

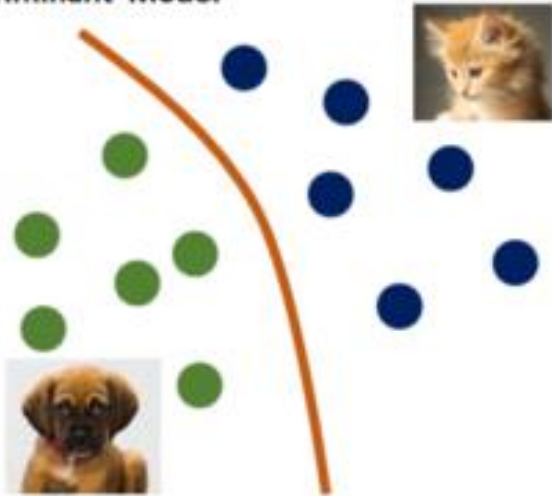
**Module No. 4****Bayesian and Computational Learning****8 Hours**

Bayes theorem , concept learning, maximum likelihood, minimum description length principle, Bayes optimal classifier, Gibbs Algorithm, Naïve Bayes Classifier, Bayesian belief network, EM algorithm, probability learning, sample complexity

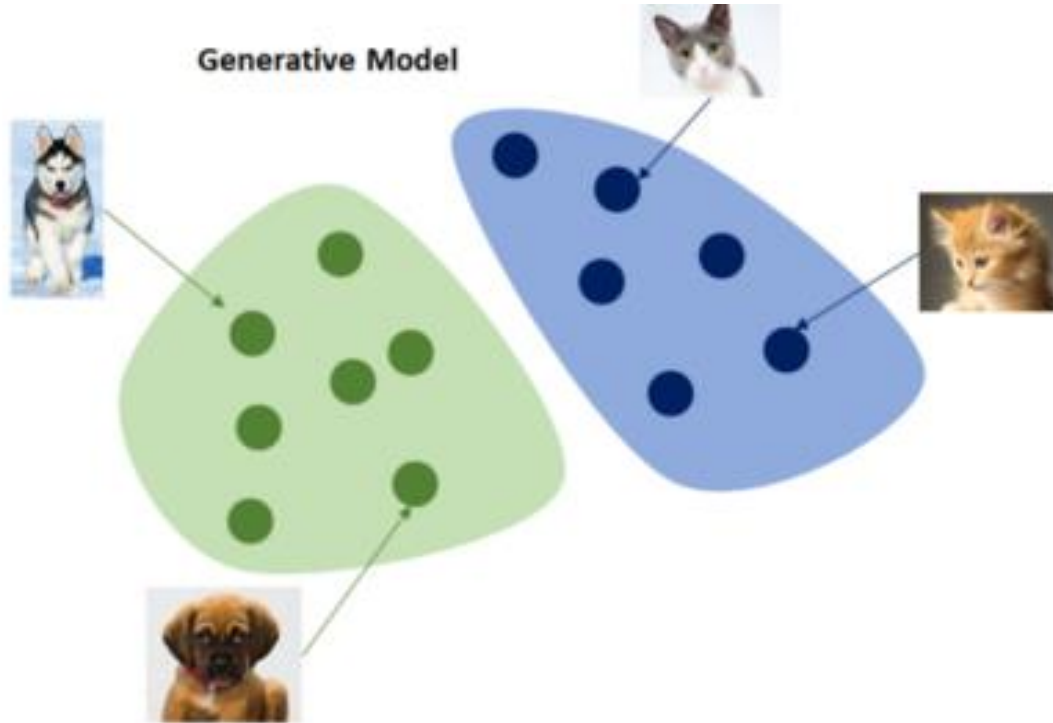
# Module 4

# Module 4

Discriminant Model



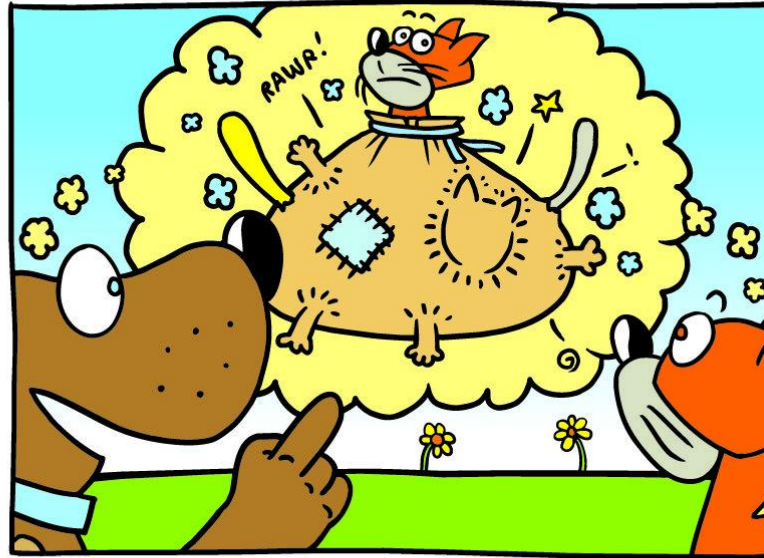
Generative Model



- Bayes theorem is given by an English statistician, philosopher, and Presbyterian minister named **Mr. Thomas Bayes** in 17<sup>th</sup> century.
- Bayes provides their thoughts in **decision theory** which is extensively used in important mathematics concepts as Probability.
- Bayes theorem is also widely used in Machine Learning where we need to **predict classes precisely and accurately**.
