Chapter 6

Bayesian Concept Learning

OBJECTIVE OF THE CHAPTER:

Principles of probability for classification are an important area of machine learning algorithms. In our practical life, our decisions are affected by our prior knowledge or belief about an event. Thus, an event that is otherwise very unlikely to occur may be considered by us seriously to occur in certain situations if we know that in the past, the event had certainly occurred when other events were observed. The same concept is applied in machine learning using Bayes' theorem and the related algorithms discussed in this chapter. The concepts of probabilities discussed in the previous chapters are used extensively in this chapter.

6.1 INTRODUCTION

In the last chapter, we discussed the rules of probability and possible uses of probability, distribution functions, and hypothesis testing principles in the machine learning domain. In this chapter, we will discuss the details of the Bayesian theorem and how it provides the basis for machine learning concepts. The technique was derived from the work of the 18th century mathematician Thomas Bayes. He developed the foundational mathematical principles, known as Bayesian methods, which describe the probability of events, and more importantly, how probabilities should be revised when there is additional information available.

6.2 WHY BAYESIAN METHODS ARE IMPORTANT?

Bayesian learning algorithms, like the naive Bayes classifier, are highly practical approaches to certain types of learning problems as they can calculate explicit probabilities for hypotheses. In many cases, they are

equally competitive or even outperform the other learning algorithms, including decision tree and neural network algorithms.

Bayesian classifiers use a simple idea that the training data are utilized to calculate an observed probability of each class based on feature values. When the same classifier is used later for unclassified data, it uses the observed probabilities to predict the most likely class for the new features. The application of the observations from the training data can also be thought of as applying our prior knowledge or prior belief to the probability of an outcome, so that it has higher probability of meeting the actual or real-life outcome. This simple concept is used in Bayes' rule and applied for training a machine in machine learning terms. Some of the real-life uses of Bayesian classifiers are as follows:

- Text-based classification such as spam or junk mail filtering, author identification, or topic categorization
- Medical diagnosis such as given the presence of a set of observed symptoms during a disease, identifying the probability of new patients having the disease
- Network security such as detecting illegal intrusion or anomaly in computer networks

One of the strengths of Bayesian classifiers is that they utilize all available parameters to subtly change the predictions, while many other algorithms tend to ignore the features that have weak effects. Bayesian classifiers assume that even if few individual parameters have small effect on the outcome, the collective effect of those parameters could be quite large. For such learning tasks, the naive Bayes classifier is most effective.

Some of the features of Bayesian learning methods that have made them popular are as follows:

- Prior knowledge of the candidate hypothesis is combined with the observed data for arriving at the final probability of a hypothesis. So, two important components are the prior probability of each candidate hy-

pothesis and the probability distribution over the observed data set for each possible hypothesis.

- The Bayesian approach to learning is more flexible than the other approaches because each observed training pattern can influence the outcome of the hypothesis by increasing or decreasing the estimated probability about the hypothesis, whereas most of the other algorithms tend to eliminate a hypothesis if that is inconsistent with the single training pattern.
- Bayesian methods can perform better than the other methods while validating the hypotheses that make probabilistic predictions. For example, when starting a new software project, on the basis of the demographics of the project, we can predict the probability of encountering challenges during execution of the project.
- Through the easy approach of Bayesian methods, it is possible to classify new instances by combining the predictions of multiple hypotheses, weighted by their respective probabilities.
- In some cases, when Bayesian methods cannot compute the outcome deterministically, they can be used to create a standard for the optimal decision against which the performance of other methods can be measured.

As we discussed above, the success of the Bayesian method largely depends on the availability of initial knowledge about the probabilities of the hypothesis set. So, if these probabilities are not known to us in advance, we have to use some background knowledge, previous data or assumptions about the data set, and the related probability distribution functions to apply this method. Moreover, it normally involves high computational cost to arrive at the optimal Bayes hypothesis.

6.3 BAYES' THEOREM

Before we discuss Bayes' theorem and its application in concept learning, we should be clear about what is **concept learning**. Let us take an example of how a child starts to learn meaning of new words, e.g. 'ball'. The child is provided with positive examples of 'objects' which are 'ball'. At

first, the child may be confused with many different colours, shapes and sizes of the balls and may also get confused with some objects which look similar to ball, like a balloon or a globe. The child's parent continuously feeds her positive examples like 'that is a ball', 'this is a green ball', 'bring me that small ball', etc. Seldom there are negative examples used for such concept teaching, like 'this is a non-ball', but the parent may clear the confusion of the child when it points to a balloon and says it is a ball by saying 'that is not a ball'. But it is observed that the learning is most influenced through positive examples rather than through negative examples, and the expectation is that the child will be able to identify the object 'ball' from a wide variety of objects and different types of balls kept together once the concept of a ball is clear to her. We can extend this example to explain how we can expect machines to learn through the feeding of positive examples, which forms the basis for concept learning.

To relate the above-mentioned learning concept with the mathematical model of Bayes, we can correlate the learning process of 'meaning of a word' as equivalent to learning, a concept using binary classification. Let us define a concept set C and a corresponding function $f(k)$. We also define $f(k) = 1$, when k is within the set C and $f(k) = 0$ otherwise. Our aim is to learn the indicator function f that defines which elements are within the set C. So, by using the function f, we will be able to classify the element either inside or outside our concept set. In Bayes' theorem, we will learn how to use standard probability calculus to determine the uncertainty about the function f, and we can validate the classification by feeding positive examples.

There are few notations that will be introduced before going into the details of Bayes' theorem. In Chapter 5, we already discussed Bayes' probability rule as given below:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

where A and B are conditionally related events and $p(A|B)$ denotes the probability of event A occurring when event B has already occurred.

Let us assume that we have a training data set D where we have noted some observed data. Our task is to determine the best hypothesis in space H by using the knowledge of D.

6.3.1 Prior

The prior knowledge or belief about the probabilities of various hypotheses in H is called Prior in context of Bayes' theorem. For example, if we have to determine whether a particular type of tumour is malignant for a patient, the prior knowledge of such tumours becoming malignant can be used to validate our current hypothesis and is a prior probability or simply called Prior.

