Naive Bayes Algorithm

# 1. Introduction

Naive Bayes is a probabilistic machine learning algorithm that can be used in a wide variety of classification tasks. Typical applications include filtering spam, classifying documents, sentiment prediction, etc.

# 2. Understanding Conditional Probability

## a) Coin Toss and Fair Dice Example

When you flip a fair coin, there is an equal chance of getting either heads or tails. So you can say the probability of getting heads is 50%.

Similarly what would be the probability of getting a 1 when you roll dice with 6 faces? Assuming the dice is fair, the probability of 1/6 = 0.166.

## b) Playing Cards Example

If you pick a card from the deck, can you guess the probability of getting a queen given the card is a spade?

Well, I have already set a condition that the card is a spade. So, the denominator (eligible population) is 13 and not 52. And since there is only one queen in spades, the probability it is a queen given the card is a spade is 1/13 = 0.077

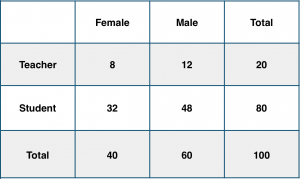
This is a classic example of conditional probability. So, when you say the conditional probability of A given B, it denotes the probability of A occurring given that B has already occurred.

Mathematically, the Conditional probability of A given B can be computed as P(A|B) = P(A AND B) / P(B)

## c) School Example

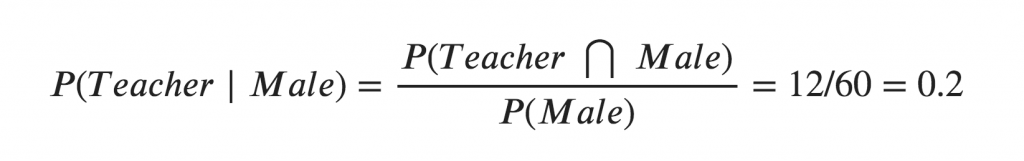
Let’s see a slightly complicated example. Consider a school with a total population of 100 persons. These 100 persons can be seen either as ‘Students’ and ‘Teachers’ or as a population of ‘Males’ and ‘Females’.

With the below tabulation of the 100 people, what is the conditional probability that a certain member of the school is a ‘Teacher’ given that he is a ‘Man’?

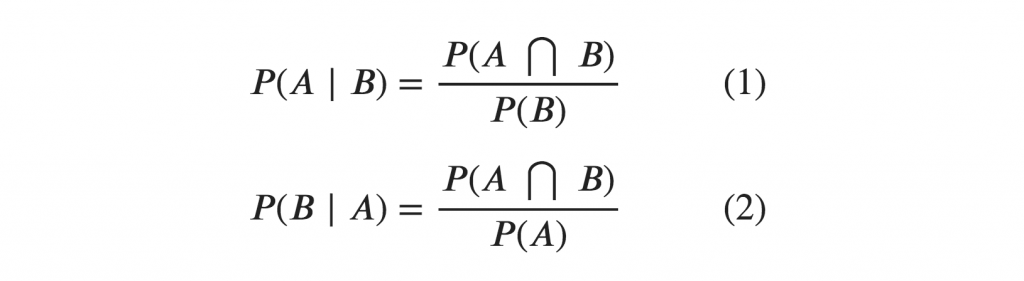


To calculate this, you may intuitively filter the sub-population of 60 males and focus on the 12 (male) teachers.

So the required conditional probability P(Teacher | Male) = 12 / 60 = 0.2.



This can be represented as the intersection of Teacher (A) and Male (B) divided by Male (B). Likewise, the conditional probability of B given A can be computed. The Bayes rule that we use for Naive Bayes, can be derived from these two notations.