- An important concept of Bayes theorem named **Bayesian method** is used to **calculate conditional probability in Machine Learning application** that includes classification tasks.
- Further, **a simplified version of Bayes theorem** (Naïve Bayes classification) is also used to **reduce computation time and average cost of the projects**.
- Bayes theorem is also known with some other name such as **Bayes rule or Bayes Law**. ***Bayes theorem helps to determine the probability of an event with random knowledge.***

# Few terms

- Uncertainty
- Probability
- Simple Probability
- Joint Probability
- Conditional Probability
- Total Probability
- Reverse Probability



**"If there are 7 cats in a sack and I draw one at random,..."**



**"... what is the probability that I will draw you?"**

© Big Ideas Learning, LLC. All Rights Reserved

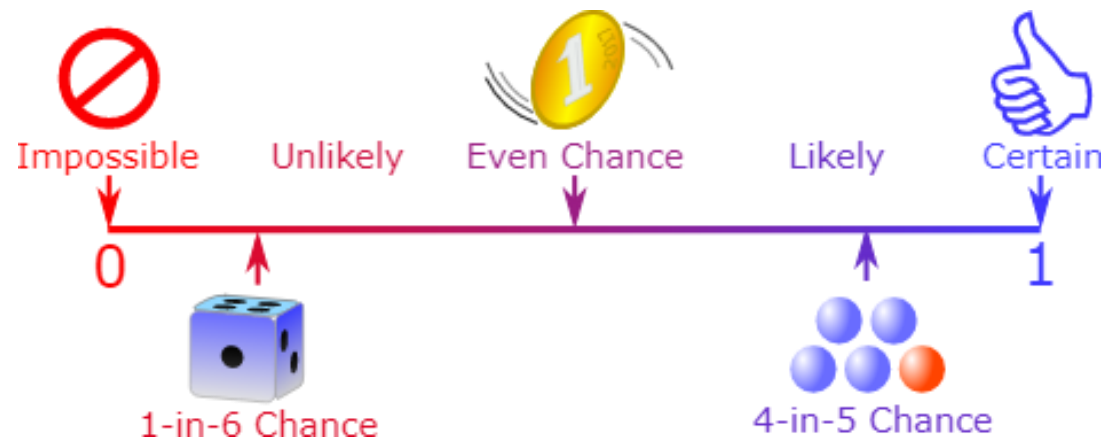
# Why use Bayes Theorem in Machine Learning?

