In the fake two-headed quarter example, $P(F)$ is called the prior probability and $P(F|H)$ is called the posterior probability.

Naive Bayes Classification

Despite its simplicity, Naive Bayes—sometimes called *Idiot's Bayes*[1]—often beats more sophisticated algorithms. It works well on large data sets, e.g. spam filtering and topic modeling.

Naive Bayes assumes conditional independence.

16 Definition (Conditional Independence)
Events A and B given C are conditionally independent if

$$P(A \cap B|C) = P(A|C)P(B|C)$$

17 Example (Naive Bayes Classification)
Use the training data below and Naive Bayes classification to predict the probabilities of the target labels for the attribute values $A = 1$ and $B = 0$, i.e. compute the probabilities:

$$P(-|A = 1, B = 0) \quad P(+|A = 1, B = 0)$$

training data

	A	B	T
1.	1	0	$+$
2.	0	1	$+$
3.	0	1	$-$
4.	0	0	$-$
5.	1	1	$-$

	A	B	T
1.	1	0	?

Solution:

[1] Many people apply Naive Bayes thinking they are using Bayes Theorem when in fact they are using an approximation of Bayes Theorem.

The Naive Bayes classifier computes the probabilities $P(\text{class}_k|\text{attributes})$, $k = 1, 2, \ldots, N$ using Bayes Theorem and selects the largest one. Bayes Theorem implies:

$$P(\text{class}_k|\text{attributes}) = \frac{P(\text{attributes}|\text{class}_k)\, P(\text{class}_k)}{P(\text{attributes})}.$$

Applying the conditional independence assumption, we have that

$$P(\text{attributes}|\text{class}_k) = P(\text{attrib}_1|\text{class}_k) P(\text{attrib}_2|\text{class}_k) \cdots P(\text{attrib}_n|\text{class}_k)$$

where n equals the number of attributes. The probability $P(\text{attrib}_j|\text{class}_k)$ equals the fraction of class_k records that have attribute attrib_j. One of the problems that can occur in practice with this conditional independence assumption is that if any attribute attrib_j is missing for class class_k, then $P(\text{attributes}|\text{class}_k) = 0$. A common fix is to add a fixed number of "pseudo-counts" (fake counts) to all the attributes so none are missing. The number of pseudo counts (which can be a fraction of a count) is a key hyper-parameter of the Naive Bayes classifier. The optimal pseudo counts to add is determined using cross-validation.