hence it exhibits the same initial conditions every time the model is run and is likely to get a single possible outcome as the solution.

Bayesian learning differs from probabilistic learning as it uses subjective probabilities (i.e., probability that is based on an individual's belief or interpretation about the outcome of an event and it can change over time) to infer parameters of a model. Two practical learning algorithms called Naïve Bayes learning and Bayesian Belief Network (BBN) form the major part of Bayesian learning. These algorithms use prior probabilities and apply Bayes rule to infer useful information. Bayesian Belief Networks (BBN) is explained in detail in Chapter 9.

Scan for information on *'Probability Theory'* and for *'Additional Examples'*

8.2 FUNDAMENTALS OF BAYES THEOREM

Naïve Bayes Model relies on Bayes theorem that works on the principle of three kinds of probabilities called prior probability, likelihood probability, and posterior probability.

Prior Probability

It is the general probability of an uncertain event before an observation is seen or some evidence is collected. It is the initial probability that is believed before any new information is collected.

Likelihood Probability

Likelihood probability is the relative probability of the observation occurring for each class or the sampling density for the evidence given the hypothesis. It is stated as P (Evidence | Hypothesis), which denotes the likeliness of the occurrence of the evidence given the parameters.

Posterior Probability

It is the updated or revised probability of an event taking into account the observations from the training data. P (Hypothesis | Evidence) is the posterior distribution representing the belief about the hypothesis, given the evidence from the training data. Therefore,

Posterior probability = prior probability + new evidence

8.3 CLASSIFICATION USING BAYES MODEL

Naïve Bayes Classification models work on the principle of Bayes theorem. Bayes' rule is a mathematical formula used to determine the posterior probability, given prior probabilities of events. Generally, Bayes theorem is used to select the most probable hypothesis from data, considering both prior knowledge and posterior distributions. It is based on the calculation of the posterior probability and is stated as:

P (Hypothesis h | Evidence E)

where, Hypothesis h is the target class to be classified and Evidence E is the given test instance.

P (Hypothesis h | Evidence E) is calculated from the prior probability P (Hypothesis h), the likelihood probability P (Evidence E | Hypothesis h) and the marginal probability P (Evidence E). It can be written as:

$$P(\text{Hypothesis } h \mid \text{Evidence } E) = \frac{P(\text{Evidence } E \mid \text{Hypothesis } h)\, P(\text{Hypothesis } h)}{P(\text{Evidence } E)} \tag{8.1}$$

where, P (Hypothesis h) is the prior probability of the hypothesis h without observing the training data or considering any evidence. It denotes the prior belief or the initial probability that the hypothesis h is correct. P (Evidence E) is the prior probability of the evidence E from the training dataset without any knowledge of which hypothesis holds. It is also called the marginal probability.

P (Evidence E | Hypothesis h) is the prior probability of Evidence E given Hypothesis h. It is the likelihood probability of the Evidence E after observing the training data that the hypothesis h is correct. P (Hypothesis h | Evidence E) is the posterior probability of Hypothesis h given Evidence E. It is the probability of the hypothesis h after observing the training data that the evidence E is correct. In other words, by the equation of Bayes Eq. (8.1), one can observe that:

Posterior Probability α Prior Probability × Likelihood Probability

Bayes theorem helps in calculating the posterior probability for a number of hypotheses, from which the hypothesis with the highest probability can be selected.

This selection of the most probable hypothesis from a set of hypotheses is formally defined as Maximum A Posteriori (MAP) Hypothesis.

