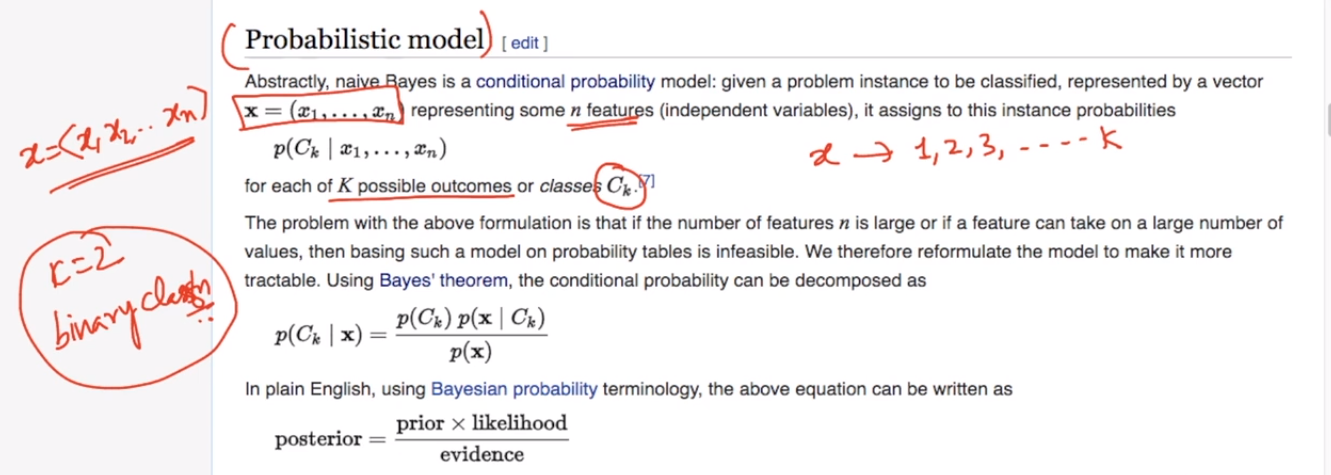
**Naïve Bayes:**

In naïve bayes term , naïve means unsophisticated or simplistic and it extensively uses Bayes theorem.

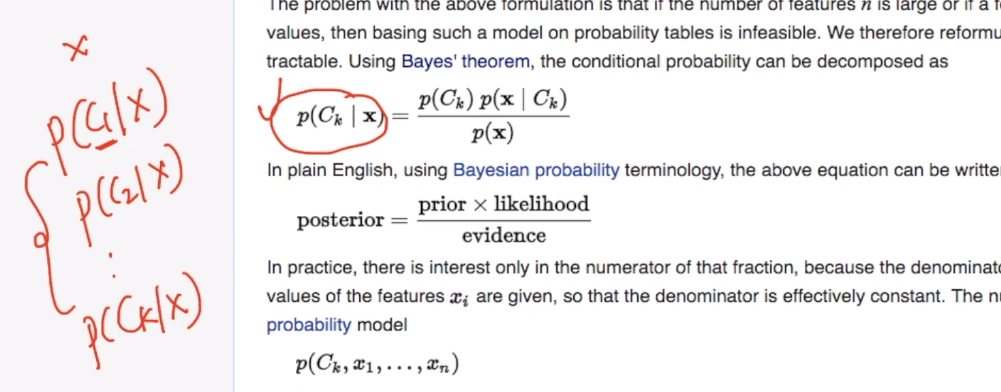
We will see how Naïve Bayes works in multiclass classification .



So what is the probability of C(k) given x which is nothing but set of features( x1,x2,….,xN features.)

According to Bayes theorem we already know

P(C(k)|x) = P( X|c(k)) P(C(k)) / P(x)

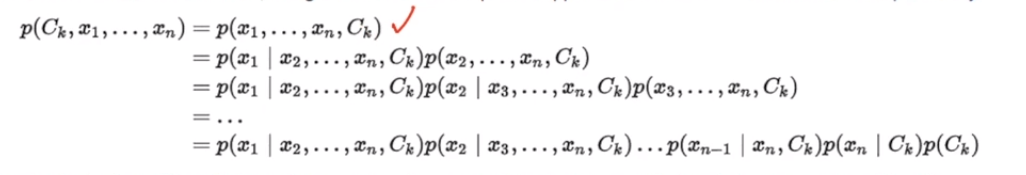


Foe whatever c(k) we will get maximum value we will select it as class label for x .