Let us introduce few notations to explain the concepts. We will assume that $P(h)$ is the initial probability of a hypothesis 'h' that the patient has a malignant tumour based only on the malignancy test, without considering the prior knowledge of the correctness of the test process or the so-called training data. Similarly, $P(T)$ is the prior probability that the training data will be observed or, in this case, the probability of positive malignancy test results. We will denote $P(T|h)$ as the probability of observing data T in a space where 'h' holds true, which means the probability of the test results showing a positive value when the tumour is actually malignant.

6.3.2 Posterior

The probability that a particular hypothesis holds for a data set based on the Prior is called the posterior probability or simply Posterior. In the above example, the probability of the hypothesis that the patient has a malignant tumour considering the Prior of correctness of the malignancy test is a posterior probability. In our notation, we will say that we are interested in finding out $P(h|T)$, which means whether the hypothesis holds true given the observed training data T. This is called the posterior proba-

bility or simply Posterior in machine learning language. So, the prior probability $P(h)$, which represents the probability of the hypothesis independent of the training data (Prior), now gets refined with the introduction of influence of the training data as $P(h|T)$.

According to Bayes' theorem

$$P(h|T) = \frac{P(T|h)P(h)}{P(T)}$$

combines the prior and posterior probabilities together.

From the above equation, we can deduce that $P(h|T)$ increases as $P(h)$ and $P(T|h)$ increases and also as $P(T)$ decreases. The simple explanation is that when there is more probability that T can occur independently of h then it is less probable that h can get support from T in its occurrence.

It is a common question in machine learning problems to find out the maximum probable hypothesis h from a set of hypotheses H ($h \in H$) given the observed training data T. This maximally probable hypothesis is called the **maximum a posteriori (MAP)** hypothesis. By using Bayes' theorem, we can identify the MAP hypothesis from the posterior probability of each candidate hypothesis:

$$\begin{aligned} h_{\text{MAP}} &= \text{argmax}_{h \in H} P(h|T) \\ &= \text{argmax}_{h \in H} \frac{P(T|h)P(h)}{P(T)} \end{aligned}$$

and as $P(T)$ is a constant independent of h, in this case, we can write

$$= \text{argmax}_{h \in H} P(T|h)P(h) \qquad (6.1)$$

6.3.3 Likelihood

In certain machine learning problems, we can further simplify equation 6.1 if every hypothesis in H has equal probable priori as $P(h_i) = P(h_j)$, and

then, we can determine $P(h|T)$ from the probability $P(T|h)$ only. Thus, $P(T|h)$ is called the likelihood of data T given h, and any hypothesis that maximizes $P(T|h)$ is called the maximum likelihood (ML) hypothesis, h_{ML}. See figure 6.1 and 6.2 for the conceptual and mathematical representation of Bayes theorem and the relationship of Prior, Posterior and Likelihood.

$$h_{ML} = \text{argmax}_{h \in H} P(T|h) \qquad (6.2)$$

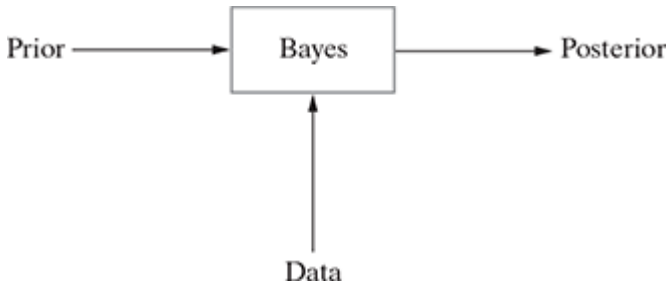

FIG. 6.1

Bayes' theorem

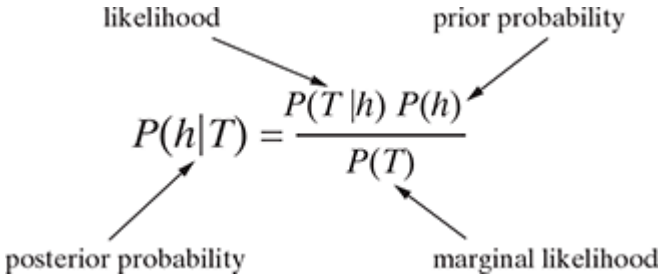

FIG. 6.2

Concept of prior, posterior, and likelihood

Points to Ponder:

Arriving at the refined probability of an event in the light of probability of a related event is a powerful concept and relates very closely with our day-to-day handling of events and using our knowledge to influence the decisions.

Example. Let us take the example of malignancy identification in a particular patient's tumour as an application for Bayes rule. We will calculate how the prior knowledge of the percentage of cancer cases in a sample population and probability of the test result being correct influence the probability outcome of the correct diagnosis. We have two alternative hypotheses: (1) a particular tumour is of malignant type and (2) a particular tumour is non-malignant type. The priori available are—1. only 0.5% of the population has this kind of tumour which is malignant, 2. the laboratory report has some amount of incorrectness as it could detect the malignancy was present only with 98% accuracy whereas could show the malignancy was not present correctly only in 97% of cases. This means the test predicted malignancy was present which actually was a false alarm in 2% of the cases, and also missed detecting the real malignant tumour in 3% of the cases.

Solution: Let us denote Malignant Tumour = MT, Positive Lab Test = PT, Negative Lab Test = NT

h_1 = the particular tumour is of malignant type = MT in our example

h_2 = the particular tumour is not malignant type = !MT in our example

$$P(\text{MT}) = 0.005 \quad P(\text{!MT}) = 0.995$$
$$P(\text{PT} \mid \text{MT}) = 0.98 \quad P(\text{PT} \mid \text{!MT}) = 0.02$$
$$P(\text{NT} \mid \text{!MT}) = 0.97 \quad P(\text{NT} \mid \text{MT}) = 0.03$$

So, for the new patient, if the laboratory test report shows positive result, let us see if we should declare this as the malignancy case or not:

$$
\begin{aligned}
P(h_1|\mathrm{PT}) &= \frac{P(\mathrm{PT}|h_1).P(h_1)}{P(\mathrm{PT})} \\
&= P(\mathrm{PT}|\mathrm{MT})P(\mathrm{MT}) \\
&= 0.98 \times 0.005 \\
&= 0.0049 \\
&= 0.49\%
\end{aligned}
$$

$$
\begin{aligned}
P(h_2|\mathrm{PT}) &= \frac{P(\mathrm{PT}|h_2).P(h_2)}{P(\mathrm{PT})} \\
&= P(\mathrm{PT}|!\mathrm{MT})P(!\mathrm{MT}) \\
&= 0.02 \times 0.995 \\
&= 0.0199 \\
&= 1.99\%
\end{aligned}
$$

As $P(h_2 \mid \mathrm{PT})$ is higher than $P(h_1 \mid \mathrm{PT})$, it is clear that the hypothesis h_2 has more probability of being true. So, hMAP = h2 = !MT.