Maximum A Posteriori (MAP) Hypothesis, h_{MAP}

Given a set of candidate hypotheses, the hypothesis which has the maximum value is considered as the *maximum probable hypothesis* or *most probable hypothesis*. This most probable hypothesis is called the Maximum A Posteriori Hypothesis h_{MAP}. Bayes theorem Eq. (8.1) can be used to find the h_{MAP}

$$\begin{aligned} h_{MAP} &= \max_{h \in H} P(Hypothesish \mid Evidence\ E) \\ &= \max_{h \in H} \frac{P(Evidence\ E \mid Hypothesis\ h) P(Hypothesis\ h)}{P(Evidence\ E)} \\ &= \max_{h \in H} P(Evidence\ E \mid Hypothesis\ h) P(Hypothesis\ h) \end{aligned} \tag{8.2}$$

Maximum Likelihood (ML) Hypothesis, h_{ML}

Given a set of candidate hypotheses, if every hypothesis is equally probable, only P (E | h) is used to find the *most probable hypothesis*. The hypothesis that gives the maximum likelihood for P (E | h) is called the Maximum Likelihood (ML) Hypothesis, h_{ML}.

$$h_{ML} = \max_{h \in H} P(Evidence\ E \mid Hypothesis\ h) \tag{8.3}$$

Correctness of Bayes Theorem

Consider two events A and B in a sample space S.

A T F T T F T T F

B F T T F T F T F

P (A) = 5/8

P (B) = 4/8

$P(A \mid B) = 2/4$

$P(B \mid A) = 2/5$

$P(A \mid B) = P(B \mid A)\, P(A) / P(B) = = 2/4$

$P(B \mid A) = P(A \mid B)\, P(B) / P(A) = = 2/5$

Let us consider a numerical example to illustrate the use of Bayes theorem now:

Example 8.1: Consider a boy who has a volleyball tournament on the next day, but today he feels sick. It is unusual that there is only a 40% chance he would fall sick since he is a healthy boy. Now, Find the probability of the boy participating in the tournament. The boy is very much interested in volley ball, so there is a 90% probability that he would participate in tournaments and 20% that he will fall sick given that he participates in the tournament.

Solution: P (Boy participating in the tournament) = 90%

P (He is sick | Boy participating in the tournament) = 20%

P (He is Sick) = 40%

The probability of the boy participating in the tournament given that he is sick is:

P (Boy participating in the tournament | He is sick) = P (Boy participating in the tournament) $\times P$ (He is sick | Boy participating in the tournament)/P (He is Sick)

P (Boy participating in the tournament | He is sick) = $(0.9 \times 0.2)/0.4$

$= 0.45$

Hence, 45% is the probability that the boy will participate in the tournament given that he is sick.

One related concept of Bayes theorem is the principle of Minimum Description Length (MDL). The minimum description length (MDL) principle is yet another powerful method like Occam's razor principle to perform inductive inference. It states that the best and most probable hypothesis is chosen for a set of observed data or the one with the minimum description. Recall from Eq. (8.2) Maximum A Posteriori (MAP) Hypothesis, h_{MAP} which says that given a set of candidate hypotheses, the hypothesis which has the maximum value is considered as the *maximum probable hypothesis* or *most probable hypothesis*. Naïve Bayes algorithm uses the Bayes theorem and applies this MDL principle to find the best hypothesis for a given problem. Let us clearly understand how this algorithm works in the following Section 8.3.1.

8.3.1 NAÏVE BAYES ALGORITHM

It is a supervised binary class or multi class classification algorithm that works on the principle of Bayes theorem. There is a family of Naïve Bayes classifiers based on a common principle. These algorithms classify for datasets whose features are independent and each feature is assumed to be given equal weightage. It particularly works for a large dataset and is very fast. It is one of the most effective and simple classification algorithms. This algorithm considers all features to be independent of each other even though they are individually dependent on the classified object. Each of the features contributes a probability value independently during classification and hence this algorithm is called as Naïve algorithm.

Some important applications of these algorithms are text classification, recommendation system and face recognition.

Algorithm 8.1: Naïve Bayes

1. Compute the prior probability for the target class.
2. Compute Frequency matrix and likelihood Probability for each of the feature.
3. Use Bayes theorem Eq. (8.1) to calculate the probability of all hypotheses.
4. Use Maximum A Posteriori (MAP) Hypothesis, h_{MAP} Eq. (8.2) to classify the test object to the hypothesis with the highest probability.