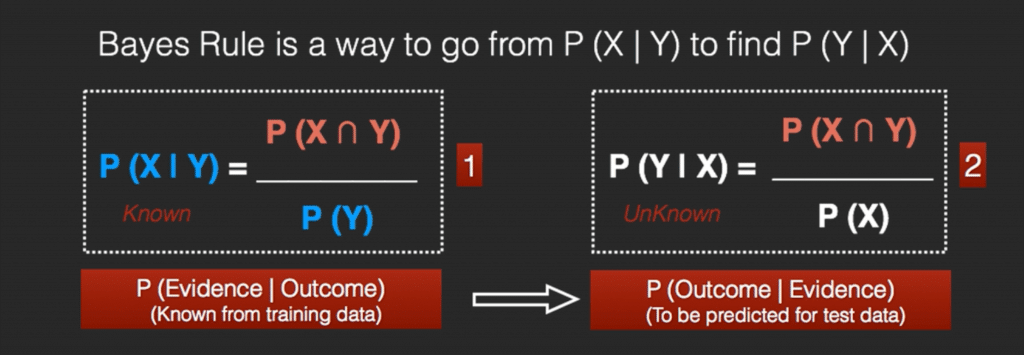
# 3. The Bayes Rule

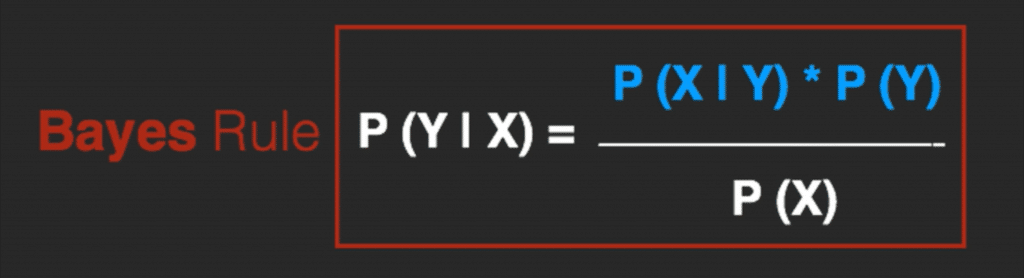
The Bayes Rule is a way of going from P(X|Y), known from the training dataset, to find P(Y|X).

To do this, we replace A and B in the above formula, with the feature X and response Y.

For observations in test or scoring data, the X would be known while Y is unknown. And for each row of the test dataset, you want to compute the probability of Y given the X has already happened.

What happens if Y has more than 2 categories? we compute the probability of each class of Y and let the highest win.