- Bayes Theorem is a method to determine conditional probabilities – that is, the probability of one event occurring given that another event has already occurred.
- Because a conditional probability includes additional conditions – in other words, more data – it can contribute to more accurate results.

# Uncertainty



- Uncertainty is hard to bear for human beings.
- But in machine learning, there are certain algorithms that help to find your way around this limitation.
- The Naive Bayes machine learning algorithm is one of the tools to deal with uncertainty with the help of probabilistic methods.



# Probability

- Probability is a field of math that enables us to reason about uncertainty and assess the likelihood of some results or events.
- When you work with predictive ML modeling, you have to predict uncertain future.
- For example, you may try to predict the performance of an Olympic champion during the next Olympics based on past results.
- Even if they won before, it doesn't mean they will win this time. Unpredictable factors such as an argument with their partner this morning or no time to have breakfast may or may not influence their results.
- Three [main sources of uncertainty](#) in machine learning are noisy data, incomplete coverage of the problem, and imperfect models

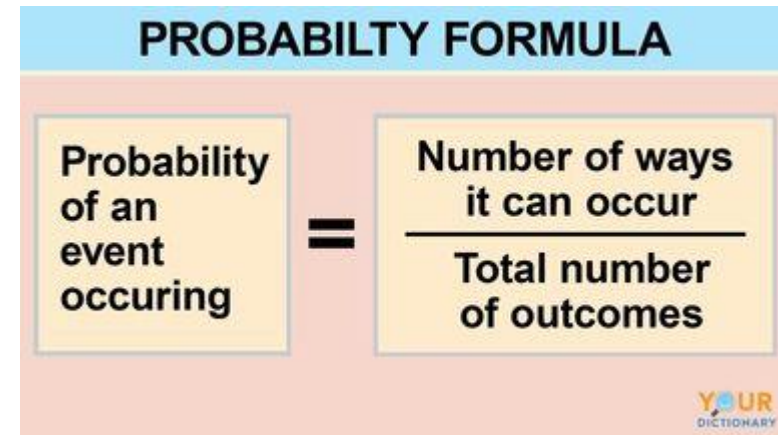
# Probability-examples

- A coin is thrown 3 times .what is the probability that atleast one head is obtained?

**Sol:** Sample space = [HHH, HHT, HTH, THH, TTH, THT, HTT, TTT]

Total number of ways =  $2 \times 2 \times 2 = 8$ . Fav. Cases = 7

$$P(A) = 7/8$$







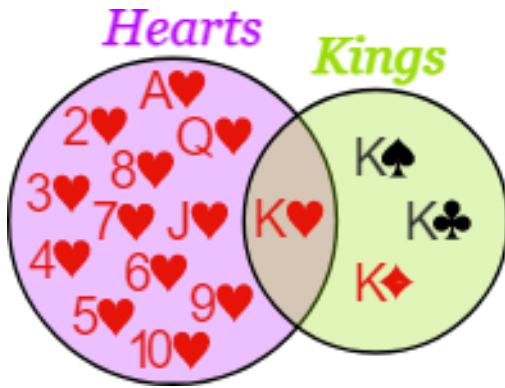
“What did the  
tomato say  
to the other  
tomato during  
a race?

# Mutually Exclusive

- When two events (call them "A" and "B") are Mutually Exclusive it is **impossible** for them to happen together:
- **$P(A \text{ and } B) = 0$**
- *"The probability of A and B together equals 0 (impossible)"*
- Example: King AND Queen
- A card cannot be a King AND a Queen at the same time!      The probability of a King **and** a Queen is **0** (Impossible)
- But, for Mutually Exclusive events, the probability of A **or** B is the sum of the individual probabilities:
- **$P(A \text{ or } B) = P(A) + P(B)$**
- *"The probability of A **or** B equals the probability of A **plus** the probability of B"*
- Example: King OR Queen
- In a Deck of 52 Cards:
- the probability of a King is  $1/13$ , so  **$P(\text{King})=1/13$**       the probability of a Queen is also  $1/13$ , so  **$P(\text{Queen})=1/13$**
- When we combine those two Events:      The probability of a King **or** a Queen is  $(1/13) + (1/13) = \mathbf{2/13}$
- Which is written like this:
- **$P(\text{King or Queen}) = (1/13) + (1/13) = 2/13$**

# Not Mutually Exclusive

But Hearts or Kings is:



all the Hearts (13 of them)

all the Kings (4 of them)

But that counts the King of Hearts twice!