Example 8.2: Assess a student's performance using Naïve Bayes algorithm with the dataset provided in Table 8.1. Predict whether a student gets a job offer or not in his final year of the course.

Table 8.1: Training Dataset

S.No.	CGPA	Interactiveness	Practical Knowledge	Communication Skills	Job Offer
1.	≥9	Yes	Very good	Good	Yes
2.	≥8	No	Good	Moderate	Yes
3.	≥9	No	Average	Poor	No
4.	<8	No	Average	Good	No
5.	≥8	Yes	Good	Moderate	Yes
6.	≥9	Yes	Good	Moderate	Yes
7.	<8	Yes	Good	Poor	No
8.	≥9	No	Very good	Good	Yes
9.	≥8	Yes	Good	Good	Yes
10.	≥8	Yes	Average	Good	Yes

Solution: The training dataset T consists of 10 data instances with attributes such as 'CGPA', 'Interactiveness', 'Practical Knowledge' and 'Communication Skills' as shown in Table 8.1. The target variable is Job Offer which is classified as Yes or No for a candidate student.

Step 1: Compute the prior probability for the target feature 'Job Offer'. The target feature 'Job Offer' has two classes, 'Yes' and 'No'. It is a binary classification problem. Given a student instance, we need to classify whether 'Job Offer = Yes' or 'Job Offer = No'.

From the training dataset, we observe that the frequency or the number of instances with 'Job Offer = Yes' is 7 and 'Job Offer = No' is 3.

The prior probability for the target feature is calculated by dividing the number of instances belonging to a particular target class by the total number of instances.

Hence, the prior probability for 'Job Offer = Yes' is 7/10 and 'Job Offer = No' is 3/10 as shown in Table 8.2.

Table 8.2: Frequency Matrix and Prior Probability of Job Offer

Job Offer Classes	No. of Instances	Probability Value
Yes	7	P (Job Offer = Yes) = 7/10
No	3	P (Job Offer = No) = 3/10

Step 2: Compute Frequency matrix and Likelihood Probability for each of the feature.

Step 2(a): Feature – CGPA

Table 8.3 shows the frequency matrix for the feature CGPA.

Table 8.3: Frequency Matrix of CGPA

CGPA	Job Offer = Yes	Job Offer = No
≥9	3	1
≥8	4	0
<8	0	2
Total	7	3

Table 8.4 shows how the likelihood probability is calculated for CGPA using conditional probability.

Table 8.4: Likelihood Probability of CGPA

CGPA	P (Job Offer = Yes)	P (Job Offer = No)
≥9	P (CGPA ≥9 | Job Offer = Yes) = 3/7	P (CGPA ≥9 | Job Offer = No) = 1/3
≥8	P (CGPA ≥8 | Job Offer = Yes) = 4/7	P (CGPA ≥8 | Job Offer = No) = 0/3
<8	P (CGPA <8 | Job Offer = Yes) = 0/7	P (CGPA <8 | Job Offer = No) = 2/3

As explained earlier the Likelihood probability is stated as the sampling density for the evidence given the hypothesis. It is denoted as P (Evidence | Hypothesis), which says how likely is the occurrence of the evidence given the parameters.

It is calculated as the number of instances of each attribute value and for a given class value divided by the number of instances with that class value.

For example P (CGPA ≥9 | Job Offer = Yes) denotes the number of instances with 'CGPA ≥9' and 'Job Offer = Yes' divided by the total number of instances with 'Job Offer = Yes'.

From the Table 8.3 Frequency Matrix of CGPA, number of instances with 'CGPA ≥9' and 'Job Offer = Yes' is 3. The total number of instances with 'Job Offer = Yes' is 7. Hence, P (CGPA ≥9 | Job Offer = Yes) = 3/7.

Similarly, the Likelihood probability is calculated for all attribute values of feature CGPA.

Step 2(b): Feature – Interactiveness

Table 8.5 shows the frequency matrix for the feature Interactiveness.

