**Naïve Bayes**

**Bayes’ Theorem:**

* Let’s say we are at a factory doing some analytics for the factory.
* There are two machines at the factory, both producing spanners.
* The machines work at different rates, and they have somewhat different characteristics, but overall, they’re producing the same spanners.
* Additional information: The spanners here are marked or tagged, so we know which machines they came from.
* And at the end of the day, we have got a whole pile of the spanners, the workers go through them, and their goal is pick out the defective spanners.
* The question we will be asking is – what’s the probability of Machine 2 producing a defective spanner?

So, if you take just a random spanner, produced by machine 2, what is the probability that the spanner is defective?

* The way we will get to that probability is through some information already given to us at the start.
* The rule that we will be using to get to that probability is called The Bayes Theorem

Text

Description automatically generated with medium confidence

* Given information:
* Machine 1 produces 30 wrenches/hr
* Machine 2 produces 20 wrenches/hr
* Out of all the produced parts, 1% are defective
* Out of all the defective parts, 50% came from Machine 1, and 50% came from Machine 2.
* So, here we can see that if we take all the defective parts, only the defective parts, and you calculate how many of them came from machine 1 and how many of them came from machine 2, you’ll see that half are from machine 1, and half from machine 2.
* So, the question is, what is the probability that a part produced by machine 2 is defective?
* **Re-writing the given information in more mathematical term:**
* M1 produces 30 wrenches/hr
* M2 produces 20 wrenches/hr
* Probability of any given wrench that you choose from the pile (defective or non-defective) that was produced by M1 -> **P(Mach1) = 30/50 = 0.6 = 60%**
* Probability of any given wrench that you choose from the pile (defective or non-defective) that was produced by M2 -> P(Mach2) = **20/50 = 0.4 = 40%**
* Out of all the defective parts, 1% is defective -> **P(Defect) = 1%**
* 50% of defective parts were produced by Machine 1, given that we are only picking up from defective pile -> **P(Mach1|Defect) = 50%**
* 50% of defective parts were produced by Machine 2, given that we are only picking up from defective pile -> **P(Mach2|Defect) = 50%**
* What is the probability that a part produced by Machine 2 is defective -> **P(Defect|Mach2) =?**
* So, we can tell from these expressions that the likelihood of a part coming from Machine 2 is 40%, meaning that it produces less wrenches.
* Whereas the likelihood of a defective part coming from Machine 2 is 50%.
* So, if you pick any wrench from the defective pile, the likelihood that it was originally produced by machine 2 is 50%.
* If you look at a part that’s defective, it’s 1%.
* But in P(Defect|Mach2), we are given a condition – parts are being produced by machine 2, and you just pick up a random one, what’s the probability of it being defective?
* Or you can think of it in terms of quantities – frequentist interpretation. Let’s say there’s a pile of wrenches that came out of machine 2, what are the number of wrenches from machine 2 that would be defective?
* So, it’s like finding the probability of any wrench produced by machine 2 is defective, or what is the portion of all these parts produced by machine 2 is going to be defective.

**Solution –**

* To begin with, we won’t need probabilities of machine 1 and probability of machine 1 given defective.
* Bayes theorem for this problem would be:

Where:

**P(Defect|Mach2)** – Probability of a defective part given that it came from machine 2

**P(Defect)** – Probability of a part being defective overall

**P(Mach2|Defect)** – Probability of a part produced by machine 2 given that it was defective

**P(Mach2)** – Probability of a part produced by machine 2

* So, when we plug in the numbers into the formula, we get:

***= 0.0125 = 1.25%***

* According to Bayes theorem, if Machine 2 produced 1,000 parts, 12.5 parts would be defective.

**Let’s look at an example –**

* We produced 1,000 wrenches in total (machine 1 + machine 2)
* 400 of those wrenches came from machine 2 (because machine produces 40% of the wrenches)
* 1% of all the total wrenches are defective = 10 are defective of all
* We also know that 50% of those 10 came from machine 2. Thus, out of all the defective wrenches produced in total, 5 of them came from machine 2.
* Percent of defective wrenches of machine 2 = 5/400 = 1.25%

**Naïve Bayes Classifier Intuition:**

* Bayes Theorem -
* How are we going to apply it to create a machine learning algorithm**Chart, scatter chart

  Description automatically generated**
* Here we have a dataset with two features, Age and Salary, along with two categories – whether the person walks to work or drives to work.
* What happens when you add a new data point to this set? How do we classify the new data point?

**Chart, scatter chart

Description automatically generated**

* Is this person going to be classified as a person who walks to work or a person who drives to work?
* And the Naïve Bayes algorithm is going to help us solve this challenge.

**Plan of Attack –**

We are going to take Bayes theorem, and apply it twice:

First time we are going to apply to find out, what is the probability that this person walks, given his features?

X in the formula is the features of that data point – Age and salary

Graphical user interface, application, Teams

Description automatically generated

In the first step we will calculate the prior probability, Marginal Likelihood, Likelihood, and Posterior Priority of the person walking to work.

Then we are going to calculate the probability of whether the new person drives to work

Graphical user interface, application, Teams

Description automatically generated

In the second step, we will calculate the prior probability, marginal likelihood, likelihood, and posterior probability of the person driving to work.

1. In the third step, we will compare the probability of the person walking to work, vs the person driving to work – P(Walk|X) vs P(Drive|X)

**Solution:**

**Step 1: -**

Prior Probability - P(Walks) =

=

Marginal Likelihood – P(X) – It tells us what’s the likelihood of any random number that we add to the dataset falling inside the circle.