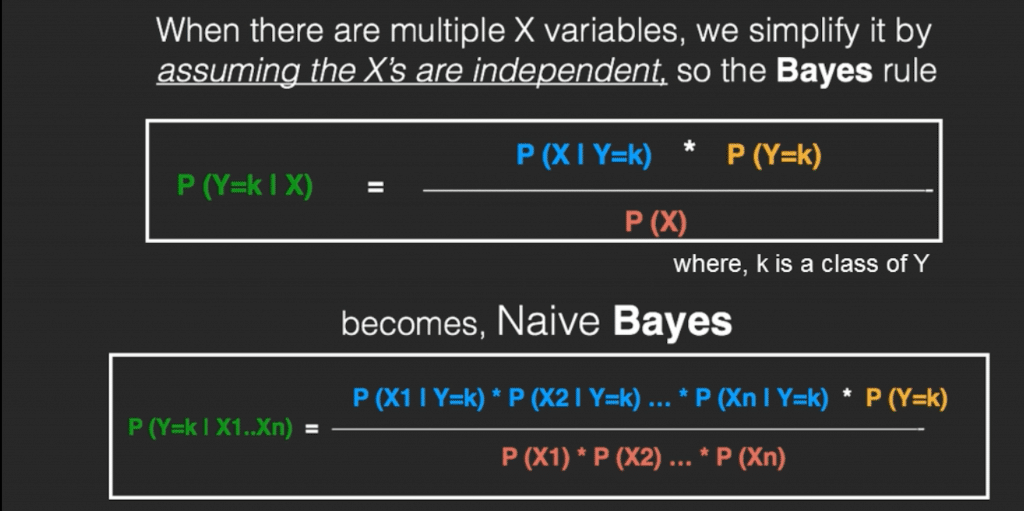


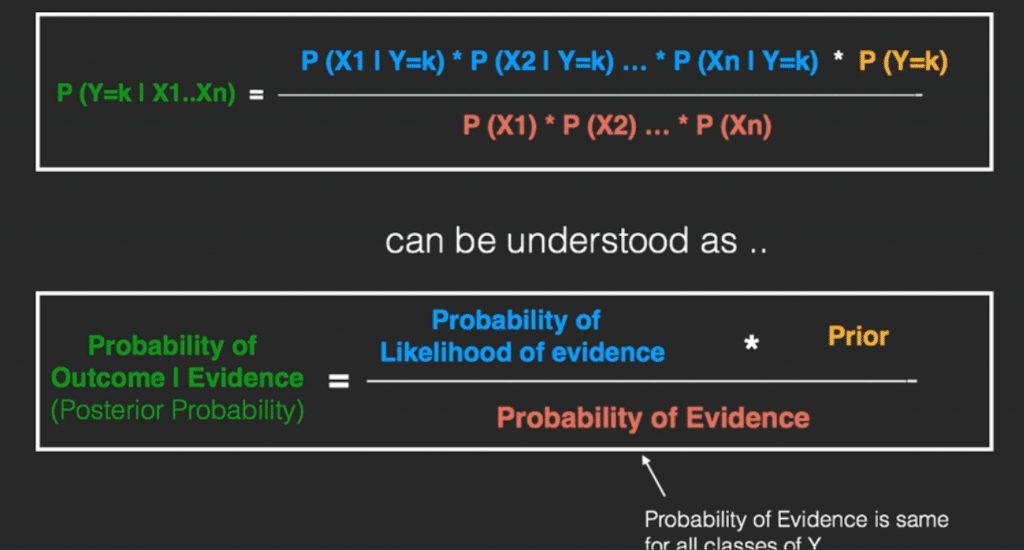
# 4. The Naive Bayes

The Bayes Rule provides the formula for the probability of Y given X. But, in real-world problems, you typically have multiple X variables.

When the features are independent, we can extend the Bayes Rule to what is called Naive Bayes.

It is called ‘Naive’ because of the naive assumption that the X’s are independent of each other. Regardless of its name, it’s a powerful formula.





In technical jargon, the left-hand-side (LHS) of the equation is understood as the posterior probability or simply the posterior

The RHS has 2 terms in the numerator.

The first term is called the **‘Likelihood of Evidence’**. It is nothing but the conditional probability of each X’s given Y is of particular class ‘c’.

Since all the X’s are assumed to be independent of each other, you can just multiply the ‘likelihoods’ of all the X’s and called it the ‘Probability of likelihood of evidence’. This is known from the training dataset by filtering records where Y=c.

The second term is called the prior which is the overall probability of Y=c, where c is a class of Y. In simpler terms, Prior = count(Y=c) / n\_Records.

# 5. Naive Bayes Example by Hand

Say you have 1000 fruits which could be either ‘banana’, ‘orange’, or ‘other’. These are the 3 possible classes of the Y variable.

We have data for the following X variables, all of which are binary (1 or 0).

* Long
* Sweet
* Yellow

The first few rows of the training dataset look like this:

| **Fruit** | **Long (x1)** | **Sweet (x2)** | **Yellow (x3)** |
| --- | --- | --- | --- |
| Orange | 0 | 1 | 0 |
| Banana | 1 | 0 | 1 |
| Banana | 1 | 1 | 1 |
| Other | 1 | 1 | 0 |
| .. | .. | .. | .. |

For the sake of computing the probabilities, let’s aggregate the training data to form a **counts table** like this.



So the objective of the classifier is to predict if a given fruit is a ‘Banana’ or ‘Orange’ or ‘Other’ when only the 3 features (long, sweet, and yellow) are known.

Let’s say you are given a fruit that is: Long, Sweet, and Yellow, can you predict what fruit it is?

This is the same as predicting the Y when only the X variables in testing data are known. Let’s solve it by hand using Naive Bayes.

