Naive Bayes: It is a classifier based on Bayes theorem ~~Bayes~~ with an assumption that features are independent of each other.

Theorem:

Given a feature vector $X = (x_1, x_2, \dots x_n)$ and a class variable C_k, Bayes theorem states that

$$P(C_k|X) = \frac{P(X|C_k)\,P(C_k)}{P(X)}, \text{ for } k = 1, 2, \dots K.$$

we call $P(C_k|X)$ is a posterior probability

$P(X|C_k)$ is likelyhood

$P(C_k) \rightarrow$ Prior Probability of class

$P(X) \rightarrow$ Prior Probability of Predictors.

Using the chain rule, the formula $(P(X|C_k))$ can be decomposed as

$$\begin{aligned} P(X|C_k) &= P(x_1, x_2, \dots x_n | C_k) \\ &= P(x_1|x_2, \dots x_n, C_k)\, P(x_2|x_3, \dots x_n, C_k) \dots P(x_{n-1}|x_n, C_k)\, P(x_n|C_k) \\ &= \prod_{i=1}^{n} P(x_i|C_k) \end{aligned}$$

$$\therefore \boxed{P(X|C_k) = \prod_{i=1}^{n} (P(x_i|C_k))}$$

Posterior Probability can be written as

$$P(C_k|X) = \frac{P(C_k) \cdot \prod_{i=1}^{n} P(x_i|C_k)}{P(X)}$$

~~since the post~~ since the prior probability of predictor $P(X)$ is constant given the input, we can get

$$P(C_K|X) \propto P(C_K)\prod_{i=1}^{n} P(x_i|C_K)$$

Now, $P(C_K|X)$ can be calculated by finding maximum likelyhood of $P(C_K)\prod_{i=1}^{n}(P(x_i|C_K))$

$$\boxed{\hat{C} = \text{argmax}\left(P(C_K)\prod_{i=1}^{n} P(x_i|C_K)\right)}$$

Eg: Problem: For a given day with "outlook", "Humidity", "wind" conditions, Predict whether there will be game play or not.

Given data:

Day	outlook	wind	Humidity	Play
D1	Sunny	High	weak	NO
D2	Sunny	High	strong	No
D3	Overcast	High	weak	Yes
D4	Rain	High	weak	Yes
D5	Rain	Normal	week	yes
D6	Rain	Normal	strong	No
D7	overcast	Normal	Strong	yes
D8	Sunney	High	week	NO
D9	Sunny	Normal	week	Yes
D10	Rain	Normal	week	yes

First we will create a frequency table using each attribute

		Play yes	NO
outlook	sunny	1	3
	overcast	2	0
	Rainy	3	1

		Play yes	NO
Humidity	weak	5	2
	strong	1	2

		Play yes	NO
wind	High	2	3
	Normal	4	1

For each frequency table we will generate the ~~frequency~~ likelyhood table.

		Play yes	NO	
outlook	sunny	1/6	3/4	$\frac{4}{10}$
	overcast	2/6	0/4	$\frac{2}{10}$
	Rainy	3/6	1/4	$\frac{4}{10}$

$P(x|c) = P(\text{sunny}|\text{yes})$

$= \frac{1}{1+2+3} = \frac{1}{6} =$

$P(\text{Sunny}|\text{No}) = \frac{3}{3+0+1} = \frac{3}{4}$

$P(\text{overcast}|\text{yes}) = \frac{2}{1+2+3} = \frac{2}{6}$

$P(x) = P(\text{Sunny}) = \frac{1+3}{6+4} = \frac{4}{10}$

		Play		
		yes	No	
Humidity	weak	5/6	2/4	7/10
	strong	1/6	2/4	3/10

$P(x_1|c) = P(\text{weak}|\text{yes}) = \frac{5}{5+1} = \frac{5}{6}$

$P(x_2|c) = P(\text{strong}|\text{yes}) = \frac{1}{5+1} = \frac{1}{6}$

Similarly

$P(\text{weak}|\text{No}) = \frac{2}{4}$

$P(\text{strong}|\text{No}) = \frac{2}{4}$

$P(\text{weak}) = \frac{5+2}{6+4} = \frac{7}{10}$

$P(\text{strong}) = \frac{1+2}{6+4} = 3/10$

		yes	No	
wind	High	2/6	3/4	5/10
	Normal	4/6	1/4	5/10

$P(\text{High}|\text{yes}) = \frac{2}{2+4} = \frac{2}{6}$

$P(\text{High}|\text{No}) = \frac{3}{3+1} = \frac{3}{4}$

$P(\text{Normal}|\text{yes}) = \frac{4}{2+4} = \frac{4}{6}$

$P(\text{Normal}|\text{No}) = \frac{1}{3+1} = \frac{1}{4}$

$P(\text{High}) = \frac{2+3}{6+4} = \frac{5}{10}$

$P(\text{Normal}) = \frac{4+1}{6+4} = \frac{5}{10}$

Now, suppose we have a Day with following values

Outlook = Rainy

Humidity = ~~High~~ Strong

Wind = ~~Weak~~ Normal

Play = ?

We have to answer the question "whether there will be a play for the day or not?"

Likelyhood of "Yes" on that day

= P(~~x~~ outlook = Rainy | Yes) * P(Humidity = "Strong" | Yes) *
P(Wind = "~~Weak~~ Normal" | Yes)

$$= \frac{1}{6} * \frac{1}{6} * \frac{4}{6} = \frac{4}{216} = 0.018518$$

Likelyhood of "No" on that day

= P(outlook = "Rainy" | No) * P(Humidity = "Strong" | No) *
P(wind = "Normal" | No)

$$= \frac{1}{4} * \frac{2}{4} * \frac{1}{4} = \frac{2}{64} = 0.03125$$

$$P(\text{yes}) = \frac{0.018518}{(0.018518 + 0.03125)} = \frac{0.018518}{0.049768} = 0.38$$

$$P(\text{No}) = \frac{0.03125}{0.018518 + 0.03125} = 0.62$$

$P(yes) = 38\%$, $P(No) = 62\%$

It means there will 68% chance that Play = No and 38% chance that play = yes.

We can cross check this with the given data.
In day 6 we have similar conditions and the o/p was "No"

<u>Different Naive Bayes classifiers:</u>

(i) Gaussian } normal classification
(ii) Multinomial
(iii) Bernoulli } mainly used in text classification

All these different classifiers differ mainly by the assumptions they make regarding the distribution of $P(x_i|y)$

<u>(i) Gaussian Naive Bayes: (GaussianNB)</u>

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

where σ_y, μ_y are estimated using maximum likelyhood.

(ii) MultiNomialNB:

The distribution is parameterized by vectors

$$\theta_y = (\theta_{y1}, \dots, \theta_{yn})$$ for each class y.

where n = no. of features (in text classification, the size of vocabulary)

θ_{yi} is the probability $P(x_i|y)$ of feature i appearing in a sample belonging to class y.

so $$\theta_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha}$$

where

$N_{yi} = \sum_{x \in T} x_i$ is the no. of times feature i appears in a sample of class y in the training set T

$N_y = \sum_{i=1}^{n} N_{yi}$ is total count of all features for class y.

The smoothing priors $\alpha \geq 0$ accounts for features not present in the training samples and prevents zero probabilities in further computations.

Setting $\alpha = 1$ is called Laplace smoothing, while $\alpha < 1$ is called Lidstone smoothing.

(iii) BernoulliNB:

~~It is~~ $P(x_i|y)$ calculated as

$$P(x_i|y) = P(i|y)x_i + (1 - P(i|y))(1 - x_i)$$