



**SNS COLLEGE OF TECHNOLOGY**

**(An Autonomous Institution)**

**Coimbatore – 35.**



**DEPARTMENT OF BIOMEDICAL ENGINEERING**

## **UNIT – 2**

### **CLASSIFICATION: BAYESIAN DECISION THEORY, PARAMETRIC AND NON-PARAMETRIC METHODS**

#### **Introduction**

**Bayesian decision theory** refers to the statistical approach based on tradeoff quantification among various classification decisions based on the concept of Probability(Bayes Theorem) and the costs associated with the decision.

It is basically a classification technique that involves the use of the Bayes Theorem which is used to find the conditional probabilities.

In **Statistical pattern Recognition**, we will focus on the statistical properties of patterns that are generally expressed in probability densities (**pdf's and pmf's**), and this will command most of our attention in this article and try to develop the fundamentals of the Bayesian decision theory.

#### **Prerequisites**

#### **Random Variable**

A random variable is a function that maps a possible set of outcomes to some values like while tossing a coin and getting head H as 1 and Tail T as 0 where 0 and 1 are random variables.

#### **Bayes Theorem**

The conditional probability of A given B, represented by  $P(A | B)$  is the chance of occurrence of A given that B has occurred.

$$P(A | B) = P(A,B)/P(B) \text{ or}$$

By Using the Chain rule, this can also be written as:

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

$$P(A | B) = P(B|A)P(A)/P(B) \quad \text{———— (1)}$$

$$\text{Where, } P(B) = P(B,A) + P(B,A') = P(B|A)P(A) + P(B|A')P(A')$$

Here, equation (1) is known as the **Bayes Theorem of probability**

Our aim is to explore each of the components included in this theorem. Let's explore step by step:

**(a) Prior or State of Nature:**

- Prior probabilities represent how likely is each Class is going to occur.
- Priors are known before the training process.
- The state of nature is a random variable  $P(w_i)$ .
- If there are only two classes, then the sum of the priors is  $P(w_1) + P(w_2)=1$ , if the classes are exhaustive.

**(b) Class Conditional Probabilities:**

- It represents the probability of how likely a feature x occurs given that it belongs to the particular class. It is denoted by,  $P(X|A)$  where x is a particular feature
- It is the probability of how likely the feature x occurs given that it belongs to the class  $w_i$ .
- Sometimes, it is also known as the **Likelihood**.
- It is the quantity that we have to evaluate while training the data. During the training process, we have input(features) X labeled to corresponding class w and we figure out the likelihood of occurrence of that set of features given the class label.

**(c) Evidence:**

- It is the probability of occurrence of a particular feature i.e.  $P(X)$ .
- It can be calculated using the chain rule as,  $P(X) = \sum_{in} P(X | w_i) P(w_i)$

- As we need the likelihood of class conditional probability is also figure out evidence values during training.

#### **(d) Posterior Probabilities:**

- It is the probability of occurrence of Class A when certain Features are given
- It is what we aim at computing in the test phase in which we have testing input or features (the given entity) and have to find how likely trained model can predict features belonging to the particular class  $w_i$ .

**For a better understanding of the above theory, we consider an example**

#### **Problem Description**

Suppose we have a classification problem statement where we have to classify among the object-1 and object-2 with the given set of features  $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]^T$ .

#### **Objective**

The main objective of designing a such classifier is to suggest actions when presented with unseen features, i.e, object not yet seen i.e, not in training data.

In this example let  $w$  denotes the state of nature with  $w = w_1$  for **object-1** and  $w = w_2$  for **object-2**. Here, we need to know that in reality, the state of nature is so unpredictable that we generally consider that was variable that is described probabilistically.

#### **Priors**

- Generally, we assume that there is some prior value  $P(w_1)$  that the next object is object-1 and  $P(w_2)$  that the next object is object-2. If we have no other object as in this problem then the sum of their prior is 1 i.e. the priors are exhaustive.
- The prior probabilities reflect the prior knowledge of how likely we will get object-1 and object-2. It is domain-dependent as the prior may change based on the time of year they are being caught.

It sounds somewhat strange and when judging multiple objects (as in a more realistic scenario) makes this decision rule stupid as we always make the same decision based on the largest prior even though we know that any other type of objective also might appear governed by the leftover prior probabilities (as priors are exhaustive in nature).

**Consider the following different scenarios:**

- If  $P(\omega_1) \gg \gg P(\omega_2)$ , our decision in favor of  $\omega_1$  will be correct most of the time we predict.
- But if  $P(\omega_1) = P(\omega_2)$ , half probable of our prediction of being right. In general, the probability of error is the minimum of  $P(\omega_1)$  and  $P(\omega_2)$ , and later in this article, we will see that under these conditions no other decision rule can yield a larger probability of being correct.

### **Feature Extraction process (Extract feature from the images)**

A suggested set of features- **Length, width, shapes of an object**, etc.