Table 8.5: Frequency Matrix of Interactiveness

Interactiveness	Job Offer = Yes	Job Offer = No
YES	5	1
NO	2	2
Total	7	3

Table 8.6 shows how the likelihood probability is calculated for Interactiveness using conditional probability.

Table 8.6: Likelihood Probability of Interactiveness

Interactiveness	P (Job Offer = Yes)	P (Job Offer = No)
YES	*P* (Interactiveness = Yes | Job Offer = Yes) = 5/7	*P* (Interactiveness = Yes | Job Offer = No) = 1/3
NO	*P* (Interactiveness = No | Job Offer = Yes) = 2/7	*P* (Interactiveness = No | Job Offer = No) = 2/3

Step 2(c): Feature – Practical Knowledge

Table 8.7 shows the frequency matrix for the feature Practical Knowledge.

Table 8.7: Frequency Matrix of Practical Knowledge

Practical Knowledge	Job Offer = Yes	Job Offer = No
Very Good	2	0
Average	1	2
Good	4	1
Total	7	3

Table 8.8 shows how the likelihood probability is calculated for Practical Knowledge using conditional probability.

Table 8.8: Likelihood Probability of Practical Knowledge

Practical Knowledge	*P* (Job Offer = Yes)	*P* (Job Offer = No)
Very Good	*P* (Practical Knowledge = Very Good | Job Offer = Yes) = 2/7	*P* (Practical Knowledge = Very Good | Job Offer = No) = 0/3
Average	*P* (Practical Knowledge = Average | Job Offer = Yes) = 1/7	*P* (Practical Knowledge = Average | Job Offer = No) = 2/3
Good	*P* (Practical Knowledge = Good | Job Offer = Yes) = 4/7	*P* (Practical Knowledge = Good | Job Offer = No) = 1/3

Step 2(d): Feature – Communication Skills

Table 8.9 shows the frequency matrix for the feature Communication Skills.

Table 8.9: Frequency Matrix of Communication Skills

Communication Skills	Job Offer = Yes	Job Offer = No
Good	4	1
Moderate	3	0
Poor	0	2
Total	7	3

Table 8.10 shows how the likelihood probability is calculated for Communication Skills using conditional probability.

Table 8.10: Likelihood Probability of Communication Skills

Communication Skills	P (Job Offer = Yes)	P (Job Offer = No)
Good	P (Communication Skills = Good | Job Offer = Yes) = 4/7	P (Communication Skills = Good | Job Offer = No) = 1/3
Moderate	P (Communication Skills = Moderate | Job Offer = Yes) = 3/7	P (Communication Skills = Moderate | Job Offer = No) = 0/3
Poor	P (Communication Skills = Poor | Job Offer = Yes) = 0/7	P (Communication Skills = Poor | Job Offer = No) = 2/3

Step 3: Use Bayes theorem Eq. (8.1) to calculate the probability of all hypotheses.

Given the test data = (CGPA ≥9, Interactiveness = Yes, Practical knowledge = Average, Communication Skills = Good), apply the Bayes theorem to classify whether the given student gets a Job offer or not.

P (Job Offer = Yes | Test data) = (P(CGPA ≥9 | Job Offer = Yes) P (Interactiveness = Yes | Job Offer = Yes) P (Practical knowledge = Average | Job Offer = Yes) P (Communication Skills = Good | Job Offer = Yes) P (Job Offer = Yes)))/(P (Test Data))

We can ignore P (Test Data) in the denominator since it is common for all cases to be considered.

Hence, P (Job Offer = Yes | Test data) = (P(CGPA ≥9 | Job Offer = Yes) P (Interactiveness = Yes | Job Offer = Yes) P (Practical knowledge = Average | Job Offer = Yes) P (Communication Skills = Good | Job Offer = Yes) P (Job Offer = Yes)

= 3/7 × 5/7 × 1/7 × 4/7 × 7/10

= 0.0175

Similarly, for the other case 'Job Offer = No',

We compute the probability,

P (Job Offer = No | Test data) = (P(CGPA ≥9 | Job Offer =No) P (Interactiveness = Yes | Job Offer = No) P (Practical knowledge = Average | Job Offer = No) P (Communication Skills = Good | Job Offer =No) P (Job Offer = No))/(P(Test Data)).