This indicates that even if the posterior probability of malignancy is significantly higher than that of non-malignancy, the probability of this patient not having malignancy is still higher on the basis of the prior knowledge. Also, it should be noted that through Bayes' theorem, we identified the probability of one hypothesis being higher than the other hypothesis, and we did not completely accept or reject the hypothesis by this theorem. Furthermore, there is very high dependency on the availability of the prior data for successful application of Bayes' theorem.

6.4 BAYES' THEOREM AND CONCEPT LEARNING

One simplistic view of concept learning can be that if we feed the machine with the training data, then it can calculate the posterior probability of the hypotheses and outputs the most probable hypothesis. This is also called brute-force Bayesian learning algorithm, and it is also observed that consistency in providing the right probable hypothesis by this algorithm is very comparable to the other algorithms.

6.4.1 Brute-force Bayesian algorithm

We will now discuss how to use the MAP hypothesis output to design a simple learning algorithm called brute-force map learning algorithm. Let us assume that the learner considers a finite hypothesis space H in which the learner will try to learn some target concept $c{:}X \rightarrow \{0,1\}$ where X is the instance space corresponding to H. The sequence of training examples is $\{(x_1, t_1), (x_2, t_2),\ldots, (x_m, t_m)\}$, where x_i is the instance of X and t_i is the target concept of x_i defined as $t_i = c(x_i)$. Without impacting the efficiency of the algorithm, we can assume that the sequence of instances of x $\{x_1,\ldots, x_m\}$ is held fixed, and then, the sequence of target values becomes $T = \{t_1,\ldots, t_m\}$.

For calculating the highest posterior probability, we can use Bayes' theorem as discussed earlier in this chapter:

Calculate the posterior probability of each hypothesis h in H:

$$P(h|T) = \frac{P(T|h)P(h)}{P(T)}$$

Identify the $h_{\mathbf{MAP}}$ with the highest posterior probability

$$h_{\mathbf{MAP}} = \text{argmax}_{h \in H} P(h \mid T)$$

Please note that calculating the posterior probability for each hypothesis requires a very high volume of computation, and for a large volume of hypothesis space, this may be difficult to achieve.

Let us try to connect the concept learning problem with the problem of identifying the $h_{\mathbf{MAP}}$. On the basis of the probability distribution of $P(h)$ and $P(T|h)$, we can derive the prior knowledge of the learning task. There are few important assumptions to be made as follows:

1. The training data or target sequence T is noise free, which means that it is a direct function of X only (i.e. $t_i = c(x_i)$)
2. The concept c lies within the hypothesis space H
3. Each hypothesis is equally probable and independent of each other

On the basis of assumption 3, we can say that each hypothesis h within the space H has equal prior probability, and also because of assumption 2, we can say that these prior probabilities sum up to 1. So, we can write

$$P(h) = \frac{1}{|H|} \text{ for all } h \text{ within } H \tag{6.3}$$

$P(T|h)$ is the probability of observing the target values t_i in the fixed set of instances $\{x_i,..., x_m\}$ in the space where h holds true and describes the concept c correctly. Using assumption 1 mentioned above, we can say that if T is consistent with h, then the probability of data T given the hypothesis h is 1 and is 0 otherwise:

$$P(T|h) = \begin{cases} 1 & \text{if } t_i = h(x_i) \text{ for all } t_i \text{ within } T \\ 0 & \text{otherwise} \end{cases} \tag{6.4}$$

Using Bayes' theorem to identify the posterior probability

$$P(h|T) = \frac{P(T|h)P(h)}{P(T)} \tag{6.5}$$

For the cases when h is inconsistent with the training data T, using 6.5 we get

$$P(h|T) = \frac{0 \times P(h)}{P(T)} = 0, \text{ when } h \text{ is inconsistent with } T,$$

and when h is consistent with T

$$P(h|T) = \frac{1 \times \frac{1}{|H|}}{P(T)} = \frac{1}{|H|P(T)} \tag{6.6}$$

Now, if we define a subset of the hypothesis H which is consistent with T as H_D, then by using the total probability equation, we get

$$\begin{aligned} P(T) &= \sum_{h_i \in H_D} P(T|h_i)P(h_i) \\ &= \sum_{h_i \in H_D} 1.\frac{1}{|H|} + \sum_{h_i \notin H_D} 0.\frac{1}{|H|} \\ &= \sum_{h_i \in H_D} 1.\frac{1}{|H|} \\ &= \frac{|H_D|}{|H|} \end{aligned}$$

This makes 6.5 as

$$\begin{aligned} P(h|T) &= \frac{1}{|H|.\frac{|H_D|}{|H|}} \\ &= \frac{1}{|H_D|} \end{aligned}$$

So, with our set of assumptions about $P(h)$ and $P(T|h)$, we get the posterior probability $P(h|T)$ as

$$P(h|T) = \begin{cases} \frac{1}{|H_D|} & \text{if } h \text{ is consistent with } T \\ 0 & \text{otherwise} \end{cases} \tag{6.7}$$

where H_D is the number of hypotheses from the space H which are consistent with target data set T. The interpretation of this evaluation is that initially, each hypothesis has equal probability and, as we introduce the training data, the posterior probability of inconsistent hypotheses becomes zero and the total probability that sums up to 1 is distributed equally among the consistent hypotheses in the set. So, under this condi-

tion, each consistent hypothesis is a MAP hypothesis with posterior probability $\frac{1}{|H_D|}$.

