This is an example of Naïve Bayes.. worked out by hand. In this case.. the Label is “fruit” – and it has three “classes” – orange, banana, and other.

(note: In our assignment, the label we are predicting is “Future entrepreneur” .. which is yes/ no – so it’s a two class problem, rather than three.)

There are three features that are being used to predict what the fruit is – whether the fruit is Long, whether it is Sweet, and whether it is Yellow – each of these are just binary 1/0 features.

(In our assignment, we have features such as gender (2 values), whether the parents had own business (yes, no) etc)

WORKED EXAMPLE, SHOWNIG CALCULATIONS.

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

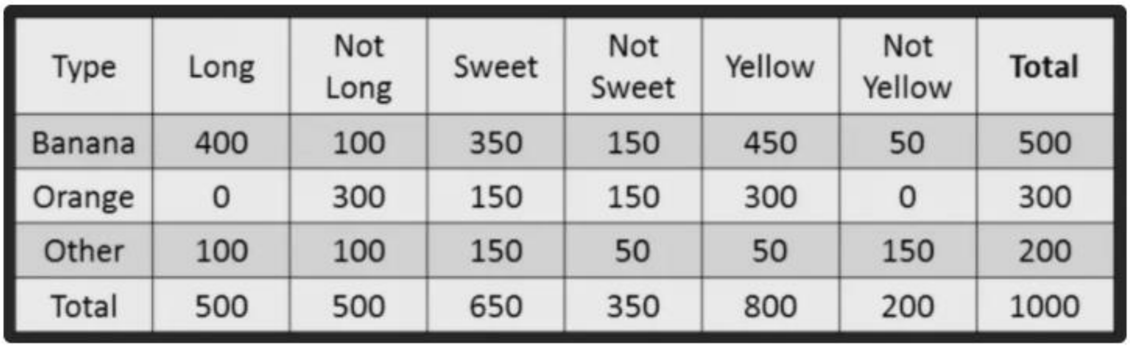
We have data for the following X variables, all of which are binary (1 or 0).

- 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.



Type	Long	Not Long	Sweet	Not Sweet	Yellow	Not Yellow	**Total**
Banana	400	100	350	150	450	50	500
Orange	0	300	150	150	300	0	300
Other	100	100	150	50	50	150	200
Total	500	500	650	350	800	200	1000

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 of 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 likelihood of evidences 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

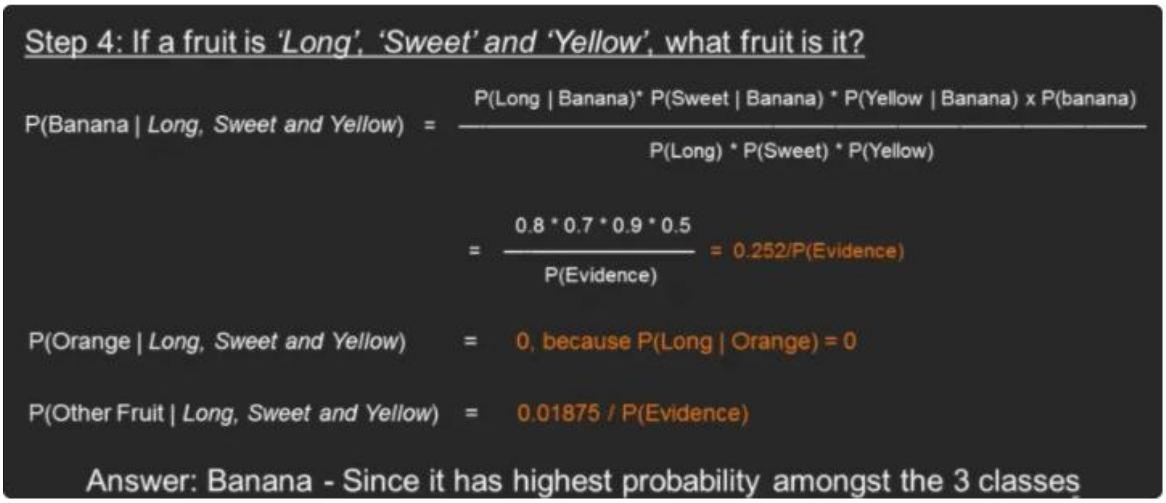
P(x1=Long | Y=Banana) = 400 / 500 = 0.80

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.



Step 4: If a fruit is *'Long', 'Sweet' and 'Yellow'*, what fruit is it?

$$P(\text{Banana} \mid \textit{Long, Sweet and Yellow}) = \frac{P(\text{Long} \mid \text{Banana})^{*}\ P(\text{Sweet} \mid \text{Banana}) * P(\text{Yellow} \mid \text{Banana}) \times P(\text{banana})}{P(\text{Long}) * P(\text{Sweet}) * P(\text{Yellow})}$$

$$= \frac{0.8 * 0.7 * 0.9 * 0.5}{P(\text{Evidence})} = 0.252/P(\text{Evidence})$$

$$P(\text{Orange} \mid \textit{Long, Sweet and Yellow}) = 0\text{, because } P(\text{Long} \mid \text{Orange}) = 0$$

$$P(\text{Other Fruit} \mid \textit{Long, Sweet and Yellow}) = 0.01875 / P(\text{Evidence})$$

Answer: Banana - Since it has highest probability amongst the 3 classes

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.