P (CGPA ≥9 | Job Offer = No) P (Interactiveness = Yes | Job Offer = No) P (Practical knowledge = Average | Job Offer = No) P (Communication Skills = Good | Job Offer = No) P (Job Offer = No)

= 1/3 × 1/3 × 2/3 × 1/3 × 3/10

= 0.0074

Step 4: Use Maximum A Posteriori (MAP) Hypothesis, h_{MAP} Eq. (8.2) to classify the test object to the hypothesis with the highest probability.

Since P (Job Offer = Yes | Test data) has the highest probability value, the test data is classified as '**Job Offer = Yes**'.

Zero Probability Error

In Example 8.1, consider the test data to be (CGPA ≥8, Interactiveness = Yes, Practical knowledge = Average, Communication Skills = Good)

When computing the posterior probability,

P (Job Offer = Yes | Test data) = (P(CGPA ≥8 | Job Offer = Yes) P (Interactiveness = Yes | Job Offer = Yes) P (Practical knowledge = Average | Job Offer = Yes) P (Communication Skills = Good | Job Offer = Yes) P (Job Offer = Yes)))/((P(Test Data))

P (Job Offer = Yes | Test data) = (P(CGPA ≥8 | Job Offer = Yes) P(Interactiveness = Yes | Job Offer = Yes) P (Practical knowledge = Average | Job Offer = Yes) P (Communication Skills = Good | Job Offer = Yes) P (Job Offer = Yes)

= 4/7 × 5/7 × 1/7 × 4/7 × 7/10

= 0.0233

Similarly, for the other case 'Job Offer = No',

When we compute the probability:

P (Job Offer = No | Test data) = (P(CGPA ≥8 | Job Offer = No) P (Interactiveness = Yes | Job Offer = No) P (Practical knowledge = Average | Job Offer = No) P (Communication Skills = Good | Job Offer = No) P (Job Offer = No))/(P(Test Data))

= P (CGPA ≥8 | Job Offer =No) P (Interactiveness = Yes | Job Offer = No) P (Practical knowledge = Average | Job Offer = No) P (Communication Skills = Good | Job Offer =No) P (Job Offer = No)

= 0/3 × 1/3 × 2/3 × 1/3 × 3/10

= 0

Since the probability value is zero, the model fails to predict, and this is called as Zero-Probability error. This problem arises because there are no instances in the given Table 8.1 for the attribute value CGPA ≥8 and Job Offer = No and hence the probability value of this case is zero. This zero-probability error can be solved by applying a smoothing technique called Laplace correction which means given 1000 data instances in the training dataset, if there are zero instances for a particular value of a feature we can add 1 instance for each attribute value pair of that feature which will not make much difference for 1000 data instances and the overall probability does not become zero.

Now, let us scale the values given in Table 8.1 for 1000 data instances. The scaled values without Laplace correction are shown in Table 8.11.

Table 8.11: Scaled Values to 1000 without Laplace Correction

CGPA	P (Job Offer = Yes)	P (Job Offer = No)
≥9	P (CGPA ≥9 | Job Offer = Yes) = 300/700	P (CGPA ≥9 | Job Offer = No) = 100/300
≥8	P (CGPA ≥8 | Job Offer = Yes) = 400/700	P (CGPA ≥8 | Job Offer = No) = 0/300
<8	P (CGPA <8 | Job Offer = Yes) = 0/700	P (CGPA <8 | Job Offer = No) = 200/300

Now, add 1 instance for each CGPA-value pair for 'Job Offer = No'. Then,

P (CGPA ≥9 | Job Offer = No) = 101/303 = 0.333

P (CGPA ≥8 | Job Offer = No) = 1/303 = 0.0033

P (CGPA <8 | Job Offer = No) = 201/303 = 0.6634

With scaled values to 1003 data instances, we get

P (Job Offer = Yes | Test data) = (P(CGPA ≥8 | Job Offer = Yes) P (Interactiveness = Yes | Job Offer = Yes) P (Practical knowledge = Average | Job Offer = Yes) P (Communication Skills = Good | Job Offer = Yes) P (Job Offer = Yes)