The circle is drawn by selecting a radius and draw a circle around our observation. The radius can be selected by us. And then we look at the points inside the radius and deem them similar. Anybody who falls in that vicinity will be deemed similar to the new data point we add to the dataset.

Chart, scatter chart, bubble chart

Description automatically generated

Probability of X – P(X) – is the probability of a new point that we add to our dataset being similar in features to the point that we are adding to our dataset. Basically, it’s the probability of the new point or any random point to fall into that circle.

Likelihood – P(X|Walks) - The probability of someone who walks, and exhibits features X. For this, we will draw the same circle again, and anything we add to the circle will be deemed like the point that we are adding. So, what is the probability that a randomly selected data point from our dataset will be like the datapoint that we are adding. So, what is the probability that a randomly selected datapoint from the circle will be from the circle given that person walks. The other way to think about this is to think that we are only working with people who walk to work i.e., the red dots.

Therefore,

So, now if we plug all in, we will get our Posterior Probability.

**Step 2: -**

To repeat the step one for the likelihood that somebody with features X will/should be classified as a person who drives to work.

Graphical user interface, application, Teams

Description automatically generated

**Step 3: -**

* + Now we will be comparing the two probabilities.
  + P(Walks|X) vs P(Drives|X)
  + Its, 0.75 vs 0.25
  + And therefore, it is more likely that the person with features X walks to work than drives to work.

**Additional Information About Naïve Bayes: -**

* Why is it called “Naïve” Bayes Algorithm?

Because the Bayes theorem requires some independence assumptions, and the Bayes theorem is the foundation of the Naïve Bayes Machine Learning Algorithm. Therefore, the Naïve Bayes Machine Learning Algorithm also relies on these assumptions which are often incorrect. Therefore, it is naïve to assume that they are going to be correct.

* P(X):

In step 2 we took when we were calculating P(X), we took P(X), drew a circle around our new data point, removed the data point just so it’s not in the way. P(X) is the likelihood that a randomly selected point from the data set will exhibit the features like the data point that we will add. And as agreed, anything in that circle is like our data point.

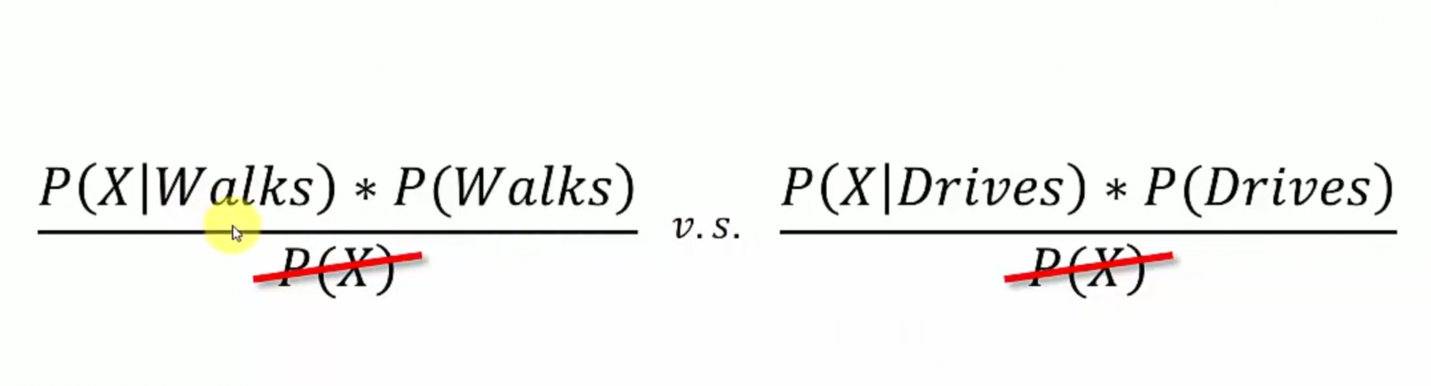
Another way to think about it is, what if we throw in a random variable, or a random data point into the dataset, what is the likelihood that it will fall into the circle – what is the likelihood that it will exhibit features like the point that we shall add in the dataset.

Where: -

Similar Observations = Observations similar to the points that we are about to add.

Total Observations = All the observations in the dataset.

Note – P(X) is the same irrespective of the probability that we are calculating i.e., irrespective of whether we are calculating the probability of someone walking to work, or commuting to work by a car, P(X) in our case would remain the same.



As P(X) is a common factor in both the equations, and the value of P(X) is same for both, we can multiply both sides by P(X) which will help us cancel out the denominators of both the equations without changing the signs, making the equation simpler to solve. And we can only compare the numerators of both the equations.

But we should **only** do that when comparing two probabilities. If we want to calculate the value, we can’t do that, as it would change the whole value of the result. Or if you are comparing the same situation from a different dataset – both the probabilities have P(X) but are from different datasets and have different values.

* What happens when we have more than 2 classes?

Remember, in this scenario we only had 2 classes, we had the green (people who walk to work) and the red (people who drive to work).

When we have only 2 classes, we compare the probability that the person who exhibits features X walks to work – that is the new data point that we added, and what is the probability that the person walks to work – vs what’s the probability that the new data point drives to work. It is a straightforward approach when we only have 2 classes, it will always add up to one.