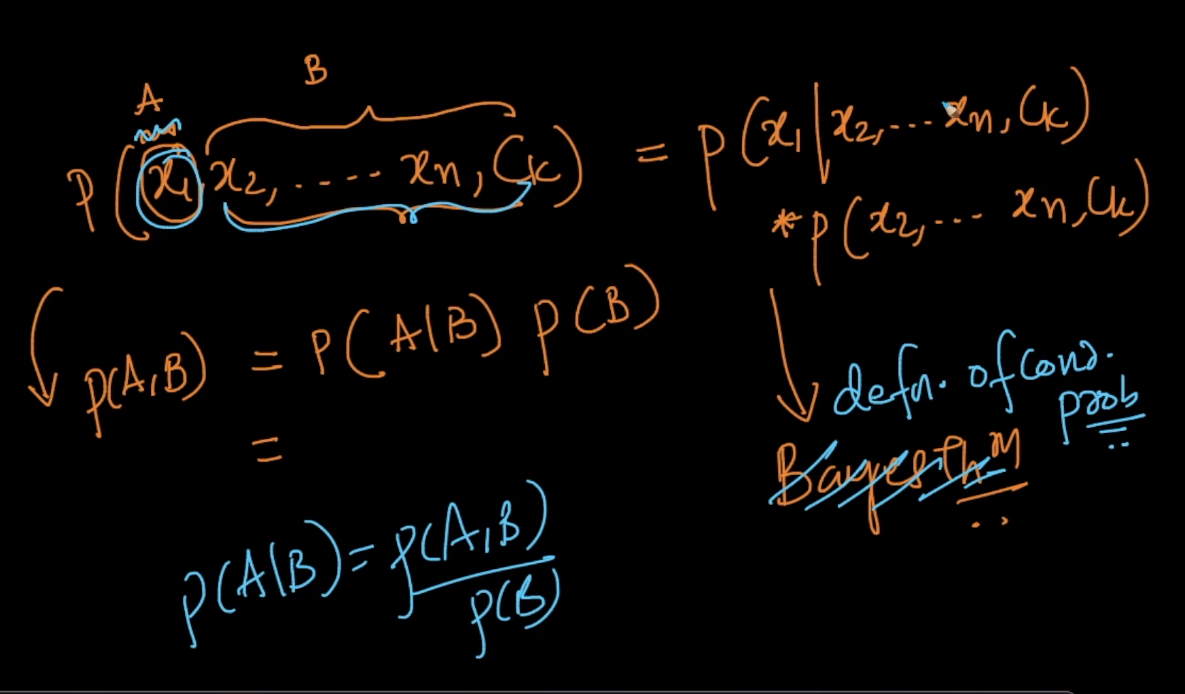
Since for every value of c(k) ,P(x) will be same so we will ignore it in the case and focus on P(C(k))\* P(x|C(k)) which is nothing but P(x and C(k))

So now we can say that P(C(k)|x) is proportional to P(C(k) and x)