6.4.2 Concept of consistent learners

From the above discussion, we understand the behaviour of the general class of learner whom we call as consistent learners. So, the group of learners who commit zero error over the training data and output the hypothesis are called *consistent learners*. If the training data is noise free and deterministic (i.e. $P(D|h) = 1$ if D and h are consistent and 0 otherwise) and if there is uniform prior probability distribution over H (so, $P(h_m) = P(h_n)$ for all m, n), then every consistent learner outputs the MAP hypothesis. An important application of this conclusion is that Bayes' theorem can characterize the behaviour of learning algorithms even when the algorithm does not explicitly manipulate the probability. As it can help to identify the optimal distributions of $P(h)$ and $P(T|h)$ under which the algorithm outputs the MAP hypothesis, the knowledge can be used to characterize the assumptions under which the algorithms behave optimally.

Though we discussed in this section a special case of Bayesian output which corresponds to the noise-free training data and deterministic predictions of hypotheses where $P(T|h)$ takes on value of either 1 or 0, the theorem can be used with the same effectiveness for noisy training data and additional assumptions about the probability distribution governing the noise.

6.4.3 Bayes optimal classifier

In this section, we will discuss the use of the MAP hypothesis to answer the question what is the most probable classification of the new instance given the training data. To illustrate the concept, let us assume three hypotheses h_1, h_2, and h_3 in the hypothesis space H. Let the posterior proba-

bility of these hypotheses be 0.4, 0.3, and 0.3, respectively. There is a new instance x, which is classified as true by h_1, but false by h_2 and h_3.

Then the most probable classification of the new instance (x) can be obtained by combining the predictions of all hypotheses weighed by their corresponding posterior probabilities. By denoting the possible classification of the new instance as c_i from the set C, the probability $P(c_i|T)$ that the correct classification for the new instance is c_i is

The optimal classification is for which $P(c_i|T)$ is maximum is

Points to Ponder:

The approach in the Bayes optimal classifier is to calculate the most probable classification of each new instance on the basis of the combined predictions of all alternative hypotheses, weighted by their posterior probabilities.

So, extending the above example,

The set of possible outcomes for the new instance x is within the set C = {True, False} and

$$P(h_1 \mid T) = 0.4,\ P(\text{False} \mid h_1) = 0,\ P(\text{True} \mid h_1) = 1$$
$$P(h_2 \mid T) = 0.3,\ P(\text{False} \mid h_2) = 1,\ P(\text{True} \mid h_2) = 0$$
$$P(h_3 \mid T) = 0.3,\ P(\text{False} \mid h_3) = 1,\ P(\text{True} \mid h_3) = 0$$

Then,

and

This method maximizes the probability that the new instance is classified correctly when the available training data, hypothesis space and the prior probabilities of the hypotheses are known. This is thus also called Bayes optimal classifier.

6.4.4 Naïve Bayes classifier

Naïve Bayes is a simple technique for building classifiers: models that assign class labels to problem instances. The basic idea of Bayes rule is that the outcome of a hypothesis can be predicted on the basis of some evidence (E) that can be observed.

From Bayes rule, we observed that

1. A prior probability of hypothesis h or $P(h)$: This is the probability of an event or hypothesis before the evidence is observed.
2. A posterior probability of h or $P(h|D)$: This is the probability of an event after the evidence is observed within the population D.

Posterior Probability is of the format 'What is the probability that a particular object belongs to class i given its observed feature values?'

For example, a person has height and weight of 182 cm and 68 kg, respectively. What is the probability that this person belongs to the class 'basketball player'? This can be predicted using the Naïve Bayes classifier. This is known as probabilistic classifications.

In machine learning, a probabilistic classifier is a classifier that can be foreseen, given a perception or information (input), a likelihood calculation over a set of classes, instead of just yielding (outputting) the most likely class that the perception (observation) should belong to. Parameter estimation for Naïve Bayes models uses the method of ML.

Bayes' theorem is used when new information can be used to revise previously determined probabilities. Depending on the particular nature of the probability model, Naïve Bayes classifiers can be trained very professionally in a supervised learning setting.

Let us see the basis of deriving the principles of Naïve Bayes classifiers. We take a learning task where each instance x has some attributes and the target function ($f(x)$) can take any value from the finite set of classification values C. We also have a set of training examples for target function, and the set of attributes $\{a_1, a_2,..., a_n\}$ for the new instance are known to us. Our task is to predict the classification of the new instance.

According to the approach in Bayes' theorem, the classification of the new instance is performed by assigning the most probable target classification $C_{\mathbf{MAP}}$ on the basis of the attribute values of the new instance $\{a_1, a_2,..., a_n\}$. So,

which can be rewritten using Bayes' theorem as

As combined probability of the attributes defining the new instance fully is always 1

So, to get the most probable classifier, we have to evaluate the two terms $P(a_1, a_2, \text{c}, a_n|c_i)$ and $P(ci)$. In a practical scenario, it is possible to calculate $P(ci)$ by calculating the frequency of each target value c_i in the training data set. But the $P(a_1, a_2, \text{c}, a_n|c_i)$ cannot be estimated easily and needs a very high effort of calculation. The reason is that the number of these terms is equal to the product of number of possible instances and the number of possible target values, and thus, each instance in the instance space needs to be visited many times to arrive at the estimate of the occurrence. Thus, the Naïve Bayes classifier makes a simple assumption that the attribute values are conditionally independent of each other for the target value. So, applying this simplification, we can now say that for a target value of an instance, the probability of observing the combination $a_1,a_2,\ldots, a_n$ is the product of probabilities of individual attributes $P(a_i|c_j)$.

Then, from equation 6.7, we get the approach for the Naïve Bayes classifier as

Here, we will be able to compute $P(a_i|c_j)$ as we have to calculate this only for the number of distinct attributes values (a_i) times the number of distinct target values (c_j), which is much smaller set than the product of

both the sets. The most important reason for the popularity of the Naïve Bayes classifier approach is that it is not required to search the whole hypothesis space for this algorithm, but rather we can arrive at the target classifier by simply counting the frequencies of various data combinations within the training example.