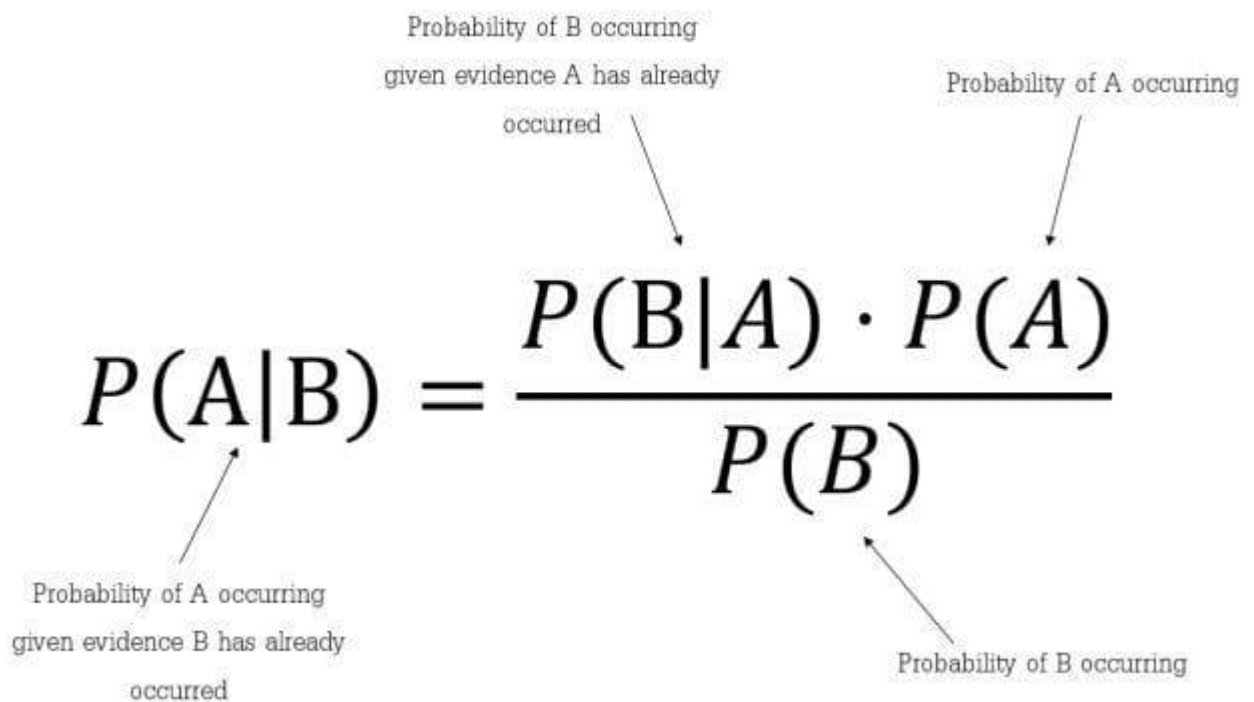
In our example, we use the **width x**, which is more **discriminatory** to improve the decision rule of our classifier. The different objects will yield different variable-width readings and we usually see this variability in probabilistic terms and also we consider  $x$  to be a continuous random variable whose distribution depends on the type of object  $w_j$ , and is expressed as  $p(x|\omega_j)$  (probability distribution function pdf as a continuous variable) and known as the class-conditional probability density function. Therefore,

The pdf  $p(x|\omega_1)$  is the probability density function for feature  $x$  given that the state of nature is  $\omega_1$  and the same interpretation for  $p(x|\omega_2)$ .

Fig. Picture Showing pdf for both classes

**Image Source: Google Images**

Suppose that we are well aware of both the prior probabilities  $P(\omega_j)$  and the conditional densities  $p(x|\omega_j)$ . Now, we can arrive at the Bayes formula for finding posterior probabilities:



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

Probability of B occurring given evidence A has already occurred

Probability of A occurring

Probability of A occurring given evidence B has already occurred

Probability of B occurring

Fig. Formula of Bayes Theorem

**Image Source: Google Images**

Bayes' formula gives us intuition that by observing the measurement of  $x$  we can convert the prior  $P(\omega_j)$  to the posteriors, denoted by  $P(\omega_j|x)$  which is the probability of  $\omega_j$  given that feature value  $x$  has been measured.

$p(x|\omega_j)$  is known as the likelihood of  $\omega_j$  with respect to  $x$ .

The evidence factor,  $p(x)$ , works as merely a scale factor that guarantees that the posterior probabilities sum up to one for all the classes.

### **Bayes' Decision Rule**

The decision rule given the posterior probabilities is as follows

If  $P(w_1|x) > P(w_2|x)$  we would decide that the object belongs to class  $w_1$ , or else class  $w_2$ .

