



Chapter 8

Bayesian Learning

"In science, progress is possible. In fact, if one believes in Bayes' theorem, scientific progress is inevitable as predictions are made and as beliefs are tested and refined."

— Nate Silver

Bayesian Learning is a learning method that describes and represents knowledge in an uncertain domain and provides a way to reason about this knowledge using probability measure. It uses Bayes theorem to infer the unknown parameters of a model. Bayesian inference is useful in many applications which involve reasoning and diagnosis such as game theory, medicine, etc. Bayesian inference is much more powerful in handling missing data and for estimating any uncertainty in predictions.

Learning Objectives

- Understand the basics of probability-based learning and probability theory
- Learn the fundamentals of Bayes theorem
- Introduce Bayes Classification models such as Brute Force Bayes learning algorithm, Bayes Optimal classifier, and Gibbs algorithm
- Introduce Naïve Bayes Classification models that work on the principle of Bayes theorem
- Explore the Naïve Bayes classification algorithm
- Study about Naïve Bayes Algorithm for continuous attributes using Gaussian distribution
- Introduce other popular types of Naive Bayes classifiers such as Bernoulli Naive Bayes classifier, Multinomial Naive Bayes classifier, and Multi-class Naïve Bayes classifier

8.1 INTRODUCTION TO PROBABILITY-BASED LEARNING

Probability-based learning is one of the most important practical learning methods which combines prior knowledge or prior probabilities with observed data. Probabilistic learning uses the concept of probability theory that describes how to model randomness, uncertainty, and noise to predict future events. It is a tool for modelling large datasets and uses Bayes rule to infer unknown quantities, predict and learn from data. In a *probabilistic model*, randomness plays a major role which gives probability distribution a solution, while in a *deterministic model* there is no randomness and

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 $P(\text{Evidence} \mid \text{Hypothesis})$, 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. $P(\text{Hypothesis} \mid \text{Evidence})$ is the posterior distribution representing the belief about the hypothesis, given the evidence from the training data. Therefore,

$$\text{Posterior probability} = \text{prior probability} + \text{new evidence}$$

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:

$$P(\text{Hypothesis } h \mid \text{Evidence } E)$$

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

$P(\text{Hypothesis } h \mid \text{Evidence } E)$ is calculated from the prior probability $P(\text{Hypothesis } h)$, the likelihood probability $P(\text{Evidence } E \mid \text{Hypothesis } h)$ and the marginal probability $P(\text{Evidence } E)$. It can be written as:

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

where, $P(\text{Hypothesis } h)$ 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. $P(\text{Evidence } E)$ 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.

$P(\text{Evidence } E \mid \text{Hypothesis } h)$ 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. $P(\text{Hypothesis } h \mid \text{Evidence } E)$ 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 \propto Prior Probability \times 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} .

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

Maximum Likelihood (ML) Hypothesis, h_{ML}

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

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

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

$$P(A) = 5/8$$

$$P(B) = 4/8$$

$$P(A | B) = 2/4$$

$$P(B | A) = 2/5$$

$$P(A | B) = P(B | A) P(A) / P(B) = 2/4$$

$$P(B | A) = P(A | B) P(B) / P(A) = 2/5$$

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 volleyball, 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: $P(\text{Boy participating in the tournament}) = 90\%$

$$P(\text{He is sick} | \text{Boy participating in the tournament}) = 20\%$$

$$P(\text{He is Sick}) = 40\%$$

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

$$P(\text{Boy participating in the tournament} | \text{He is sick}) = P(\text{Boy participating in the tournament})$$

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

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

$$= 0.45$$

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	$P(\text{Job Offer} = \text{Yes}) = 7/10$
No	3	$P(\text{Job Offer} = \text{No}) = 3/10$

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	$P(\text{Job Offer} = \text{Yes})$	$P(\text{Job Offer} = \text{No})$
≥9	$P(\text{CGPA} \geq 9 \mid \text{Job Offer} = \text{Yes}) = 3/7$	$P(\text{CGPA} \geq 9 \mid \text{Job Offer} = \text{No}) = 1/3$
≥8	$P(\text{CGPA} \geq 8 \mid \text{Job Offer} = \text{Yes}) = 4/7$	$P(\text{CGPA} \geq 8 \mid \text{Job Offer} = \text{No}) = 0/3$
<8	$P(\text{CGPA} < 8 \mid \text{Job Offer} = \text{Yes}) = 0/7$	$P(\text{CGPA} < 8 \mid \text{Job Offer} = \text{No}) = 2/3$

As explained earlier the Likelihood probability is stated as the sampling density for the evidence given the hypothesis. It is denoted as $P(\text{Evidence} \mid \text{Hypothesis})$, 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 $P(\text{CGPA} \geq 9 \mid \text{Job Offer} = \text{Yes})$ 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, $P(\text{CGPA} \geq 9 \mid \text{Job Offer} = \text{Yes}) = 3/7$.

