4.4.1 How it Works

The naïve Bayesian algorithm is built on Bayes' theorem, named after Reverend Thomas Bayes. Bayes' work is described in "Essay Towards Solving a Problem in the Doctrine of Chances" (1763), published posthumously in the *Philosophical Transactions of the Royal Society of London* by Richard Price. Bayes' theorem is one of the most influential and important concepts in statistics and probability theory. It provides a mathematical expression for how a degree of subjective belief changes to account for new evidence. First, let's discuss the terminology used in Bayes' theorem.

Assume **X** is the evidence (or factors or attribute set) and Y is the outcome (or target or label class). Here **X** is a set, not individual attributes, hence **X** = $\{X_1, X_2, X_3, \ldots, X_n\}$, where X_i is an individual attribute, such as credit rating. The probability of outcome P(Y) is called *prior probability*, which can be calculated from the data set. Prior probability shows the likelihood of an outcome in a given data set. For example, in the mortgage case, P(Y) is the default rate of a home mortgage, which is 2%. P(Y|**X**) is called the *conditional probability*, which provides the probability of an outcome given the evidence when we know the value of **X**. Again, using the mortgage example, P(Y|**X**) is the average rate of default given that an individual's credit history is known. If the credit history is excellent, then the probability of default is likely to be less than 2%. P(Y|**X**) is also called *posterior probability*. Calculating posterior probability is the objective of predictive analytics using Bayes' theorem. This is the likelihood of an outcome as we learn the values of the input attributes.

Bayes' theorem states that

$$P(Y|\mathbf{X}) = \frac{P(Y) * P(\mathbf{X}|Y)}{P(\mathbf{X})} \tag{4.13}$$

P(**X**|Y) is another conditional probability, called the *class conditional probability*. P(**X**|Y) is the probability that an attribute assumes a particular value given the class label. Like P(Y), P(**X**|Y) can be calculated from the data set as well. If we know the training set of loan defaults, we can calculate the probability of an "excellent" credit rating given that the default is a "yes." As indicated in Bayes' theorem, class conditional probability is crucial in calculating posterior probability. P(**X**) is basically the probability of the evidence. In the mortgage example, this is simply the proportion of individuals with a given credit rating. To classify a new record, we can compute P(Y|**X**) for *each class of Y* and see which probability "wins." Class label Y with the highest value of P(Y|**X**) wins for a particular attribute value **X**. Since P(**X**) is the same for every class value of the outcome, we don't have to calculate this and assume it as a constant. More generally, in an example set with n attributes **X** = $\{X_1, X_2, X_3 \ldots X_n\}$,

$$P(Y|\mathbf{X}) = \frac{P(Y) * \prod_{i=1}^{n} P(\mathbf{X_i}|Y)}{P(\mathbf{X})} \tag{4.14}$$

If we know how to calculate class conditional probability $P(\mathbf{X}|Y)$ or $\prod_{i=1}^{n} P(X_i \mid Y)$, then it is easy to calculate posterior probability $P(Y|\mathbf{X})$. Since $P(X)$ is constant for every value of Y, it is enough to calculate the numerator of the equation $P(Y) * \prod_{i=1}^{n} P(X_i \mid Y)$ for every class value.

To further explain how the naïve Bayesian algorithm works, let's use the modified Golf data set shown in Table 4.4. The Golf table is an artificial data set with four attributes and one class label. Note that we are using the nominal data table for easier explanation (temperature and humidity have been converted from the numeric type). In Bayesian terms, weather condition is the evidence and decision to play or not play is the belief. Altogether there are 14 examples with 5 examples of Play = no and nine examples of Play = yes. The objective is to predict if the player will Play (yes or no), given the information about a few weather-related measures, based on learning from the data set in Table 4.4. Here is the step-by-step explanation of how the Bayesian model works.

Step 1: Calculating Prior Probability P(Y)

Prior probability P(Y) is the probability of an outcome. In this example set there are two possible outcomes: Play = yes and Play = no. From Table 4.4, out of 14 records there are 5 records with the "no" class and 9 records with the "Yes" class. The probability of outcome is

$$P(Y = \text{no}) = 5/14$$
$$P(Y = \text{yes}) = 9/14$$

Table 4.4 Golf Data Set with Modified Temperature and Humidity Attributes

No.	Temperature X_1	Humidity X_2	Outlook X_3	Wind X_4	Play (Class Label) Y
1	high	med	sunny	false	no
2	high	high	sunny	true	no
3	low	low	rain	true	no
4	med	high	sunny	false	no
5	low	med	rain	true	no
6	high	med	overcast	false	yes
7	low	high	rain	false	yes
8	low	med	rain	false	yes
9	low	low	overcast	true	yes
10	low	low	sunny	false	yes
11	med	med	rain	false	yes
12	med	low	sunny	true	yes
13	med	high	overcast	true	yes
14	high	low	overcast	false	yes

Since the probability of an outcome is calculated from the data set, it is important that the data set used for data mining is *representative* of the population, if sampling is used. A class-stratified sampling of data from the population will not be compatible for naïve Bayesian modeling.