So we correct our answer, by subtracting the extra "and" part:  
set hearts kings sum

16 Cards = 13 Hearts + 4 Kings – the 1 extra King of Hearts

Count them to make sure this works!

As a formula this is:  $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

"The probability of A or B equals the probability of A plus the probability of B minus the probability of A and B"

Here is the same formula, but using  $\cup$  and  $\cap$ :

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

# Mutually exclusive vs Independent events

- Mutually exclusive events are those that cannot happen simultaneously, whereas independent events are those whose probabilities do not affect one another.
- Events A and B are said to be independent if the probability of B occurring is unaffected by the occurrence of the event A happening.
- For example, now suppose that we are tossing a coin twice. Let A be the event that the first coin toss lands on heads.
- In addition, let B be the event that the second coin toss lands on heads. Clearly the result of the first coin toss does not affect the result of the second coin toss.

$$P(A \text{ and } B) = P(A) P(B)$$

# Conditional Probability

- We can try to answer the question of what is the probability § that an athlete wins the race given their results for the past 10 years.
- **This is how conditional probability is defined:** the probability of a, given b = the joint probability of both a and b happening, divided by the probability of b.

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

# Conditional Probability

## Definition 8.2: Conditional Probability

If events are dependent, then their probability is expressed by conditional probability. The probability that  $A$  occurs given that  $B$  is denoted by  $P(A|B)$ .

Suppose,  $A$  and  $B$  are two events associated with a random experiment. The probability of  $A$  under the condition that  $B$  has already occurred and  $P(B) \neq 0$  is given by

$$\begin{aligned} P(A|B) &= \frac{\text{Number of events in } B \text{ which are favourable to } A}{\text{Number of events in } B} \\ &= \frac{\text{Number of events favourable to } A \cap B}{\text{Number of events favourable to } B} \\ &= \frac{P(A \cap B)}{P(B)} \end{aligned}$$

	Vegetarian	Not Vegetarian	Total
Women	15	32	47
Men	29	24	53
Total	44	56	100

- if A represents being vegetarian and B represents being a woman, then  $P(A|B)$  and  $P(B|A)$  express different events
- The likelihood of a woman being vegetarian is:
- $P(A|B)=15/47$        $P(A|B)=15/47$
- Meanwhile, the likelihood of a vegetarian being a woman is:
- $P(B|A)=15/44$        $P(B|A)=15/44$
- Based on these outcomes, we can conclude that it is more likely for a vegetarian to be female, than for a woman *not* to eat meat. This goes to show that in probability theory things are never straightforward.

## Conditional Probability-Example

A die is rolled twice and two numbers are obtained, let X be the outcome of first roll and Y be the outcome of the second roll. Given that  $X+Y=5$ , what is the probability of  $X=4$  or  $Y=4$ ?

Assume, A be the event the getting 4 as X or Y, and B be the event of  $X+Y=5$ , therefore