= 400/700 × 500/700 × 100/700 × 400/700 × 700/1003

= 0.02325

P (Job Offer = No | Test data) = P(CGPA ≥8 | Job Offer = No) P (Interactiveness = Yes | Job Offer = No) P (Practical knowledge = Average | Job Offer = No) P (Communication Skills = Good | Job Offer = No) P (Job Offer = No)

= 1/303 × 100/300 × 200/300 × 100/300 × 303/1003

= 0.00007385

Thus, using Laplace Correction, Zero Probability error can be solved with Naïve Bayes classifier.

8.3.2 Brute Force Bayes Algorithm

Applying Bayes theorem, Brute Force Bayes algorithm relies on the idea of concept learning wherein given a hypothesis space H for the training dataset T, the algorithm computes the posterior probabilities for all the hypothesis $h_i \in H$. Then, Maximum A Posteriori (MAP) Hypothesis, h_{MAP} is used to output the hypothesis with maximum posterior probability. The algorithm is quite expensive since it requires computations for all the hypotheses. Although computing posterior probabilities is inefficient, this idea is applied in various other algorithms which is also quite interesting.

8.3.3 Bayes Optimal Classifier

Bayes optimal classifier is a probabilistic model, which in fact, uses the Bayes theorem to find the most probable classification for a new instance given the training data by combining the predictions of all posterior hypotheses, This is different from Maximum A Posteriori (MAP) Hypothesis, h_{MAP} which chooses the maximum probable hypothesis or the most probable hypothesis.

Here, a new instance can be classified to a possible classification value C_i by the following Eq. (8.4).

$$= \max_{C_i} \sum_{h_i \in H} P(C_i \mid h_i)\, P(h_i \mid T) \tag{8.4}$$

Example 8.3: Given the hypothesis space with 4 hypothesis h_1, h_2, h_3 and h_4. Determine if the patient is diagnosed as COVID positive or COVID negative using Bayes Optimal classifier.

Solution: From the training dataset T, the posterior probabilities of the four different hypotheses for a new instance are given in Table 8.12.

Table 8.12: Posterior Probability Values

$P(h_i \mid T)$	P (COVID Positive $\mid h_i$)	P (COVID Negative $\mid h_i$)
0.3	0	1
0.1	1	0
0.2	1	0
0.1	1	0

h_{MAP} chooses h_1 which has the maximum probability value 0.3 as the solution and gives the result that the patient is COVID negative. But Bayes Optimal classifier combines the predictions of h_2, h_3 and h_4 which is 0.4 and gives the result that the patient is COVID positive.

$$\sum_{h_i \in H} P(\text{COVID Negative} \mid h_i) P(h_i \mid T) = 0.3 \times 1 = 0.3$$

$$\sum_{h_i \in H} P(\text{COVID Positive} \mid h_i) P(h_i \mid T) = 0.1 \times 1 + 0.2 \times 1 + 0.1 \times 1 = 0.4$$

Therefore, $\max_{c_i \in \{\text{COVID Positive, COVID Negative}\}} \sum_{h_i \in H} P(C_i|h_i)P(h_i|T)$ = COVID Positive.

Thus, this algorithm, diagnoses the new instance to be COVID positive.

8.3.4 Gibbs Algorithm

The main drawback of Bayes optimal classifier is that it computes the posterior probability for all hypotheses in the hypothesis space and then combines the predictions to classify a new instance.

Gibbs algorithm is a sampling technique which randomly selects a hypothesis from the hypothesis space according to the posterior probability distribution and classifies a new instance. It is found that the prediction error occurs twice with the Gibbs algorithm when compared to Bayes Optimal classifier.

