Bayes Theorem

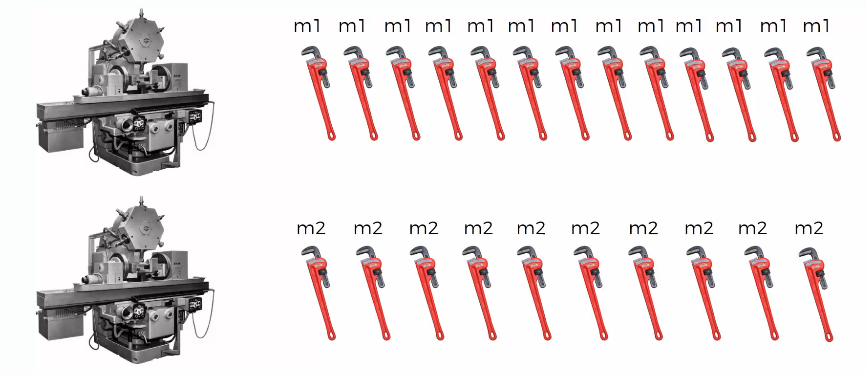
Lets say we are doing some analytics for a factory. We have two machines. Both the machines produce spanners. The spanners from machine1 are tagged as m1 and those from machine2 are tagged as m2 as shown. 

Figure 1 Tagging Spanners

The goal of the workers at the factory is to, at the end of the day, remove all the defective spanners from m1 and m2 together.

Now, one questions that we can ask is what is the probability that spanner from machine 2 is defective? To find the answer, we use the Bayes Theorem

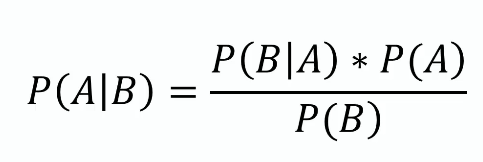


Figure 2 Bayes Theorem

The following information is provided to us:

* Machine 1: 30 wrenches/hr
* Machine 2: 20 wrenches/hr

After checking all the produced spanners at the end of the day, we can see that 1% are defective. Out of all the defective parts, we can see that 50% came from machine 1 and 50% came from machine 2.

Both machines 1 and 2 together produce 50 wrenches. Therefore, picking a random wrench from the pile being from machine 1 or machine 2 at any moment is

***P(machine 1) = 30/50 = 0.6***

***P(machine 2) = 20/60 = 0.4***

Also, from the given information

***P(Defect) = 1%***

Next, as stated, 50% of all defective parts came from machine 1. In mathematical terms we write it as

***P(Machine1 | Defect) = 50%***

Similarly

***P(Machine 2 | Defect) = 50%***

Finally, coming to our question, probability that a part produced by machine 2 is defective can be written as

***P(Defect | Machine 2) = [P(Machine 2 |Defect) \* P(Defect)] / P(Machine2)***

***= (0.5 \* 0.01) / 0.4***

***= 0.0125***

Intuition of Bayes Theorem

Lets say we have 1000 wrenches. So we know:

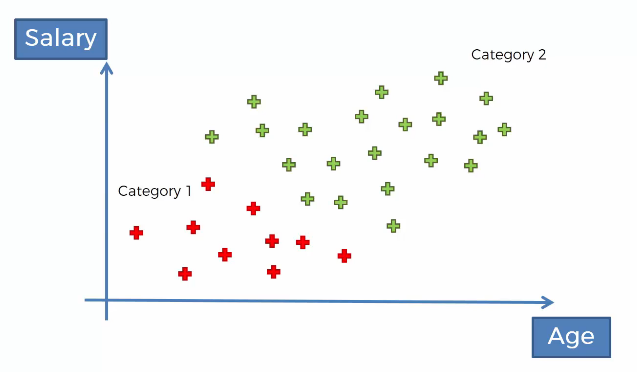
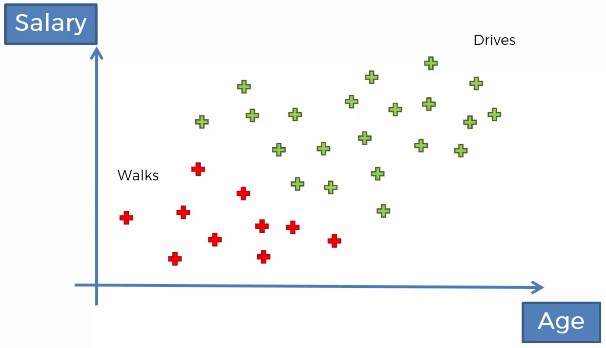
* 400 came from machine 2
* 1% have a defect = 10
* 50% of those 10 came from machine 2 = 5
* % of defective parts from machine 2 = 5/400 = 1.25%

The logical steps that we took here are the same as Bayes theorem.