To summarize, a Naïve Bayes classifier is a primary probabilistic classifier based on a view of applying Bayes' theorem (from Bayesian inference with strong naive) independence assumptions. The prior probabilities in Bayes' theorem that are changed with the help of newly available information are classified as posterior probabilities.

A key benefit of the naive Bayes classifier is that it requires only a little bit of training information (data) to gauge the parameters (mean and differences of the variables) essential for the classification (arrangement). In the Naïve Bayes classifier, independent variables are always assumed, and only the changes (variances) of the factors/variables for each class should be determined and not the whole covariance matrix. Because of the rather naïve assumption that all features of the dataset are equally important and independent, this is called Naïve Bayes classifier.

Naïve Bayes classifiers are direct linear classifiers that are known for being the straightforward, yet extremely proficient result. The modified version of Naïve Bayes classifier originates from the assumption that information collection (data set) is commonly autonomous (mutually independent). In most of the practical scenarios, the 'independence' assumption is regularly violated. However, Naïve Bayes classifiers still tend to perform exceptionally well.

Some of the key strengths and weaknesses of Naïve Bayes classifiers are described in Table 6.1.

Table 6.1 *Strengths and Weaknesses of Bayes Classifiers*

Example. Let us assume that we want to predict the outcome of a football world cup match on the basis of the past performance data of the playing teams. We have training data available (refer Fig. 6.3) for actual match outcome, while four parameters are considered – Weather Condition (Rainy, Overcast, or Sunny), how many matches won were by this team out of the last three matches (one match, two matches, or three matches), Humidity Condition (High or Normal), and whether they won the toss (True or False). Using Naïve Bayesian, you need to classify the conditions when this team wins and then predict the probability of this team winning a particular match when Weather Conditions = Rainy, they won two of the last three matches, Humidity = Normal and they won the toss in the particular match.

FIG. 6.3 Training data for the Naïve Bayesian method

6.4.4.1 Naïve Bayes classifier steps

Step 1: First construct a frequency table. A frequency table is drawn for each attribute against the target outcome. For example, in Figure 6.3, the various attributes are (1) Weather Condition, (2) How many matches won by this team in last three matches, (3) Humidity Condition, and (4) whether they won the toss and the target outcome is will they win the match or not?

Step 2: Identify the cumulative probability for 'Won match = Yes' and the probability for 'Won match = No' on the basis of all the attributes. Otherwise, simply multiply probabilities of all favourable conditions to derive 'YES' condition. Multiply probabilities of all non-favourable conditions to derive 'No' condition.

Step 3: Calculate probability through normalization by applying the below formula

P(Yes) will give the overall probability of favourable condition in the given scenario.

P(No) will give the overall probability of non-favourable condition in the given scenario.

Solving the above problem with Naive Bayes

Step 1: Construct a frequency table. The posterior probability can be easily derived by constructing a frequency table for each attribute against the target. For example, frequency of Weather Condition variable with values 'Sunny' when the target value Won match is 'Yes', is, 3/(3+4+2) = 3/9.

Figure 6.4 shows the frequency table thus constructed.

Step 2:

To predict whether the team will win for given weather conditions (a_1) = Rainy, Wins in last three matches (a_2) = 2 wins, Humidity (a_3) = Normal and Win toss (a_4) = True, we need to choose 'Yes' from the above table for the given conditions.

From Bayes' theorem, we get

This equation becomes much easier to resolve if we recall that Naïve Bayes classifier assumes independence among events. This is specifically

true for class-conditional independence, which means that the events are independent so long as they are conditioned on the same class value. Also, we know that if the events are independent, then the probability rule says, $P(A \cap B) = P(A)\ P(B)$, which helps in simplifying the above equation significantly as

$$P(\text{Win match} \mid a_1 \cap a_2 \cap a_3 \cap a_4)$$

This should be compared with

$$P(!\text{Win match} \mid a_1 \cap a_2 \cap a_3 \cap a_4)$$

FIG. 6.4 Construct frequency table

Step 3: by normalizing the above two probabilities, we can ensure that the sum of these two probabilities is 1.

Conclusion: This shows that there is 58% probability that the team will win if the above conditions become true for that particular day. Thus, Naïve Bayes classifier provides a simple yet powerful way to consider the influence of multiple attributes on the target outcome and refine the uncertainty of the event on the basis of the prior knowledge because it is able to simplify the calculation through independence assumption.

6.4.5 Applications of Naïve Bayes classifier

Text classification: Naïve Bayes classifier is among the most successful known algorithms for learning to classify text documents. It classifies the document where the probability of classifying the text is more. It uses the above algorithm to check the permutation and combination of the probability of classifying a document under a particular 'Title'. It has various applications in document categorization, language detection, and sentiment detection, which are very useful for traditional retailers, e-retailors, and other businesses on judging the sentiments of their clients on the basis of keywords in feedback forms, social media comments, etc.

Spam filtering: Spam filtering is the best known use of Naïve Bayesian text classification. Presently, almost all the email providers have this as a built-in functionality, which makes use of a Naïve Bayes classifier to iden-

tify spam email on the basis of certain conditions and also the probability of classifying an email as 'Spam'. Naïve Bayesian spam sifting has turned into a mainstream mechanism to recognize illegitimate a spam email from an honest-to-goodness email (sometimes called 'ham'). Users can also install separate email filtering programmes. Server-side email filters such as DSPAM, Spam Assassin, Spam Bayes, and ASSP make use of Bayesian spam filtering techniques, and the functionality is sometimes embedded within the mail server software itself.

Hybrid Recommender System: It uses Naïve Bayes classifier and collaborative filtering. Recommender systems (used by e-retailors like eBay, Alibaba, Target, Flipkart, etc.) apply machine learning and data mining techniques for filtering unseen information and can predict whether a user would like a given resource. For example, when we log in to these retailer websites, on the basis of the usage of texts used by the login and the historical data of purchase, it automatically recommends the product for the particular login persona. One of the algorithms is combining a Naïve Bayes classification approach with collaborative filtering, and experimental results show that this algorithm provides better performance regarding accuracy and coverage than other algorithms.