8.4 NAÏVE BAYES ALGORITHM FOR CONTINUOUS ATTRIBUTES

There are two ways to predict with Naive Bayes algorithm for continuous attributes:

1. Discretize continuous feature to discrete feature.
2. Apply Normal or Gaussian distribution for continuous feature.

Gaussian Naive Bayes Algorithm

In Gaussian Naive Bayes, the values of continuous features are assumed to be sampled from a Gaussian distribution.

Example 8.4: Assess a student's performance using Naïve Bayes algorithm for the continuous attribute. Predict whether a student gets a job offer or not in his final year of the course. The training dataset T consists of 10 data instances with attributes such as 'CGPA' and 'Interactiveness' as shown in Table 8.13. The target variable is Job Offer which is classified as Yes or No for a candidate student.

Table 8.13: Training Dataset with Continuous Attribute

S.No.	CGPA	Interactiveness	Job Offer
1.	9.5	Yes	Yes
2.	8.2	No	Yes
3.	9.3	No	No
4.	7.6	No	No
5.	8.4	Yes	Yes
6.	9.1	Yes	Yes
7.	7.5	Yes	No
8.	9.6	No	Yes
9.	8.6	Yes	Yes
10.	8.3	Yes	Yes

Solution:

Step 1: Compute the prior probability for the target feature 'Job Offer'.

Prior probabilities of both the classes are calculated using the same formula (refer to Table 8.14).

Table 8.14: Prior Probability of Target Class

Job Offer Classes	No. of Instances	Probability Value
Yes	7	P (Job Offer = Yes) = 7/10
No	3	P (Job Offer = No) = 3/10

Step 2: Compute Frequency matrix and Likelihood Probability for each of the feature.

Likelihood probabilities for a continuous attribute is obtained from Gaussian (Normal) Distribution. In the above data set, CGPA is a continuous attribute for which we need to apply Gaussian distribution to calculate the likelihood probability.

Gaussian distribution for continuous feature is calculated using the given formula,

$$P(X_i = x_k \mid C_j) = g(x_k, \mu_{ij}, \sigma_{ij}) \tag{8.5}$$

where,

X_i is the i^{th} continuous attribute in the given dataset and x_k is a value of the attribute.

C_j denotes the j^{th} class of the target feature.

μ_{ij} denotes the mean of the values of that continuous attribute X_i with respect to the class j of the target feature.

σ_{ij} denotes the standard deviation of the values of that continuous attribute X_i with respect to the class j of the target feature.

Hence, the normal distribution formula is given as:

$$P(X_i = x_k \mid C_j) = \frac{1}{\sigma_{ij}\sqrt{2\pi}} e^{\frac{(x_k - \mu_{ij})^2}{2\sigma_{ij}^2}} \tag{8.6}$$

Step 2(a): Consider the feature CGPA

In this example CGPA is a continuous attribute,

To calculate the likelihood probability for this continuous attribute, first compute the mean and standard deviation for CGPA with respect to the target class 'Job Offer'.

Here, X_i = CGPA

C_j = 'Job Offer = Yes'

Mean and Standard Deviation for class 'Job Offer = Yes' are given as:

$\mu_{ij} = \mu_{CGPA - YES} = 8.814286$

$\sigma_{ij} = \sigma_{CGPA - YES} = 0.58146$

Mean and Standard Deviation for class 'Job Offer = No' are given as:

C_j = 'Job Offer = No'

$\mu_{ij} = \mu_{CGPA - NO} = 8.133333$

$\sigma_{ij} = \sigma_{CGPA - NO} = 1.011599$

Once Mean and Standard Deviation are computed, the likelihood probability for any test value using Gaussian distribution formula can be calculated.

Step 2(b): Consider the feature Interactiveness

Interactiveness is a discrete feature whose probability is calculated as earlier.

Table 8.15 shows the frequency matrix for the feature Interactiveness.

Table 8.15: Frequency Matrix of Interactiveness

Interactiveness	Job Offer = Yes	Job Offer = No
YES	5	1
NO	2	2
Total	7	3

Table 8.16 shows how the likelihood probability is calculated for Interactiveness using conditional probability.