Now the obvious question is *“If the items are labeled, why couldn’t we just count the number of defective wrenches that came from machine 2 and divide by the total number that came from machine 2?”*

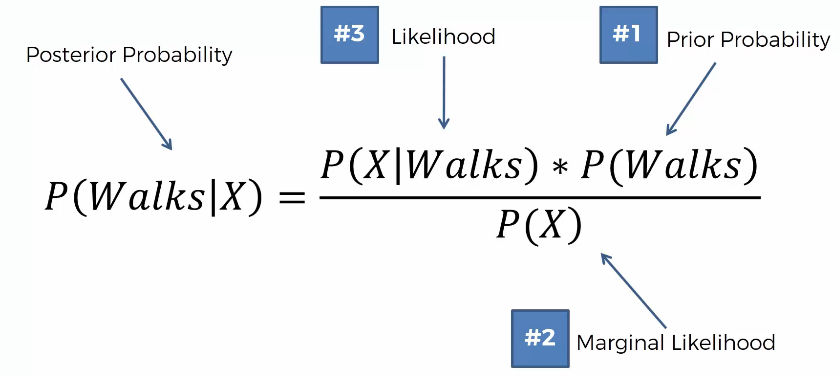
If we look at the equation of Bayes, we are indeed doing that. In the numerator, we are calculating the number of defective wrenches that came from machine 2 and in the denominator we have the total number of wrenches that came from machine 2.

Naïve Bayes Classifier Intuition

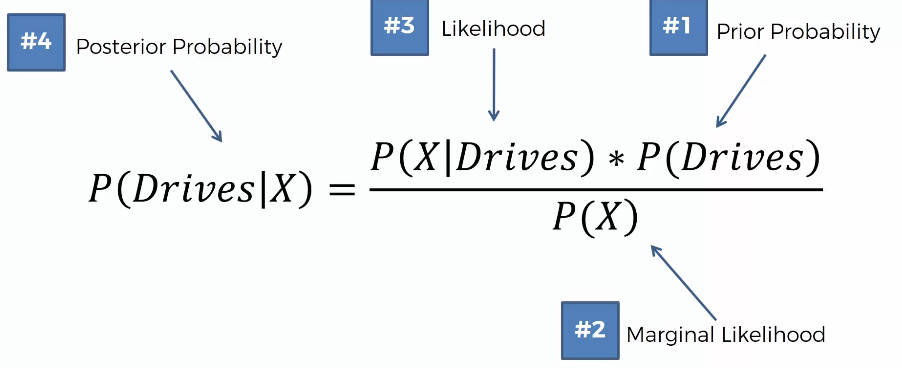
Lets say we have the following dataset.

We have to design a classification algorithm that will classify a new data point into either of the categories

We are going to apply the Bayes theorem twice. First, we are going to calculate the probability that this person walks given his features.

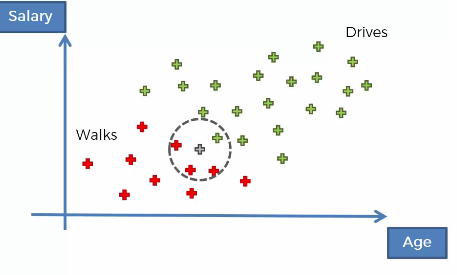


Second, we are going to calculate the probability that this person drives given his features.

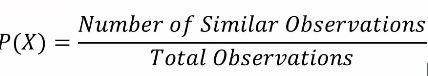


Finally, we are going to compare the above two probabilities.

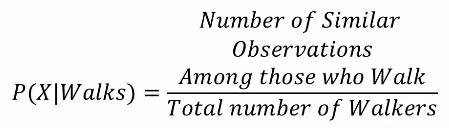
To find the marginal likelihood (#2 in the image above), we are going to select a radius and draw a circle with the new data point as the center.



After drawing the circle, we are going to look at all the points inside it. We will conclude that all the points in the circle are deemed to be similar in terms of features to our new data point.



NOTE: Number of Similar Observations in the above formula are all the points that fall inside the circle. Therefore P(X) = 4/30 in the above example



In the above formulae, Number of Similar Observations Among those who Walk is the number of walkers inside the circle. Thus, P(X | Walks) = 3/10

The highest probability between the two (P(Walks | X) vs P(Drives | X)) is the class into which the new data point gets classified. In the above example, since P (Walks | X) was 0.75 and P(Drives | X) was 0.25, the new data point will fall in the “Walks” category.

Additional Comments

1. Why “Naïve”? Since Bayes theorem requires that the variables (age and salary in this case) need to be independent. But these variables are not exactly independent. There is some correlation between the age and the salary of a person. This is the reason why the algorithm is called Naïve Bayes.

2. P(X) remains same while calculating P(Walks|X) as well as P(Drives|X). Therefore, we can remove P(X) during comparison. The final comparison then comes out to be