The idea is to compute the 3 probabilities, that is the probability of the fruit being a banana, orange or other. Whichever fruit type gets the highest probability wins.

All the information to calculate these probabilities is present in the above tabulation.

### Step 1: Compute the ‘Prior’ probabilities for each of the class of fruits.

That is the proportion of each fruit class out of all the fruits from the population. You can provide the ‘Priors’ from prior information about the population. Otherwise, it can be computed from the training data.

For this case, let’s compute from the training data. Out of 1000 records in training data, you have 500 Bananas, 300 Oranges, and 200 Others. So the respective priors are 0.5, 0.3, and 0.2.

P(Y=Banana) = 500 / 1000 = 0.50

P(Y=Orange) = 300 / 1000 = 0.30

P(Y=Other) = 200 / 1000 = 0.20

### Step 2: Compute the probability of evidence that goes in the denominator.

This is nothing but the product of P of Xs for all X. This is an optional step because the denominator is the same for all the classes and so will not affect the probabilities.

P(x1=Long) = 500 / 1000 = 0.50

P(x2=Sweet) = 650 / 1000 = 0.65

P(x3=Yellow) = 800 / 1000 = 0.80

### Step 3: Compute the probability of the likelihood of evidence that goes in the numerator.

It is the product of conditional probabilities of the 3 features. If you refer back to the formula, it says P(X1 |Y=k). Here X1 is ‘Long’ and k is ‘Banana’. That means the probability the fruit is ‘Long’ given that it is a Banana. In the above table, you have 500 Bananas. Out of that 400 is long. So, P(Long | Banana) = 400/500 = 0.8.

Here, I have done it for Banana alone.

**Probability of Likelihood for Banana**

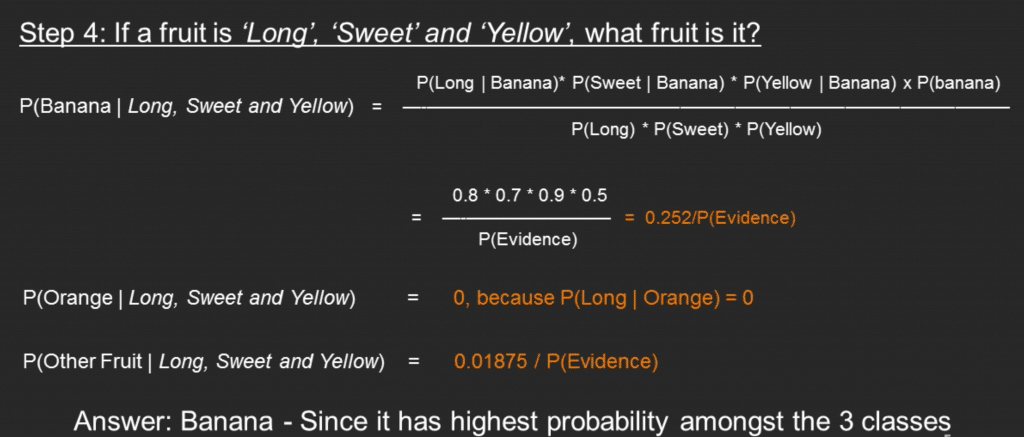
P(x1=Long | Y=Banana) = 400 / 500 = 0.80

P(x2=Sweet | Y=Banana) = 350 / 500 = 0.70

P(x3=Yellow | Y=Banana) = 450 / 500 = 0.90

So, the overall probability of Likelihood of evidence for Banana = 0.8 \* 0.7 \* 0.9 = 0.504

### Step 4: Substitute all the 3 equations into the Naive Bayes formula, to get the probability that it is a banana.

****

Similarly, you can compute the probabilities for ‘Orange’ and ‘Other fruit’. The denominator is the same for all 3 cases, so it’s optional to compute.

Clearly, Banana gets the highest probability, so that will be our predicted class.

# 6. Build a Naive Bayes classifier:

Please refer to the file that is named Naive\_Bayes.ipynb.