Table 8.16: Likelihood Probability of Interactiveness

Interactiveness	P (Job Offer = Yes)	P (Job Offer = No)
YES	P (Interactiveness = Yes | Job Offer = Yes) = 5/7	P (Interactiveness = Yes | Job Offer = No) = 1/3
NO	P (Interactiveness = No | Job Offer = Yes) = 2/7	P (Interactiveness = No | Job Offer = No) = 2/3

Step 3: Use Bayes theorem to calculate the probability of all hypotheses.

Consider the test data to be (CGPA = 8.5, Interactiveness = Yes).

For the hypothesis 'Job Offer = Yes':

P (Job Offer = Yes | Test data) = (P(CGPA = 8.5 | Job Offer = Yes) × P (Interactiveness = Yes | Job Offer = Yes) × P (Job Offer = Yes)

To compute P (CGPA = 8.5 | Job Offer = Yes) use Gaussian distribution formula:

$$P(X_i = x_k \mid C_j) = g(x_k, \mu_{ij}, \sigma_{ij})$$

$$P(X_{CGPA} = 8.5 \mid C_{\text{Job Offer = Yes}}) = \frac{1}{\sigma_{ij}\sqrt{2\pi}} e^{\frac{(x_k - \mu_{ij})^2}{2\sigma_{ij}^2}}$$

$$P(X_{CGPA} = 8.5 \mid C_{\text{Job Offer = Yes}}) = \frac{1}{\sigma_{CGPA-YES}\sqrt{2\pi}} e^{\frac{(8.5 - \mu_{CGPA-YES})^2}{2\sigma_{CGPA\ YES}^2}}$$

$$P(\text{CGPA} = 8.5 \mid \text{Job Offer} = \text{Yes}) = g(x_k = 8.5, \mu_{ij} = 8.814, \sigma_{ij} = 0.581)$$

$$= \frac{1}{0.581\sqrt{2\pi}} e^{\frac{(8.5 - 8.814)^2}{2 \times 0.581^2}} = 0.594$$

P (Interactiveness = Yes | Job Offer = Yes) = 5/7

P (Job Offer = Yes) = 7/10

Hence:

P (Job Offer = Yes | Test data) = (P(CGPA = 8.5 | Job Offer = Yes) × P (Interactiveness = Yes | Job Offer = Yes) × P (Job Offer = Yes)

$= 0.594 \times 5/7 \times 7/10$

$= \mathbf{0.297}$

Similarly, for the hypothesis 'Job Offer = No':

P (Job Offer = No | Test data) = P (CGPA = 8.5 | Job Offer = No) × P (Interactiveness = Yes | Job Offer = No) × P (Job Offer = No)

P (CGPA = 8.5 | Job Offer = No) = g (x_k = 8.5, μ_{ij} = 8.133 σ_{ij} = 1.0116)

$$P(X_{CGPA} = 8.5 \mid C_{\text{Job Offer = Yes}}) = \frac{1}{\sigma_{CGPA-NO}\sqrt{2\pi}} e^{\frac{-(8.5-\mu_{CGPA-NO})^2}{2\sigma_{CGPA-NO}^2}}$$

$$= \frac{1}{1.0116\sqrt{2\pi}} e^{\frac{(8.5-8.133)^2}{2\times 1.0116^2}} = 0.369$$

P (Interactiveness = Yes | Job Offer = No) = 1/3

= P (Job Offer = No) = 0.369

Hence,

P (Job Offer = No | Test data) = P (CGPA = 8.5 | Job Offer = No) P (Interactiveness = Yes | Job Offer = No) × P (Job Offer = No)

$= 0.369 \times 1/3 \times 3/10$

$= \mathbf{0.0369}$

Step 4: Use Maximum A Posteriori (MAP) Hypothesis, h_{MAP} to classify the test object to the hypothesis with the highest probability.

Since P (Job Offer = Yes | Test data) has the highest probability value of 0.297,the test data is classified as '**Job Offer = Yes**'.