There is something called chain rule in probability and it can be written as

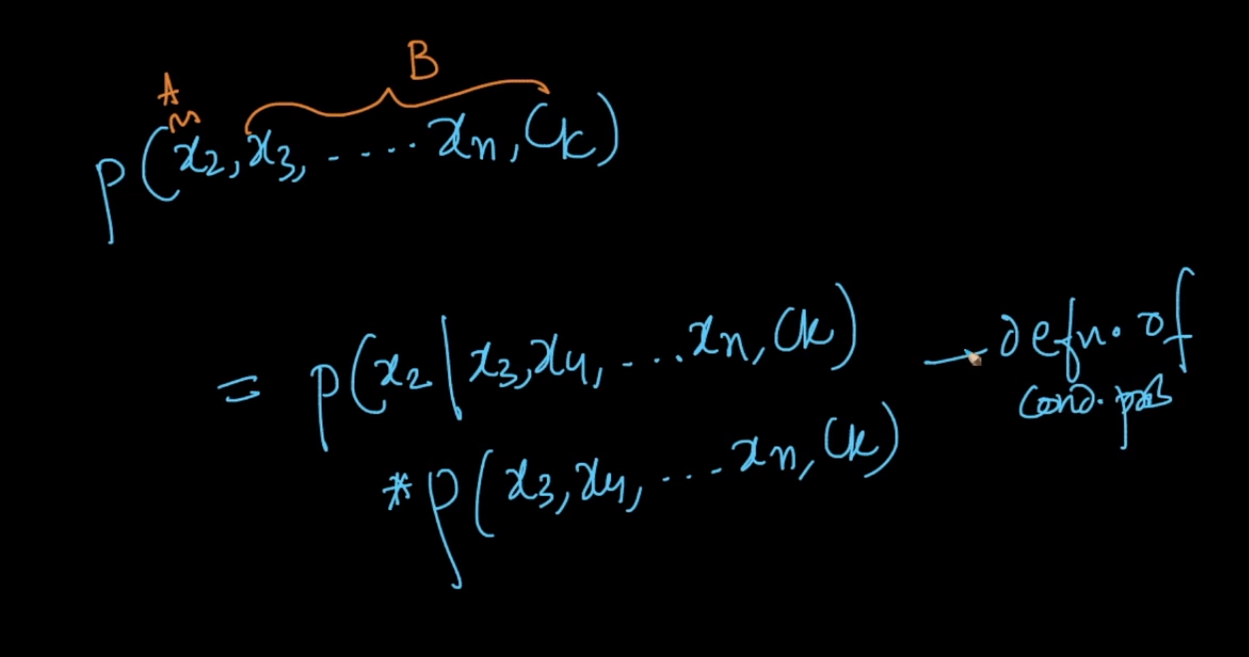


Now lets try to break out term:

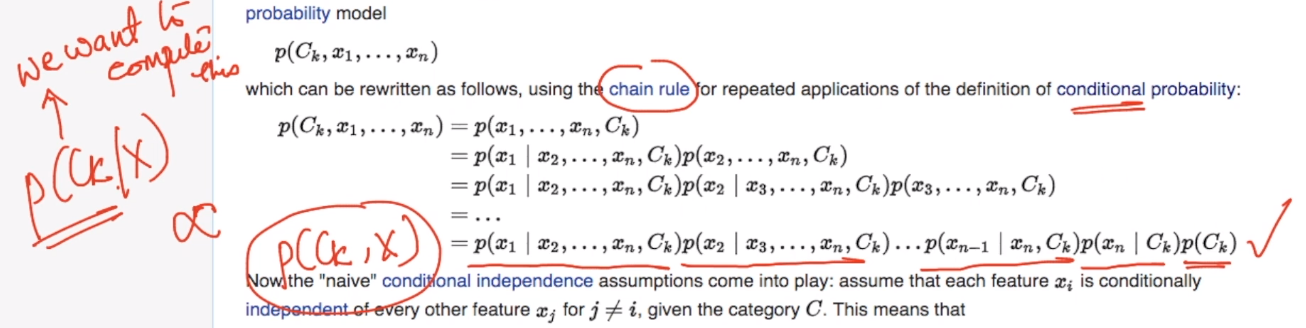


P(x1,x2,x3,….,xn,c(k)) = P(x1|x2,…,xn,ck) \* P(x2,…..,xn,ck) --------- just simple definition of cond. probability

Now lets try to break our second term using same method.

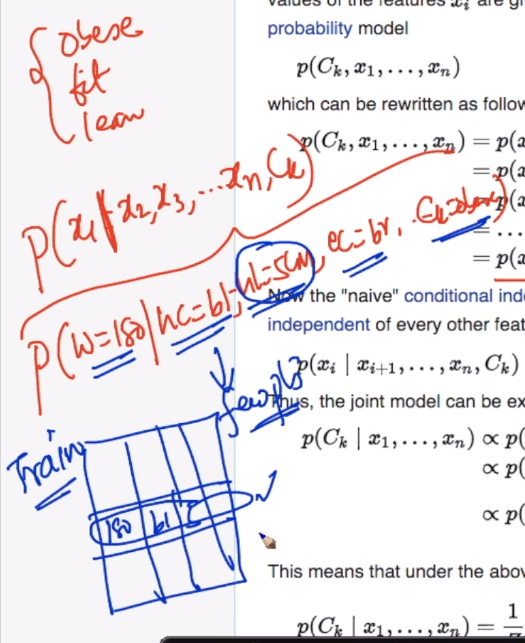


And applying same rule in continuous manner or say chain manner we get our final output as



And this is same what we want to compute.

