

Bayesian Classification(or Naïve Bayesian classifier)

- **Bayesian classification** is a statistical technique used to classify the data based on probabilistic reasoning.
- It is a type of probabilistic classification that uses **Bayes' theorem** to predict the probability of a data point belonging to a certain class.
- The basic idea behind Bayesian classification is ***to assign a class label*** to a new data record based on the probability that it belongs to a particular class.

- The Bayesian classification is a powerful technique for probabilistic inference and decision-making and is widely used in various applications such as ***medical diagnosis, spam classification, fraud detection***, etc.

- ***Naïve Bayesian classifier:***

Bayesian classification is based on **Bayes' theorem**, a simple Bayesian classifier known as the ***naïve Bayesian classifier***.

- **Bayes' theorem:**

Bayes' theorem is named after ***Thomas Bayes***, an 18th-century British mathematician who first formulated it.

In data mining, *Bayes' theorem* is used to ***compute*** the ***probability of a hypothesis*** (such as a class label) for given some observed event (such as a set of attributes or features).

The ***formula for Bayes' Theorem*** is as follows:

$$P(H|X) = \frac{P(X|H) * P(H)}{P(X)}$$

Bayes' Theorem consists of two types of probabilities:

the ***posterior probability (P(H/X))*** and

the ***prior probability (P(H))***,

where **X** represents the data tuple and **H** represents a hypothesis(class Label).

- **Posterior probability:**

$P(H/X)$ is the *posterior probability* of hypothesis **H** conditioned on data tuple **X**.

For example, suppose we have data tuples from the fruits database described by the **attributes** *color* and *shape*, respectively, and that **X** is **red color** fruit with a shape of **round**. Suppose that **H** is the hypothesis that the fruit will an **Apple**.

Then the $P(H/X)$ represents the probability of fruit **X** is **Apple**, considering the fruit's *color* and *shape* information.

- **Prior probability:**

$P(H)$ is the **prior probability** of hypothesis H .

For example, this is the probability that any given fruit will *be an **Apple***, regardless of color, shape, or any other information, for that matter.

i.e., The *prior probability*(H), is ***independent*** of data tuple X .

- Similarly, $P(X/H)$ is the *posterior probability* of X conditioned on hypothesis H . That is, it is the probability that a fruit, X , is **red color** and shape is **round**, given that we know the fruit will be an **Apple**.
- Similarly, $P(X)$ is the *prior probability* of X . Using our example, it is the probability that a fruit from our data set of fruits is **red** color and **round** in shape.

Naive Bayes Classifier Algorithm:

- It is a classification technique based on ***Baye's Theorem*** and It is a simple Bayesian classifier.
- ***For example***, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter.
- Even if these features depend on each other, all of these properties independently contribute to the probability that this fruit is ***an apple*** and that is why it is known as ***“Naïve”***.

Working of Naïve Baye's Classifier:

This algorithm involves the following steps:

Step 1: Convert the given dataset into *frequency tables*.

Step 2: Generate *Likelihood table* by finding the probabilities of given features.

Step 3: Now, use *Bayes theorem* to calculate the *posterior probability*.

Example:

Suppose we have a dataset of *weather conditions* and corresponding target variable "***Play***". So, using the dataset we need to decide that whether we should ***play*** or ***not*** on a particular day according to the **weather conditions**

Problem: If the Weather is ***Sunny***, then the Player should ***play*** or ***not***?

Data Set:

	Weather	Humidity	Play
1	Rainy	High	Yes
2	Sunny	High	Yes
3	Overcast	High	Yes
4	Overcast	High	Yes
5	Sunny	Normal	No
6	Rainy	Normal	Yes
7	Sunny	Normal	Yes
8	Overcast	High	Yes
9	Rainy	Normal	No
10	Sunny	Normal	No
11	Sunny	Normal	Yes
12	Rainy	High	No
13	Overcast	Normal	Yes
14	Overcast	High	Yes

First, Convert the above dataset into ***frequency table*** as follows,

Weather	Yes	No
<i>Overcast</i>	5	0
<i>Rainy</i>	2	2
<i>Sunny</i>	3	2
Total	10	4

Next, Generate ***Likelihood table*** by finding the probabilities of given features as follows,

Weather	Yes	No	P(Yes)	P(No)	p
<i>Overcast</i>	5	0	5 / 10	0 / 4	5/14= 0.35
<i>Rainy</i>	2	2	2 / 10	2 / 4	4/14= 0.29
<i>Sunny</i>	3	2	3 / 10	2 / 4	5/14= 0.35
All	10/14= 0.71	4/14= 0.29			

Applying Bayes' theorem:

If the data tuple(**X**) is ***Sunny***, and the hypothesis(**H**) is **Yes** *then* the probability of ***Play*** *is*:

$$P(\text{Yes}|\text{Sunny}) = \frac{P(\text{Sunny}|\text{Yes}) * P(\text{Yes})}{P(\text{Sunny})}$$

Where,

$$P(\text{Sunny} | \text{Yes}) = 3/10 = \mathbf{0.3}$$

$$P(\text{Yes}) = 10/14 = \mathbf{0.71}$$

$$P(\text{Sunny}) = 5/14 = \mathbf{0.35}$$

Let us substitute the values in the above formula,

$$\begin{aligned} P(\text{Yes} | \text{Sunny}) &= (0.3 * 0.71) / 0.35 \\ &= \mathbf{0.60} \end{aligned}$$

If the Weather is ***Sunny***, then the probability of ***Play (Yes)*** is **0.60**

If the data tuple(**X**) is ***Sunny***, and the hypothesis(**H**) is **No** *then* the probability of ***Play*** is:

$$P(\text{No}|\text{Sunny}) = \frac{P(\text{Sunny}|\text{No}) * P(\text{No})}{P(\text{Sunny})}$$

Where,

$$P(\text{Sunny} | \text{No}) = 2/4 = \mathbf{0.5}$$

$$P(\text{No}) = 4/14 = \mathbf{0.29}$$

$$P(\text{Sunny}) = 5/14 = \mathbf{0.35}$$

Let us substitute the values in the above formula,

$$\begin{aligned} P(\text{No} | \text{Sunny}) &= (\mathbf{0.5 * 0.29}) / \mathbf{0.35} \\ &= \mathbf{0.41} \end{aligned}$$

If the Weather is ***Sunny***, then the probability of ***Play (No)*** is **0.41**.

- So, as we can see from the above calculation that $P(\text{Yes} | \text{Sunny}) > P(\text{No} | \text{Sunny})$
- Hence on a **Sunny** day, Player can **play the game**.

Exercise:

Problem: If the Weather is ***Sunny*** and Humidity is ***High***, then the Player should ***play*** or ***not***?

Solution:

- If the data tuple(**X**) is (***Sunny, High***), and the hypothesis(**H**) is **Yes** *then* the probability of ***Play is*** :

$$P(\text{Yes} | (\text{Sunny}, \text{High})) = \frac{P(\text{Sunny} | \text{Yes}) * P(\text{High} | \text{Yes}) * P(\text{Yes})}{P(\text{Sunny}) * P(\text{High})}$$

- If the data tuple(**X**) is (***Sunny, High***), and the hypothesis(**H**) is **No** *then* the probability of ***Play is*** :

$$P(\text{No} | (\text{Sunny}, \text{High})) = \frac{P(\text{Sunny} | \text{No}) * P(\text{High} | \text{No}) * P(\text{No})}{P(\text{Sunny}) * P(\text{High})}$$