$$A = \{(4,1), (4,2), (4,3), (4,4), (4,5), (4,6), (1,4), (2,4), (3,4), (4,4), (5,4), (6,4)\}$$

$$B = \{(1,4), (4,1), (2,3), (3,2)\}$$

We are interested in finding the probability of A given B

$$A \cap B = \{(1,4), (4,1)\}$$

As die is rolled out two times, total sample space = 36

$$P(A \cap B) = 2/36$$

$$P(B) = 4/36$$

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

$$= 2/4, \text{ or}$$

$$= 1/2.$$



# Total Probability

- It is a useful way to find the probability of some event  $A$  when we don't directly know the probability of  $A$  but we do know that events  $B_1, B_2, B_3 \dots$  form a partition of the sample space  $S$ .
- Let  $E_1, E_2, \dots, E_n$  be  $n$  mutually exclusive and exhaustive events associated with a random experiment. If  $A$  is any event which occurs with  $E_1$  or  $E_2$  or  $\dots, E_n$ , then
- $$P(A) = P(E_1) \cdot P(A|E_1) + P(E_2) \cdot P(A|E_2) + \dots + P(E_n) \cdot P(A|E_n)$$

Suppose there are two bags in a box, which contain the following marbles:

**Bag 1:** 7 red marbles and 3 green marbles

**Bag 2:** 2 red marbles and 8 green marbles

If we randomly select one of the bags and then randomly select one marble from that bag, what is the probability that it's a green marble?

let  $P(G)$  = probability of choosing a green marble.

$$P(G|B_1) = 3/10 = 0.3$$

$$P(G|B_2) = 8/10 = 0.8$$

Thus, using the law of total probability we can calculate the probability of choosing a green marble as:

$$P(G) = \sum P(G|B_i) * P(B_i)$$

$$P(G) = P(G|B_1) * P(B_1) + P(G|B_2) * P(B_2)$$

$$P(G) = (0.3) * (0.5) + (0.8) * (0.5)$$

$$P(G) = \mathbf{0.55}$$

# Note

- This theorem of total probability is the foundation of Baye's Theorem.
- The Baye's Theorem is defined for a sample space  $S$  containing a set of events  $E_1, E_2, E_3, \dots, E_n$ , which together constitutes the sample space  $S$ , such that  $E_1 \cup E_2 \cup E_3, \dots \cup E_n = S$ .
- The events are pair-wise disjoint, exhaustive, and with non-zero probabilities.

(a) Frequency Table

	Bread	No Bread	Total
Jam	40	10	50
No Jam	50	900	950
Total	90	910	1000

/ 1000 →

(b) Probability Table

	Bread	No Bread	Total
Jam	0.040	0.010	0.050
No Jam	0.050	0.900	0.950
Total	0.090	0.910	1

Joint    Marginal

buying Bread A, buying Jam B.

Marginal probability is the proportion of customers who bought Bread regardless of whether they bought Jam or not.

It's called marginal because it occurs at the margins of the probability table

*Marginal Probability*

$$P(A) = \frac{n(\text{Cust with Bread})}{n(\text{Cust})} = \frac{90}{1000}$$

$$P(B) = \frac{n(\text{Cust with Jam})}{n(\text{Cust})} = \frac{50}{1000}$$

### *Joint Probability*

$$\begin{aligned} & P(\text{Cust with Bread + Jam}) \\ &= P(A \text{ and } B) = \frac{n(A \text{ and } B)}{n(\text{Cust})} = \frac{40}{1000} \end{aligned}$$

### *Conditional Probability*

$$\begin{aligned} & P(\text{Cust with Bread given Jam}) \\ &= P(A|B) = \frac{n(A \text{ and } B)}{n(B)} = \frac{40}{50} \end{aligned}$$

$$\begin{aligned} & P(\text{Cust with Jam given Bread}) \\ &= P(B|A) = \frac{n(A \text{ and } B)}{n(A)} = \frac{40}{90} \end{aligned}$$

- This **alternate calculation** of the conditional probability is referred to as Bayes Rule or Bayes Theorem, named for Reverend Thomas Bayes, who is credited with first describing it.
- It is grammatically correct to refer to it as Bayes' Theorem (with the apostrophe), but it is common to omit the apostrophe for simplicity.

# Bayes theorem

Hence, Bayes Theorem can be written as:

**posterior = likelihood \* prior / evidence**

Probabilities that we are calculated before the new information are called **Prior**, and probabilities that we are calculated after the new information are called **Posterior**.