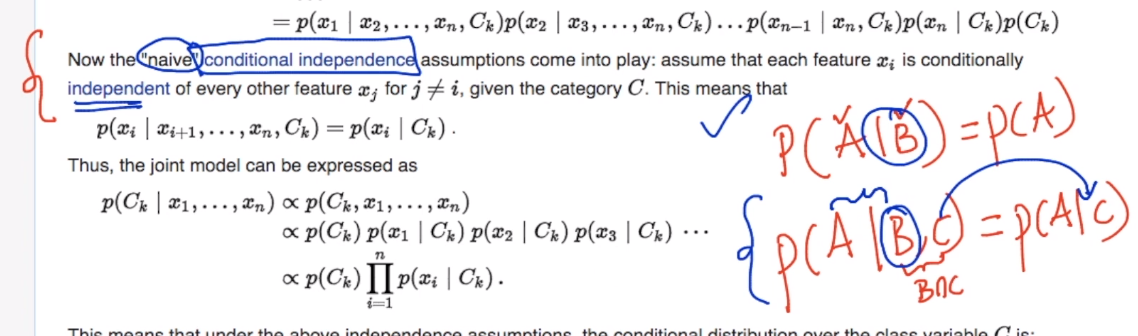
Taking a small real example:



What is the probability of weight = 180 when hair color = black , hair length = 5cm and eye color = brown and class label = obese

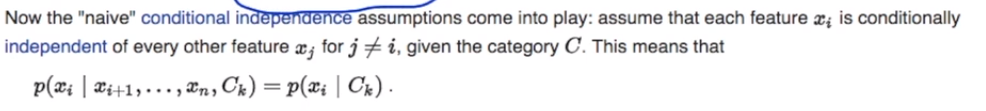
SO probability of finding exactly same data is itself very low and so it is very difficult to calculate such probabilities.