Step 2: Calculating Class Conditional Probability $P(X_i \mid Y)$

Class conditional probability is the probability of *each* attribute value for an attribute, for each outcome value. This calculation is repeated for all the attributes: Temperature (X_1), Humidity (X_2), Outlook (X_3), and Wind(X_4), and for every distinct outcome value. Let's calculate the class conditional probability of Temperature (X_1). For each value of the Temperature attribute, we can calculate $P(X_1|Y = \text{no})$ and $P(X_1|Y = \text{yes})$ by constructing a probability table as shown in Table 4.5. From the data set there are five Y = no records and nine Y = yes records. Out of the five Y = no records, we can also calculate the probability of occurrence when the temperature is high, medium, and low. The values will be 2/5, 1/5, and 2/5, respectively. We can repeat the same process when the outcome Y = yes.

Similarly, we can repeat the calculation to find the class conditional probability for the other three attributes: Humidity (X_2), Outlook (X_3), and Wind(X_4). This class conditional probability table is shown in Table 4.6.

Table 4.5 Class Conditional Probability of Temperature

Temperature (X_1)	**P(X_1|Y = no)**	**P(X_1|Y = yes)**
high	2/5	2/9
med	1/5	3/9
low	2/5	4/9

Table 4.6 Conditional Probability of Humidity, Outlook, and Wind

Humidity (X_2)	**P(X_1|Y = no)**	**P(X_1|Y = yes)**
high	2/5	2/9
low	1/5	4/9
med	2/5	3/9
Outlook (X_3)	**P(X_1|Y = no)**	**P(X_1|Y = yes)**
overcast	0/5	4/9
Rain	2/5	3/9
sunny	3/5	2/9
Wind (X_4)	**P(X_1|Y = no)**	**P(X_1|Y = yes)**
false	2/5	6/9
true	3/5	3/9

Table 4.7 Test Record

No.	Temperature X_1	Humidity X_2	Outlook X_3	Wind X_4	Play (Class Label) Y
Unlabeled Test	high	low	sunny	false	?

Step 3: Predicting the Outcome Using Bayes' Theorem

We are all set with preparing class conditional probability tables and now they can be used in the future prediction task. If a new, unlabeled test record (Table 4.7) has the attribute values Temperature= high, Humidity = low, Outlook = sunny, and Wind = false, what would be the class label prediction? Play = Yes or No? The outcome class can be predicted based on Bayes' theorem by calculating the posterior probability $P(Y|\mathbf{X})$ for both values of Y. Once $P(Y = \text{yes}|\mathbf{X})$ and $P(Y = \text{no}|\mathbf{X})$ are calculated, we can determine which outcome has higher probability and the predicted outcome is the one that has the highest probability. While calculating both class conditional probabilities using Equation 4.14, it is sufficient to just calculate $P(Y) * \prod_{i=1}^{n} P(Xi|\, Y)$ as $P(\mathbf{X})$ is going to be same for both the outcome classes.

$$P(Y = \text{yes}|\mathbf{X}) = \frac{P(Y) * \prod_{i=1}^{n} P(Xi\,|\,Y)}{P(X)}$$

$$= P(Y = \text{yes}) * \{P(\text{Temp} = \text{high}|Y = \text{yes}) * P(\text{Humidity} = \text{low}|Y = \text{yes}) *$$
$$P(\text{Outlook} = \text{sunny}|\ Y = \text{yes}) * P(\text{Wind} = \text{false}|Y = \text{yes})\}/P(X)$$
$$= 9/14 * \{2/9 * 4/9 * 2/9 * 6/9\}/P(X)$$
$$= 0.0094/P(X)$$
$$P(Y = \text{no}|\mathbf{X}) = 5/14 * \{2/5 * 4/5 * 3/5 * 2/5\}$$
$$= 0.0274/P(X)$$

We normalize both the estimates by dividing both by (0.0094 + 0.027) to get

$$\text{Likelihood of (Play = yes)} = \frac{0.0094}{0.0274 + 0.0094} = 26\%$$

$$\text{Likelihood of (Play = no)} = \frac{0.0094}{0.0274 + 0.0094} = 74\%$$

In this case $P(Y = \text{yes}|X) < P(Y = \text{no}|X)$, hence the prediction for the unlabeled test record will be Play = no.

Bayesian modeling is relatively simple to understand once you get past the notation (for beginners) and easy to implement in practically any programing language. The computation for model building is quite simple and involves the creation of a lookup table of probabilities. Bayesian modeling is quite robust in handling missing values. If the test example set does not contain a