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	$P(\text{Job Offer} = \text{Yes})$	$P(\text{Job Offer} = \text{No})$
YES	$P(\text{Interactiveness} = \text{Yes} \mid \text{Job Offer} = \text{Yes}) = 5/7$	$P(\text{Interactiveness} = \text{Yes} \mid \text{Job Offer} = \text{No}) = 1/3$
NO	$P(\text{Interactiveness} = \text{No} \mid \text{Job Offer} = \text{Yes}) = 2/7$	$P(\text{Interactiveness} = \text{No} \mid \text{Job Offer} = \text{No}) = 2/3$

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	$P(\text{Job Offer} = \text{Yes})$	$P(\text{Job Offer} = \text{No})$
Very Good	$P(\text{Practical Knowledge} = \text{Very Good} \mid \text{Job Offer} = \text{Yes}) = 2/7$	$P(\text{Practical Knowledge} = \text{Very Good} \mid \text{Job Offer} = \text{No}) = 0/3$
Average	$P(\text{Practical Knowledge} = \text{Average} \mid \text{Job Offer} = \text{Yes}) = 1/7$	$P(\text{Practical Knowledge} = \text{Average} \mid \text{Job Offer} = \text{No}) = 2/3$
Good	$P(\text{Practical Knowledge} = \text{Good} \mid \text{Job Offer} = \text{Yes}) = 4/7$	$P(\text{Practical Knowledge} = \text{Good} \mid \text{Job Offer} = \text{No}) = 1/3$

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	$P(\text{Job Offer} = \text{Yes})$	$P(\text{Job Offer} = \text{No})$
Good	$P(\text{Communication Skills} = \text{Good} \text{Job Offer} = \text{Yes}) = 4/7$	$P(\text{Communication Skills} = \text{Good} \text{Job Offer} = \text{No}) = 1/3$
Moderate	$P(\text{Communication Skills} = \text{Moderate} \text{Job Offer} = \text{Yes}) = 3/7$	$P(\text{Communication Skills} = \text{Moderate} \text{Job Offer} = \text{No}) = 0/3$
Poor	$P(\text{Communication Skills} = \text{Poor} \text{Job Offer} = \text{Yes}) = 0/7$	$P(\text{Communication Skills} = \text{Poor} \text{Job Offer} = \text{No}) = 2/3$

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

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

$$P(\text{Job Offer} = \text{Yes} | \text{Test data}) = (P(\text{CGPA} \geq 9 | \text{Job Offer} = \text{Yes}) P(\text{Interactiveness} = \text{Yes} | \text{Job Offer} = \text{Yes}) P(\text{Practical knowledge} = \text{Average} | \text{Job Offer} = \text{Yes}) P(\text{Communication Skills} = \text{Good} | \text{Job Offer} = \text{Yes}) P(\text{Job Offer} = \text{Yes})) / (P(\text{Test Data}))$$

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

$$\begin{aligned} \text{Hence, } P(\text{Job Offer} = \text{Yes} | \text{Test data}) &= (P(\text{CGPA} \geq 9 | \text{Job Offer} = \text{Yes}) P(\text{Interactiveness} = \text{Yes} | \text{Job Offer} = \text{Yes}) \\ &\quad P(\text{Practical knowledge} = \text{Average} | \text{Job Offer} = \text{Yes}) P(\text{Communication Skills} = \text{Good} | \text{Job Offer} = \text{Yes}) P(\text{Job Offer} = \text{Yes})) \\ &= 3/7 \times 5/7 \times 1/7 \times 4/7 \times 7/10 \\ &= 0.0175 \end{aligned}$$

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

We compute the probability,

$$\begin{aligned} P(\text{Job Offer} = \text{No} | \text{Test data}) &= (P(\text{CGPA} \geq 9 | \text{Job Offer} = \text{No}) P(\text{Interactiveness} = \text{Yes} | \text{Job Offer} = \text{No}) \\ &\quad P(\text{Practical knowledge} = \text{Average} | \text{Job Offer} = \text{No}) P(\text{Communication Skills} = \text{Good} | \text{Job Offer} = \text{No}) P(\text{Job Offer} = \text{No})) / (P(\text{Test Data})) \end{aligned}$$

$$\begin{aligned} \text{P(CGPA} \geq 9 | \text{Job Offer} = \text{No}) P(\text{Interactiveness} = \text{Yes} | \text{Job Offer} = \text{No}) P(\text{Practical knowledge} \\ = \text{Average} | \text{Job Offer} = \text{No}) P(\text{Communication Skills} = \text{Good} | \text{Job Offer} = \text{No}) P(\text{Job Offer} = \text{No}) \\ &= 1/3 \times 1/3 \times 2/3 \times 1/3 \times 3/10 \\ &= 0.0074 \end{aligned}$$

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 $P(\text{Job Offer} = \text{Yes} | \text{Test data})$ has the highest probability value, the test data is classified as 'Job Offer = Yes'.