Online Sentiment Analysis: The online applications use supervised machine learning (Naïve Bayes) and useful computing. In the case of sentiment analysis, let us assume there are three sentiments such as nice, nasty, or neutral, and Naïve Bayes classifier is used to distinguish between them. Simple emotion modelling combines a statistically based classifier with a dynamical model. The Naïve Bayes classifier employs 'single words' and 'word pairs' like features and determines the sentiments of the users. It allocates user utterances into nice, nasty, and neutral classes, labelled as +1, −1, and 0, respectively. This binary output drives a simple first-order dynamical system, whose emotional state represents the simulated emotional state of the experiment's personification.

6.4.6 Handling Continuous Numeric Features in Naïve Bayes Classifier

In the above example, we saw that the Naïve Bayes classifier model uses a frequency table of the training data for its calculation. Thus, each attribute data should be categorical in nature so that the combination of class and feature values can be created. But this is not possible in the case of continuous numeric data as it does not have the categories of data.

The workaround that is applied in these cases is discretizing the continuous data on the basis of some data range. This is also called binning as the individual categories are termed as bins. For example, let us assume we want to market a certain credit card to all the customers who are visiting a particular bank. We have to classify the persons who are visiting a bank as either interested candidate for taking a new card or non-interested candidate for a new card, and on the basis of this classification, the representative will approach the customer for sale. In this case, the customers visit the bank continuously during banking hours and have different values for the attributes we want to evaluate before classifying them into the interested/non-interested categories.

If we plot the number of customers visiting the bank during the 8 hours of banking time, the distribution graph will be a continuous graph. But if we introduce a logic to categorize the customers according to their time of entering the bank, then we will be able to put the customers in ‘bins’ or buckets for our analysis. We can then try to assess what time range is best suited for targeting the customers who will have interest in the new credit card. The bins created by categorizing the customers by their time of entry looks like Figure 6.5.

This creates eight natural bins for us (or we may change the number of bins by changing our categorizing criteria), which can now be used for Bayes analysis.

FIG. 6.5 The distribution of bins based on the time of entry of customers in the bank

6.5 BAYESIAN BELIEF NETWORK

We must have noted that a significant assumption in the Naïve Bayes classifier was that the attribute values a_1, a_2,..., a_n are conditionally independent for a target value. The Naïve Bayes classifier generates optimal output when this condition is met. Though this assumption significantly reduces the complexity of computation, in many practical scenarios, this requirement of conditional independence becomes a difficult constraint for the application of this algorithm. So, in this section, we will discuss the approach of Bayesian Belief network, which assumes that within the set of attributes, the probability distribution can have conditional probability relationship as well as conditional independence assumptions. This is different from the Naïve Bayes assumption of conditional independence of all the attributes as the belief network provides the flexibility of declaring a subset of the attributes as conditionally dependent while leaving rest of the attributes to hold the assumptions of conditional independence. The prior knowledge or belief about the influence of one attribute over the

other is handled through joint probabilities as discussed later in this section.

Let us refresh our mind on the concept of conditional probability. If an uncertain event A is conditional on a knowledge or belief K, then the degree of belief in A with the assumption that K is known is expressed as $P(A|K)$. Traditionally, conditional probability is expressed by joint probability as follows:

Rearranging (6.9), we get the product rule

$$P(A, K) = P(A|K)P(K)$$

This can be extended for three variables or attributes as

$$P(A, K, C) = P(A|K, C)P(K, C) = P(A|K, C)P(K|C)P(C)$$

For a set of n attributes, the generalized form of the product rule becomes

$$P(A_1, A_2, \ldots, A_n) = P(A_1|A_2, \ldots, A_n)P(A_2|A_3, \ldots, A_n)P(A_n - 1|A_n)P(A_n) \quad (6.10)$$

This generalized version of the product rule is called the Chain Rule.

FIG. 6.6 Chain rule

Let us understand the chain rule by using the diagram in Figure 6.6. From the joint probability formula 6.10, we can write

$$P(A, B, C, D, E) = P(A \mid B, C, D, E)P(B \mid C, D, E)P(C \mid D, E)P(D \mid E)P(E)$$

But from Figure 6.6, it is evident that E is not related to C and D, which means that the probabilities of variables C and D are not influenced by E and vice versa. Similarly, A is directly influenced only by B. By applying this knowledge of independence, we can simplify the above equation as

$$P(A, B, C, D, E) = P(A \mid B)P(B \mid C, D)P(C \mid D)P(D)P(E)$$

Let us discuss this concept of independence and conditional independence in detail in the next section.

6.5.1 Independence and conditional independence

We represent the conditional probability of A with knowledge of K as $P(A \mid K)$. The variables A and K are said to be independent if $P(A \mid K) = P(A)$,

which means that there is no influence of K on the uncertainty of A. Similarly, the joint probability can be written as $P(A,K) = P(A)P(K)$.

Extending this concept, the variables A and K are said to be conditionally independent given C if $P(A \mid C) = P(A \mid K, C)$.

This concept of conditional independence can also be extended to a set of attributes. We can say that the set of variables $A_1, A_2,..., A_n$ is conditionally independent of the set of variables $B_1, B_2,..., B_m$ given the set of variables $C_1, C_2,..., C_l$ if

$$P(A_1, A_2,..., A_n \mid B_1, B_2,..., Bm, C_1, C_2,..., C_1) = P(A_1, A_2,..., A_n \mid C_1, C_2,..., C_l)$$

If we compare this definition with our assumption in the Naïve Bayes classifier, we see that the Naïve Bayes classifier assumes that the instance attribute A_1 is conditionally independent of the instance attribute A_2, given the target value V, which can be written using the general product rule and application of conditional independence formula as

$$P(A_1, A_2 \mid V) = P(A_1 \mid A_2, V)P(A_2, V) = P(A_1, V)P(A_2, V)$$

A Bayesian Belief network describes the joint probability distribution of a set of attributes in their joint space. In Figure 6.7, a Bayesian Belief network is presented. The diagram consists of nodes and arcs. The nodes represent the discrete or continuous variables for which we are interested to calculate the conditional probabilities. The arc represents the causal relationship of the variables.