SO now we will see assumption about conditional independence

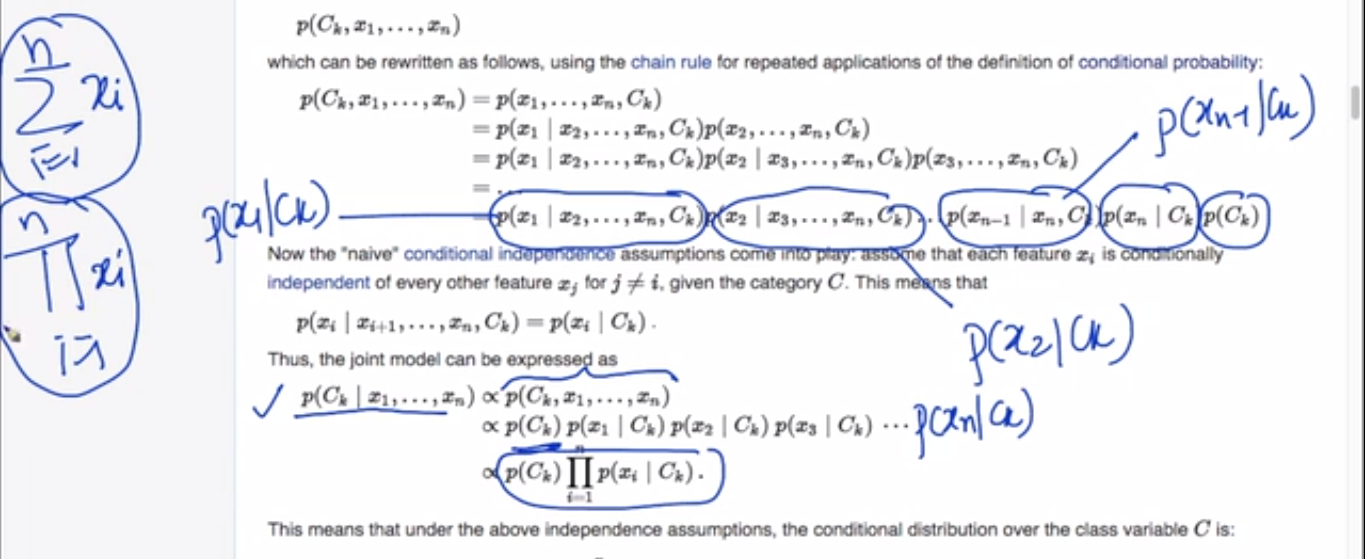


Here it means P(A|B,c) = P(A|C) means A in conditionally independent of B.

Now according to this conditional independence theory what we get is xi is conditionally independent of every other feature xj such that I is not equal to j

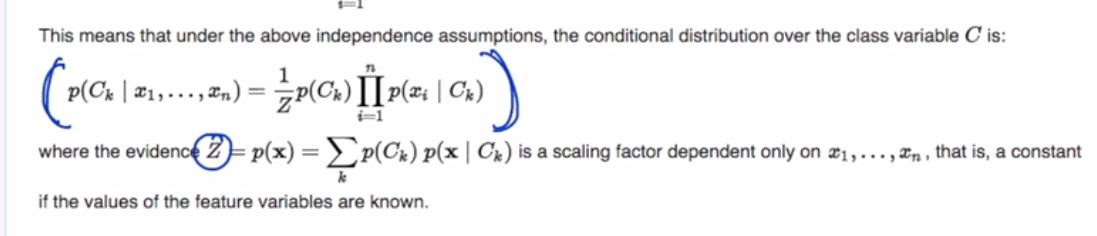


SO finally what we are left with in our equation of chair reaction is



**Pie means product here**

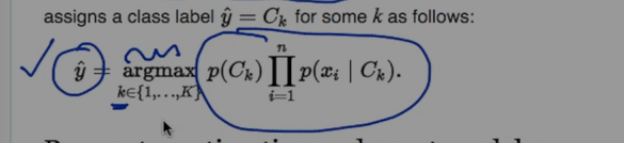
Now we get the final equation for P(c(k) |x) as



Z is the constant term which ignored just to simplify our calculations.

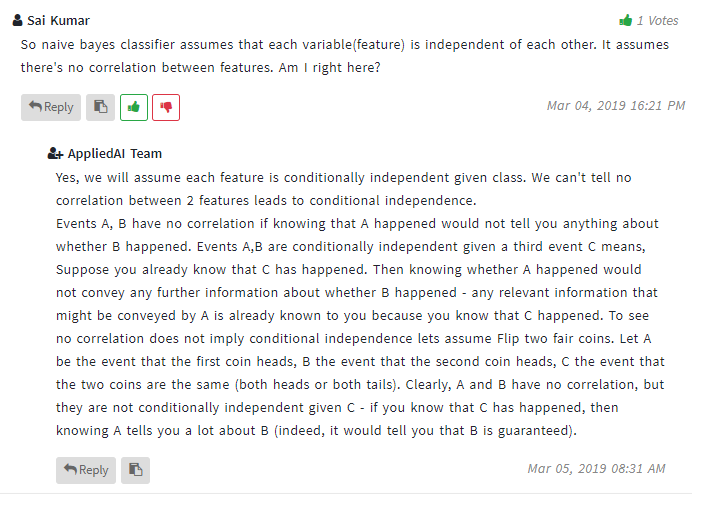
SO now I can calculate every (p(c1 |x), P(c2|x), ….,P(Cn|X))

And whichever gives us maximum value, we will call it as class label for x. and this phenomenon is called “**Maximum a posteriori”**



This is the mathematical part for Naïve Bayes algorithm.

Comments:



if you knew two variables were independent then their correlation must be zero, but it does not work the other way around.

<https://www.quora.com/Does-a-partial-correlation-of-0-imply-conditional-independence>

More explanation at : <https://www.youtube.com/watch?v=RixQygYyDKI&t=1s>

Naïve bayes performs well even if features are dependent.

