According to Bayes' theorem, the probability of a datapoint having class y given input features X is:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

where:

- $P(y|X)$ (**Posterior**) describes the prbability of finding y given these input features X
- $P(X|y)$ (**Likelihood**) describes the probability of finding the input features X given a target y
  - in terms of training data, this can be calculated from $\frac{\text{number of times feature } x \text{is found when class is } y}{\text{number of occurances of class } y}$
- $P(y)$ (**Prior**) is the probability of finding class y
- $P(X)$ (**Evidence**) is the probability of finding these input features out of the whole dataset

$P(X)$ is written explicitly as $P(x_1, x_2, x_3, \ldots)$ (probability of x_1 and x_2 and, ...)

We can factor $P(x_1, \ldots, x_n, |y)P(y)$ using the chain rule (see below) which gives us

$$\begin{aligned}
P(x_1, \ldots, x_n, y) &= P(x_1|x_2, \ldots, x_n, y)P(x_2, \ldots x_n, y) \\
&= P(x_1|x_2, \ldots, x_n, y)P(x_2|x_3, \ldots, x_n, y)P(x_3, \ldots, x_n, y) \\
&\ldots \\
&= P(x_1|x_2, \ldots, x_n, y)P(x_2|x_3, \ldots, x_n, y) \ldots P(x_n|y)p(y)
\end{aligned}$$

Here is where the naive assumption comes in
The term $P(x_1|x_2, \ldots, x_n, y)$ could be read as "probability of x_1 given x_2 and ... and x_n and y". If $x_1 \ldots, x_n$ are conditionally independent on y, that is knowing about x_2 or x_3, ... tells you nothing about x_1, then we can say that $P(x_1|x_2, \ldots, x_n, y) = P(x_1|y)$

So this whole factorization reduces to:

$$P(x_1, \ldots, x_n|y)P(y) = P(x_1, \ldots, x_n, y) = P(x_1|y)P(x_2|y) \ldots P(x_n|y)P(y) = P(y) \prod_{i=1}^{n} P(x_i|y)$$

Plugging this back into Baye's rule:

$$P(y|x_1, \ldots, x_n) = \frac{P(y) \prod P(x_i|y)}{P(X)}$$

since $P(X)$ does not change for a given class, we usually leave it out of the calculation:

$$P(y|x_1, \ldots, x_n) \propto P(y) \prod P(x_i|y)$$

So now all we need to do is find y that maximizes this proability and that's our prediction

$$y_{pred} = \text{argmax}_y P(y) \prod_{i=1}^{n} P(x_i|y)$$

Types of naive bayes

Multinomial
multiple categorical variables

Bernoulli
predictors are boolean values

Gaussian
predictors are continuous - *assume they come from a gaussian distribution* (**This means if you have continous variables you should normalize them before running the model**)

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \text{exp}(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2})$$

Advantages

- fast
- easy to implement

Disadvantages

- relies on predictors being independent, this is usually not the case (but sometimes you can get decent performance even if those assumptions are not met)

Chain Rule
$P(A, B) = P(A|B)P(B)$
$P(A, B, C) = P(A|B, C)P(B, C) = P(A|B, C)P(B|C)P(C)$

Refs

https://en.wikipedia.org/wiki/Naive_Bayes_classifier
https://en.wikipedia.org/wiki/Conditional_independence
https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c