

ML assignment 2

Group:

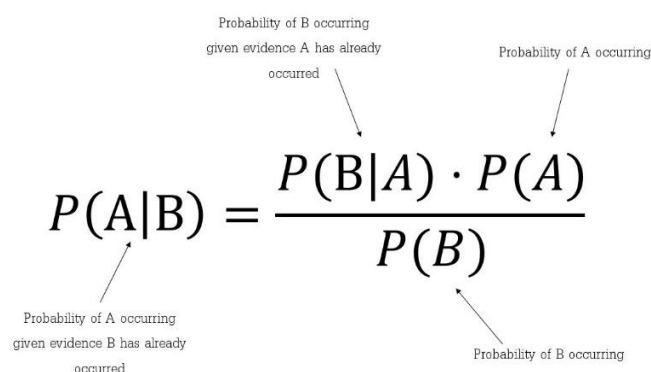
1. Achyuta Krishna V 2018A7PS0165H
2. J Alvin Ronnie 2018A7PS0029H
3. Shriram R 2018B3A70948H

Naïve Bayes Classifier:

Naïve Bayes Classifiers are classification algorithms based on the Bayes theorem where every pair of features are independent of each other. It revolves around the assumption that every feature contributes to the outcome independently and equally.

Bayes Theorem:

The mathematical form of the theorem is as follows:



The diagram shows the Bayes Theorem formula $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$ with four arrows pointing to its components and their meanings:

- An arrow from $P(A|B)$ points to the text: "Probability of A occurring given evidence B has already occurred".
- An arrow from $P(B|A)$ points to the text: "Probability of B occurring given evidence A has already occurred".
- An arrow from $P(A)$ points to the text: "Probability of A occurring".
- An arrow from $P(B)$ points to the text: "Probability of B occurring".

This then can be applied to our dataset as:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

where y is the class variable and X is a dependent feature where

$$X = (x_1, x_2, x_3, \dots, x_n)$$

The Naïve Bayes Classifier implemented classifies based on the following results :

$$P(\text{Spam}|w_1, w_2, \dots, w_n) \propto P(\text{Spam}) \cdot \prod_{i=1}^n P(w_i|\text{Spam}) \quad - (i)$$

$$P(\text{Ham}|w_1, w_2, \dots, w_n) \propto P(\text{Ham}) \cdot \prod_{i=1}^n P(w_i|\text{Ham}) \quad -(ii)$$

where w_i is the i th word

If the result in (i) is greater than that of (ii), then the statement is classified as spam

Accuracy of the model over each fold:

```
Accuracy: 84.61538461538461
Accuracy: 81.81818181818181
Accuracy: 82.51748251748252
Accuracy: 84.61538461538461
Accuracy: 83.91608391608392
Accuracy: 75.52447552447552
Accuracy: 83.80281690140845
```

Average accuracy:

```
Average accuracy: 82.40140141548594
```

Limitation of Naïve Bayes Classifier:

- Assumes that every feature is independent of each other but in real life, it is almost impossible to find a set of independent predictors
- When there is Data Scarcity, classifier fails to give accurate results