The diagram illustrates Bayes' Theorem with the following components:

- LIKELIHOOD**: The probability of "B" being True, given "A" is True. An arrow points from this label to the numerator term  $P(B|A)$ .
- PRIOR**: The probability "A" being True. This is the knowledge. An arrow points from this label to the numerator term  $P(A)$ .
- POSTERIOR**: The probability of "A" being True, given "B" is True. An arrow points from this label to the denominator term  $P(A|B)$ .
- MARGINALIZATION**: The probability "B" being True. An arrow points from this label to the denominator term  $P(B)$ .

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

## Likelihood

How probable is the evidence  
given that our hypothesis is true?

## Prior

How probable was our hypothesis  
before observing the evidence?

$$P(H | e) = \frac{P(e | H) P(H)}{P(e)}$$

## Posterior

How probable is our hypothesis  
given the observed evidence?  
(Not directly computable)

## Marginal

How probable is the new evidence  
under all possible hypotheses?  
 $P(e) = \sum P(e | H_i) P(H_i)$



# Prior and Posterior Probabilities

- $P(A)$  and  $P(B)$  are called prior probabilities
- $P(A|B)$ ,  $P(B|A)$  are called posterior probabilities

## Example 8.6: Prior versus Posterior Probabilities

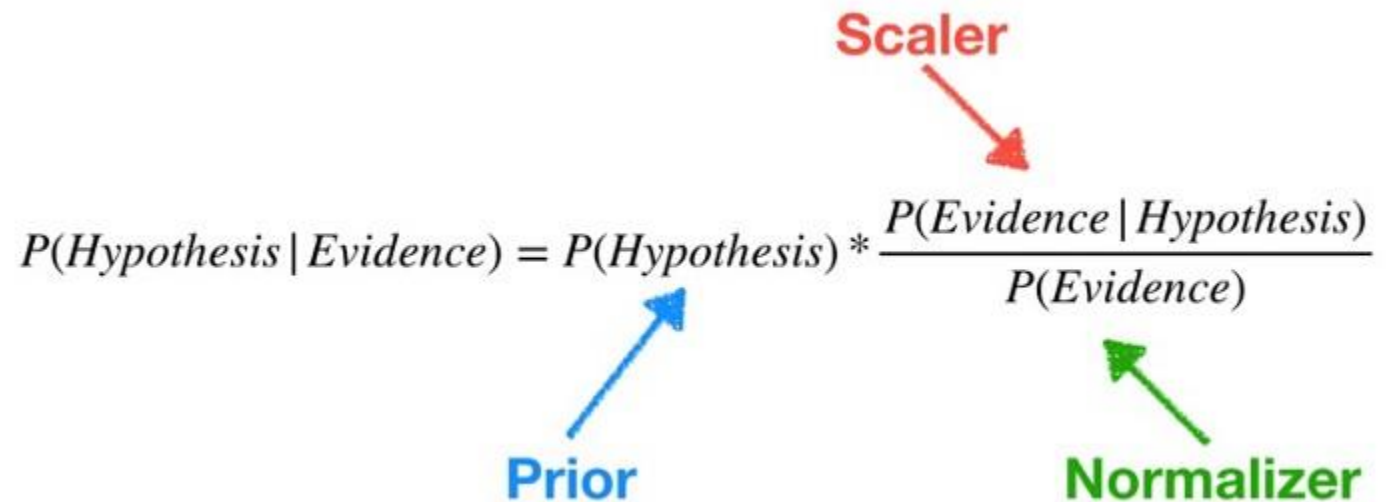
- This table shows that the event  $Y$  has two outcomes namely  $A$  and  $B$ , which is dependent on another event  $X$  with various outcomes like  $x_1$ ,  $x_2$  and  $x_3$ .
- **Case1:** Suppose, we don't have any information of the event  $A$ . Then, from the given sample space, we can calculate  $P(Y = A) = \frac{5}{10} = 0.5$ .
- **Case2:** Now, suppose, we want to calculate  $P(X = x_2/Y = A) = \frac{2}{5} = 0.4$ .