### **Probability of Error**

To justify our decision we look at the probability of error, whenever we observe  $x$ , we have,

**$P(\text{error}|x) = P(\omega_1|x)$  if we decide  $\omega_2$ , and  $P(\omega_2|x)$  if we decide  $\omega_1$**

As they are exhaustive and if we choose the correct nature of an object by probability  $P$  then the leftover probability  $(1-P)$  will show how probable is the decision that it is not the decided object.

We can minimize the probability of error by deciding the one which has a greater posterior and the rest as the probability of error will be minimum as possible. So we finally get,

**$P(\text{error}|x) = \min [P(\omega_1|x), P(\omega_2|x)]$**

And our Bayes decision rule as,

**Decide  $\omega_1$  if  $P(\omega_1|x) > P(\omega_2|x)$ ; otherwise decide  $\omega_2$**

This type of decision rule highlights the role of the posterior probabilities. With the help of Bayes theorem, we can express the rule in terms of conditional and prior probabilities.

The evidence is unimportant as far as the decision is concerned. As we discussed earlier it is working as just a scale factor that states how frequently we will measure the feature with value  $x$ ; it assures  $P(\omega_1|x) + P(\omega_2|x) = 1$ .

So by eliminating the unrequired scale factor in our decision rule we have, the similar decision rule by Bayes theorem as,

**Decide  $\omega_1$  if  $p(x|\omega_1)P(\omega_1) > p(x|\omega_2)P(\omega_2)$ ; otherwise decide  $\omega_2$**

**Now, let's consider 2 cases:**

- **Case-1:** If class conditionals are equal i.e,  $p(x|\omega_1) = p(x|\omega_2)$ , then we arrive at our premature decision rule governed by just priors.
- **Case-2:** On the other hand, if priors are equal i.e,  $P(\omega_1) = P(\omega_2)$  then the decision is entirely based on class conditionals  $p(x|\omega_j)$ .

***This completes our example formulation!***

## **Generalization of the preceding ideas for Multiple Features and Classes**

Bayes classification: Posterior, likelihood, prior, and evidence

$$P(w_i | X) = P(X | w_i) P(w_i) / P(X)$$

**Posterior = Likelihood\* Prior/Evidence**

We now discuss those cases which have multiple features as well as multiple classes,

Let the Multiple Features be  $X_1, X_2, \dots, X_n$  and Multiple Classes be  $w_1, w_2, \dots, w_n$ , then:

$$P(w_i | X_1, \dots, X_n) = P(X_1, \dots, X_n | w_i) * P(w_i) / P(X_1, \dots, X_n)$$

Where,

$$\text{Posterior} = P(w_i | X_1, \dots, X_n)$$

$$\text{Likelihood} = P(X_1, \dots, X_n | w_i)$$

$$\text{Prior} = P(w_i)$$

$$\text{Evidence} = P(X_1, \dots, X_n)$$

In cases of the same incoming patterns, we might need to use a drastically different cost function, which will lead to different actions altogether. Generally, different decision tasks may require features and yield boundaries quite different from those useful for our original categorization problem.

So, In the later articles, we will discuss the **Cost function, Risk Analysis, and decisive action** which will further help to understand the Bayes decision theory in a better way.

**Parametric Methods:** The basic idea behind the parametric method is that there is a set of fixed parameters that uses to determine a probability model that is used in Machine Learning as well. Parametric methods are those methods for which we priory knows that the population is normal, or if not then we can easily approximate it using a normal distribution which is possible by invoking the Central Limit Theorem. Parameters for using the normal distribution is as follows:

- Mean
- Standard Deviation

Eventually, the classification of a method to be parametric is completely depends on the presumptions that are made about a population. There are many parametric methods available some of them are:

- Confidence interval used for – population mean along with known standard deviation.
- The confidence interval is used for – population means along with the unknown standard deviation.
- The confidence interval for population variance.
- The confidence interval for the difference of two means, with unknown standard deviation.

**Nonparametric Methods:** The basic idea behind the parametric method is no need to make any assumption of parameters for the given population or the population we are studying. In fact, the methods don't depend on the population. Here there is no fixed set of parameters are available, and also there is no distribution (normal distribution, etc.) of any kind is available for use. This is also the reason that nonparametric methods are also referred to as distribution-free methods. Nowadays Non-parametric methods are gaining popularity and an impact of influence some reasons behind this fame is:

- The main reason is that there is no need to be mannered while using parametric methods.
- The second important reason is that we do not need to make more and more assumptions about the population given (or taken) on which we are working on.
- Most of the nonparametric methods available are very easy to apply and to understand also i.e. the complexity is very low.

Reference:

1. <https://www.analyticsvidhya.com/blog/2021/05/an-intuitive-introduction-to-bayesian-decision-theory/>
2. <https://www.geeksforgeeks.org/difference-between-parametric-and-non-parametric-methods/>