The two important information points we get from this network graph are used for the determining the joint probability of the variables. First,

the arcs assert that the node variables are conditionally independent of its non-descendants in the network given its immediate predecessors in the network. If two variables *A* and *B* are connected through a directed path, then *B* is called the descendent of *A*. Second, the conditional probability table for each variable provides the probability distribution of that variable given the values of its immediate predecessors. We can use Bayesian probability to calculate different behaviours of the variables in Figure 6.7.

FIG. 6.7 Bayesian belief network

1. The unconditional probability that Tim is late to class –

   P(Tim is late to class)

   = P(Tim late|Rain Today)P(Rain Today) + P(Tim late|No Rain Today)*P(No Rain Today)

   = (0.8 × 0.1) × (0.1 × 0.9)

   = 0.17

   From this unconditional probability, the most important use of the Bayesian Belief network is to find out the revised probability on the basis of the prior knowledge. If we assume that there was rain today,

then the probability table can quickly provide us the information about the probability of Paul being late to class or the probability of Tim being late to class from the probability distribution table itself. But if we do not know whether there was rain today or not, but we only know that Tim is late to class today, then we can arrive at the following probabilities –

2. The revised probability that there was rain today –

3. The revised probability that Paul will be late to class today –

   P(Paul late to class today)

   = P(Paul late | Rain today)P(Rain today) + P(Paul late | No rain today)P(No rain today)

   = (0.6 × 0.47) + (0.5 × (1-0.47))

   = 0.55

Here, we used the concept of hard evidence and soft evidence. Hard evidence (instantiation) of a node is evidence that the state of the variable is definitely as a particular value. In our above example, we had hard evidence that 'Tim is late to class'. If a particular node is instantiated, then it will block propagation of evidence further down to its child nodes. Soft evidence for a node is the evidence that provides the prior probability values for the node. The node 'Paul is late to class' is soft evidenced with the prior knowledge that 'Tim is late to class'.

Note

There can be two main scenarios faced in the Bayesian Belief network learning problem. First, the network structure might be available in advance or can be inferred from the training data. Second, all the network variables either are directly observable in each training example or some

of the variables may be unobservable. Learning the conditional probability tables is a straightforward problem when the network structure is given in advance and the variables are fully observable in the training examples. But in the case the network structure is available and only *some* of the variable values are observable in the training data, then the learning problem is more difficult. This is topic of much research, and some of the advanced topics for identifying the node values include algorithms such as Gradient Ascent Training and the EM algorithm.

The Bayesian Belief network can represent much more complex scenarios with dependence and independence concepts. There are three types of connections possible in a Bayesian Belief network.

Diverging Connection: In this type of connection, the evidence can be transmitted between two child nodes of the same parent provided that the parent is not instantiated. In Figure 6.7, we already saw the behaviour of diverging connection.

Serial Connection: In this type of connection, any evidence entered at the beginning of the connection can be transmitted through the directed path provided that no intermediate node on the path is instantiated (see Fig. 6.8 for illustration).

FIG. 6.8 Serial connection

Converging Connection: In this type of connection, the evidence can only be transmitted between two parents when the child (converging) node has received some evidence and that evidence can be soft or hard (see Fig. 6.9 for illustration).

FIG. 6.9 Convergent connection

As discussed above, by using the Bayesian network, we would like to infer the value of a target variable on the basis of the observed values of some other variables. Please note that it will not be possible to infer a single value in the case of random variables we are dealing with, but our intention is to infer the probability distribution of the target variable given the observed values of other variables. In general, the Bayesian network

can be used to compute the probability distribution of any subset of node variables given the values or distribution of the remaining variables.

6.5.2 Use of the Bayesian Belief network in machine learning

We have seen that the Bayesian network creates a complete model for the variables and their relationships and thus can be used to answer probabilistic queries about them. A common use of the network is to find out the updated knowledge about the state of a subset of variables, while the state of the other subset (known as the evidence variables) is observed. This concept, often known as probabilistic inference process of computing the posterior distribution of variables, given some evidences, provides a universal sufficient statistic for applications related to detections. Thus if one wants to choose the values for a subset of variables in order to minimize some expected loss functions or decision errors, then this method is quiet effective. In other words, the Bayesian network is a mechanism for automatically applying Bayes' theorem to complex problems. Bayesian networks are used for modelling beliefs in domains like computational biology and bioinformatics such as protein structure and gene regulatory networks, medicines, forensics, document classification, information retrieval, image processing, decision support systems, sports betting and gaming, property market analysis and various other fields.

6.6 SUMMARY

- Bayesian methods introduced a basis for probabilistic learning methods that consider the knowledge about the **prior** probabilities of alternative hypotheses and the probability of **likelihood** or observing various training data given the hypothesis. It then assigns a **posterior** probability to each candidate hypothesis on the basis of the assumed priors and the observed data.
- The **MAP** hypothesis is the most probable hypothesis given the data. As no other hypothesis is more likely, this is also the optimal hypothesis.

- The Bayes optimal classifier calculates the most probable classification of each new instance by combining the predictions of all alternative hypotheses, weighted by their posterior probabilities.
- **Naïve Bayes classifier** makes the naïve assumption that the attribute values are conditionally independent given the classification of the instance. This simplifying assumption considerably reduces the calculation overhead without losing the effectiveness of the outcome. With this assumption in effect, the Naïve Bayes classifier outputs the MAP classification.
- The Naïve Bayes classifier has been found to be useful in many practical applications and is considered as one of the powerful learning methods. Even when the assumption of conditional independence is not met, the Naïve Bayes classifier is quite effective in providing a standard for the other learning methods.
- **Bayesian Belief network** provides the mechanism to handle more practical scenario of considering the conditional independence of a subset of variables while considering the joint probability distribution of the remaining variables given the observation.

SAMPLE QUESTIONS

MULTIPLE CHOICE QUESTIONS (1 MARK EACH)

1. Three companies X, Y, and Z supply 40%, 45%, and 15% of the uniforms to a school. Past experience shows that 2%, 3%, and 4% of the uniforms supplied by these companies are defective. If a uniform was found to be defective, what is the probability that the uniform was supplied by Company X?

    1.

    2.

    3. 16/55

    4.

