

Naive Bayesian

The Naive Bayesian classifier is based on Bayes' theorem with the independence assumptions between predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods.

Algorithm

Bayes theorem provides a way of calculating the posterior probability, $P(c|x)$, from $P(c)$, $P(x)$, and $P(x/c)$. Naive Bayes classifier assume that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence.

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

Likelihood Class Prior Probability
↓ ↓
Posterior Probability Predictor Prior Probability

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \dots \times P(x_n|c) \times P(c)$$

- $P(c|x)$ is the posterior probability of *class (target)* given *predictor (attribute)*.
- $P(c)$ is the prior probability of *class*.
- $P(x/c)$ is the likelihood which is the probability of *predictor* given *class*.
- $P(x)$ is the prior probability of *predictor*.

Example 1:

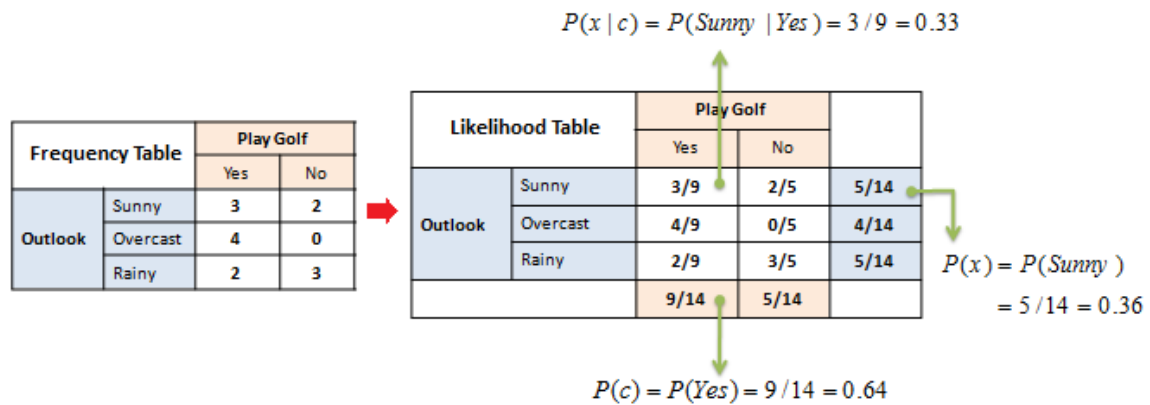
We use the same simple Weather dataset here.

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

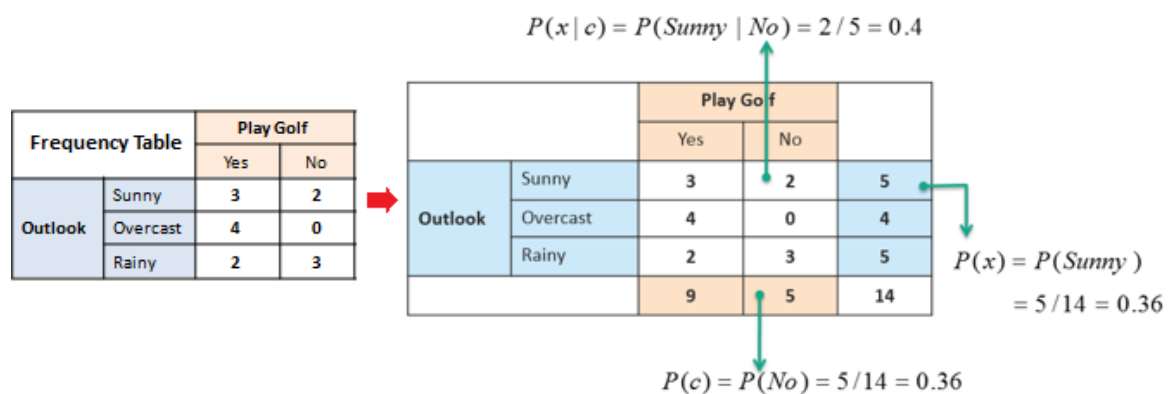
*** This material is taken on the reference webpage. ***

Reference: https://www.saedsayad.com/naive_bayesian.htm

The posterior probability can be calculated by first, constructing a frequency table for each attribute against the target. Then, transforming the frequency tables to likelihood tables and finally use the Naive Bayesian equation to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of prediction.



