://www.analyticsvidhya.com/blog/2022/03/building-naive-bayes-classifier-from-scratch-to-perform-sentiment-analysis/

Woller, M. (2022, September 9). *What's All the buzz about Bayes?* Psychology In Action. Retrieved November 14, 2022, from https://www.psychologyinaction.org/psychology-in-action-1/2021/10/19/whats-all-the-buzz-about-bayes

Jeremy Jordan. (2018, August 25). *Evaluating a machine learning model.* Jeremy Jordan. Retrieved November 14, 2022, from https://www.jeremyjordan.me/evaluating-a-machine-learning-model/

Matplotlib 3.5.0 documentation. (n.d.). Retrieved November 14, 2022, from https://matplotlib.org/3.5.0/gallery/misc/table_demo.html#sphx-glr-gallery-misc-table-demo-py