2. A box of apples contains 10 apples, of which 6 are defective. If 3 of the apples are removed from the box in succession without replacement, what is the probability that all the 3 apples are defective?
    1. (6*5*4)/ (10*10*10)
    2. (6*5*4)/ (10*9*8)
    3. (3*3*3)/ (10*10*10)
    4. (3*2*1)/ (10*9*8)
3. Two boxes containing chocolates are placed on a table. The boxes are labelled B_1 and B_2. Box B_1 contains 6 dark chocolates and 5 white chocolates. Box B_2 contains 3 dark chocolates and 8 orange chocolates. The boxes are arranged so that the probability of selecting box B_1 is $^1/_3$ and the probability of selecting box B_2 is $^2/_3$. Sneha is blindfolded and asked to select a chocolate. She will win Rs. 10,000 if she selects a dark chocolate. What is the probability that Sneha will win Rs. 10,000 (that is, she will select a dark chocolate)?
    1.
    2.
    3.
    4.
4. Two boxes containing chocolates are placed on a table. The boxes are labelled B_1 and B_2. Box B_1 contains 6 Cadbury chocolates and 5 Amul chocolates. Box B_2 contains 3 Cadbury chocolates and 8 Nestle chocolates. The boxes are arranged so that the probability of selecting box B_1 is $^1/_3$ and the probability of selecting box B_2 is $^2/_3$. Sneha is blindfolded and asked to select a chocolate. She will win Rs. 10,000 if she selects a Cadbury chocolate. If she win Rs 10,000, what is the probability that she selected a Cadbury chocolate from the first box?
    1.
    2.

3.

4.

5. In a certain basketball club, there are 4% of male players who are over 6 feet tall and 1% of female players who are over 6 feet tall. The ratio of male to female players in the total player population is male:female = 2:3. A player is selected at random from among all those who are over 6 feet tall. What is the probability that the player is a female?
   1. 3/11
   2. 2/5
   3. 2/11
   4. 1/11
6. The probability that a particular hypothesis holds for a data set based on the Prior is called
   1. Independent probabilities
   2. Posterior probabilities
   3. Interior probabilities
   4. Dependent probabilities
7. One main disadvantage of Bayesian classifiers is that they utilize all available parameters to subtly change the predictions.
   1. True
   2. False
8. In a bolt factory, machines A1, A2, and A3 manufacture respectively 25%, 35%, and 40% of the total output. Of these 5%, 4%, and 2% are defective bolts. A bolt is drawn at random from the product and is found to be defective. What is the probability that it was manufactured by machine A2?
   1. 0.0952
   2. 0.452
   3. 0.952
   4. 0.125
9. Bayesian methods can perform better than the other methods while validating the hypotheses that make probabilistic predictions.
   1. True

    2. False
10. Naïve Bayes classifier makes the naïve assumption that the attribute values are conditionally dependent given the classification of the instance.
    1. True
    2. False

SHORT ANSWER-TYPE QUESTIONS (5 MARKS EACH)

1. What is prior probability? Give an example.
2. What is posterior probability? Give an example.
3. What is likelihood probability? Give an example.
4. What is Naïve Bayes classifier? Why is it named so?
5. What is optimal Bayes classifier?
6. Write any two features of Bayesian learning methods.
7. Define the concept of consistent learners.
8. Write any two strengths of Bayes classifier.
9. Write any two weaknesses of Bayes classifier.
10. Explain how Naïve Bayes classifier is used for
    1. Text classification
    2. Spam filtering
    3. Market sentiment analysis

LONG ANSWER-TYPE QUESTIONS (10 MARKS EACH)

1. Explain the concept of Prior, Posterior, and Likelihood with an example.
2. How Bayes' theorem supports the concept learning principle?
3. Explain Naïve Bayes classifier with an example of its use in practical life.
4. Is it possible to use Naïve Bayes classifier for continuous numeric data? If so, how?
5. What are Bayesian Belief networks? Where are they used? Can they solve all types of problems?

6. In an airport security checking system, the passengers are checked to find out any intruder. Let I with $i \in \{0, 1\}$ be the random variable which indicates whether somebody is an intruder ($i = 1$) or not ($i = 0$) and A with $a \in \{0, 1\}$ be the variable indicating alarm. An alarm will be raised if an intruder is identified with probability $P(A = 1 \mid I = 1) = 0.98$ and a non-intruder with probability $P(A = 1 \mid I = 0) = 0.001$, which implies the error factor. In the population of passengers, the probability of someone is intruder is $P(I = 1) = 0.00001$. What is the probability that an alarm is raised when a person actually is an intruder?
7. An antibiotic resistance test (random variable T) has 1% false positives (i.e. 1% of those not resistance to an antibiotic show positive result in the test) and 5% false negatives (i.e. 5% of those actually resistant to an antibiotic test negative). Let us assume that 2% of those tested are resistant to antibiotics. Determine the probability that somebody who tests positive is actually resistant (random variable D).
8. For preparation of the exam, a student knows that one question is to be solved in the exam which is either of types A, B, or C. The probabilities of A, B, or C appearing in the exam are 30%, 20%, and 50% respectively. During the preparation, the student solved 9 of 10 problems of type A, 2 of 10 problems of type B, and 6 of 10 problems of type C.
    1. What is the probability that the student will solve the problem of the exam?
    2. Given that the student solved the problem, what is the probability that it was of type A?
9. A CCTV is installed in a bank to monitor the incoming customers and take a photograph. Though there are continuous flows of customers, we create bins of timeframe of 5 min each. In each time frame of 5 min, there may be a customer moving into the bank with 5% probability or there is no customer (again, for simplicity, we assume that either there is 1 customer or none, not the case of multiple customers). If there is a customer, it will be detected by the CCTV with a probability of 99%. If there is no customer, the camera will take a false photograph by detecting other thing's movement with a probability of 10%.
    1. How many customers enter the bank on average per day (10 hours)?

2. How many false photographs (there is a photograph taken even though there is no customer) and how many missed photographs (there is no photograph even though there is a customer) are there on average per day?
3. If there is a photograph, what is the probability that there is indeed a customer?

10. Draw the Bayesian Belief network to represent the conditional independence assumptions of the Naïve Bayes classifier for the match winning prediction problem of Section 6.4.4. Construct the conditional probability table associated with the node Won Toss.