Posterior Probability: $P(c | x) = P(\text{Yes} | \text{Sunny}) = 0.33 \times 0.64 \div 0.36 = 0.60$



Posterior Probability: $P(c | x) = P(\text{No} | \text{Sunny}) = 0.40 \times 0.36 \div 0.36 = 0.40$

*** This material is taken on the reference webpage. ***

Reference: https://www.saedsayad.com/naive_bayesian.htm

The likelihood tables for all four predictors.

Frequency Table		Likelihood Table	
Outlook	Sunny	Play Golf	
		Yes	No
	Overcast	3	2
	Rainy	4	0
Humidity	High	Play Golf	
		Yes	No
	Normal	3	4
		6	1
Temp.	Hot	Play Golf	
		Yes	No
	Mild	2	2
	Cool	4	2
Windy	False	Play Golf	
		Yes	No
	True	2	2
		3	1

Example 2:

In this example we have 4 inputs (predictors). The final posterior probabilities can be standardized between 0 and 1.

Outlook	Temp	Humidity	Windy	Play
Rainy	Cool	High	True	?

$$P(Yes | X) = P(Rainy | Yes) \times P(Cool | Yes) \times P(High | Yes) \times P(True | Yes) \times P(Yes)$$

$$P(Yes | X) = 2/9 \times 3/9 \times 3/9 \times 3/9 \times 9/14 = 0.00529 \rightarrow 0.2 = \frac{0.00529}{0.02057 + 0.00529}$$

$$P(No | X) = P(Rainy | No) \times P(Cool | No) \times P(High | No) \times P(True | No) \times P(No)$$

$$P(No | X) = 3/5 \times 1/5 \times 4/5 \times 3/5 \times 5/14 = 0.02057 \rightarrow 0.8 = \frac{0.02057}{0.02057 + 0.00529}$$

*** This material is taken on the reference webpage. ***

Reference: https://www.saedsayad.com/naive_bayesian.htm

The zero-frequency problem

Add 1 to the count for every attribute value-class combination (Laplace estimator) when an attribute value (Outlook=Overcast) doesn't occur with every class value (Play Golf=no).

Numerical Predictors

Numerical variables need to be transformed to their categorical counterparts (binning) before constructing their frequency tables. The other option we have is using the distribution of the numerical variable to have a good guess of the frequency. For example, one common practice is to assume normal distributions for numerical variables.

The probability density function for the normal distribution is defined by two parameters (mean and standard deviation).

$$\begin{aligned}\mu &= \frac{1}{n} \sum_{i=1}^n x_i && \text{Mean} \\ \sigma &= \left[\frac{1}{n-1} \sum_{i=1}^n (x_i - \mu)^2 \right]^{0.5} && \text{Standard deviation} \\ f(x) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} && \text{Normal distribution}\end{aligned}$$

Binning

Binning or discretization is the process of transforming numerical variables into categorical counterparts. An example is to bin values for Age into categories such as 20-39, 40-59, and 60-79. Numerical variables are usually discretized in the modeling methods based on frequency tables (e.g., decision trees). Moreover, binning may improve accuracy of the predictive models by reducing the noise or non-linearity. Finally, binning allows easy identification of outliers, invalid and missing values of numerical variables.

Example:

		Humidity										Mean	StDev
Play Golf	yes	86	96	80	65	70	80	70	90	75	79.1	10.2	
	no	85	90	70	95	91					86.2	9.7	

$$P(\text{humidity} = 74 \mid \text{play} = \text{yes}) = \frac{1}{\sqrt{2\pi}(10.2)} e^{-\frac{(74-79.1)^2}{2(10.2)^2}} = 0.0344$$

$$P(\text{humidity} = 74 \mid \text{play} = \text{no}) = \frac{1}{\sqrt{2\pi}(9.7)} e^{-\frac{(74-86.2)^2}{2(9.7)^2}} = 0.0187$$

*** This material is taken on the reference webpage. ***

Reference: https://www.saedsayad.com/naive_bayesian.htm