Assignment 2 : Naive Bayes Classifier

The Naive Bayes classification algorithm is the derived from the conditional probability formula

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

In this classification, we calculate the probability of the sentence belongs to a class which is given by below formula

$$p(\text{Spam}|w_1, \ldots, w_n) = \frac{p(w_1, \ldots, w_n|\text{Spam})p(\text{Spam})}{p(w_1, \ldots, w_n)}$$

Where w1, w2 … wn represents the words in the sentence

To help us with that equation, we can make an assumption called the Naive Bayes assumption to help us with the math, and eventually the code. The assumption is that each word is independent of all other words. In reality, this is not true!

$$p(\text{Spam}|w_1, \ldots, w_n) = \frac{p(w_1|\text{Spam})p(w_2|\text{Spam}) \ldots p(w_n|\text{Spam})p(\text{Spam})}{p(w_1, \ldots, w_n)}$$

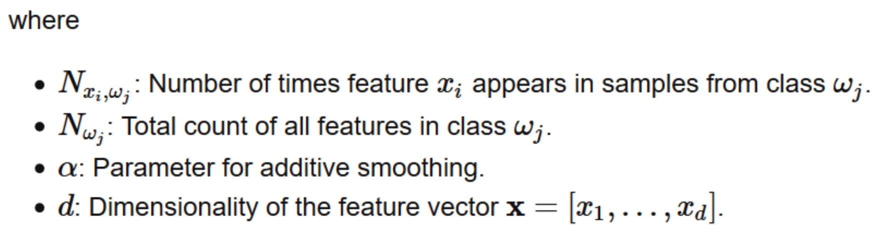
As the denominator in the right side is common for all the classes we neglect it So the final equation becomes

$$p(\text{Spam}|w_1, \ldots, w_n) \propto p(\text{Spam}) \prod_{i=1}^{n} p(w_i|\text{Spam})$$

Laplace Smoothing(apha = 1):

There might be cases where the probability of a word in the dataset is zero. So we apply additive smoothing called Laplace smoothing, in order to avoid zero probabilities. The formula is given by:

$$\hat{P}(x_i \mid \omega_j) = \frac{N_{x_i,\omega_j} + \alpha}{N_{\omega_j} + \alpha\, d} \quad (i = (1, \ldots, d))$$


where

- N_{x_i,ω_j}: Number of times feature x_i appears in samples from class ω_j.
- N_{ω_j}: Total count of all features in class ω_j.
- α: Parameter for additive smoothing.
- d: Dimensionality of the feature vector $\mathbf{x} = [x_1, \ldots, x_d]$.

The steps taken can be listed out as follows:

- Importing the python libraries like numpy, pandas, maths
- Importing the data from the file
- Splitting the data into 7 equal groups(142 rows) to perform the 7 fold validation on the data set
- Calculating the word frequency of both classes positive and negative
- Applying the naive bayes classification formula along with the laplace smoothing
- Now we do 7 fold cross validation by selecting a few chunks that were made and checking the accuracy

The accuracy of the k-folds as follows for the 5 folds

1st fold	65.49295774647888
2nd fold	64.08450704225352
3rd fold	66.90140845070422,
4th fold	60.56338028169014
5th fold	59.859154929577464
6th fold	61.97183098591549
7th fold	64.08450704225352

The overall accuracy of the model is ***63.27967806841047***

Limitations of Naive Bayes Classifier:

1. The main limitation of the classifier is the assumption of occurrence of the words in a sentence is independent of each other. In real life, it was almost impossible for the words to be independent as we know that language as well defined structure
2. The zero probability of the feature is another limitation, which can be solved by using additive smoothing like laplace smoothing