The later is the conditional or posterior probability, where as the former is the prior probability.

$X$	$Y$
$x_1$	$A$
$x_2$	$A$
$x_3$	$B$
$x_3$	$A$
$x_2$	$B$
$x_1$	$A$
$x_1$	$B$
$x_3$	$B$
$x_2$	$B$
$x_2$	$A$

# Example

- If we have to calculate the probability that there is fire given that there is smoke, then the following equation will be used:
- $P(\text{Fire}|\text{Smoke}) = P(\text{Smoke}|\text{Fire}) * P(\text{Fire}) / P(\text{Smoke})$
- Where,  $P(\text{Fire})$  is the Prior,  $P(\text{Smoke}|\text{Fire})$  is the Likelihood, and  $P(\text{Smoke})$  is the evidence.



The diagram shows the equation  $P(\text{Hypothesis} | \text{Evidence}) = P(\text{Hypothesis}) * \frac{P(\text{Evidence} | \text{Hypothesis})}{P(\text{Evidence})}$ . A blue arrow points from the word 'Prior' to  $P(\text{Hypothesis})$ . A red arrow points from the word 'Scaler' to the fraction  $\frac{P(\text{Evidence} | \text{Hypothesis})}{P(\text{Evidence})}$ . A green arrow points from the word 'Normalizer' to  $P(\text{Evidence})$  in the denominator.

$$P(\text{Hypothesis} | \text{Evidence}) = P(\text{Hypothesis}) * \frac{P(\text{Evidence} | \text{Hypothesis})}{P(\text{Evidence})}$$

Prior

Scaler

Normalizer

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

- $P(H)$  is the probability of hypothesis  $H$  being true. This is known as the prior probability.
- $P(E)$  is the probability of the evidence (regardless of the hypothesis).
- $P(E|H)$  is the probability of the evidence given that hypothesis is true.
- $P(H|E)$  is the probability of the hypothesis given that the evidence is there.

# Example 1:

- A bag I contains 4 white and 6 black balls while another Bag II contains 4 white and 3 black balls. One ball is drawn at random from one of the bags, and it is found to be black.
- Find the probability that it was drawn from Bag I.

# Example 1:

- A bag I contains 4 white and 6 black balls while another Bag II contains 4 white and 3 black balls. One ball is drawn at random from one of the bags, and it is found to be black.
- Find the probability that it was drawn from Bag I.

**Solution:**

Let  $E_1$  be the event of choosing bag I,  $E_2$  the event of choosing bag II, and A be the event of drawing a black ball.

$$\text{Then, } P(E_1) = P(E_2) = \frac{1}{2}$$

$$\text{Also, } P(A|E_1) = P(\text{drawing a black ball from Bag I}) = \frac{6}{10} = \frac{3}{5}$$

$$P(A|E_2) = P(\text{drawing a black ball from Bag II}) = \frac{3}{7}$$

By using Bayes' theorem, the probability of drawing a black ball from bag I out of two bags,

$$P(E_1|A) = \frac{P(E_1)P(A|E_1)}{P(E_1)P(A|E_1) + P(E_2)P(A|E_2)}$$

$$= \frac{\frac{1}{2} \times \frac{3}{5}}{\frac{1}{2} \times \frac{3}{5} + \frac{1}{2} \times \frac{3}{7}} = \frac{7}{12}$$

## Example 2:

- A man is known to speak the truth 2 out of 3 times. He throws a die and reports that the number obtained is a four.
- Find the probability that the number obtained is actually a four.

# Example 2:

- A man is known to speak the truth 2 out of 3 times. He throws a die and reports that the number obtained is a four.
- Find the probability that the number obtained is actually a four.

**Solution:**

Let  $A$  be the event that the man reports that number four is obtained.

Let  $E_1$  be the event that four is obtained and  $E_2$  be its complementary event.

Then,  $P(E_1)$  = Probability that four occurs =  $\frac{1}{6}$

$P(E_2)$  = Probability that four does not occur =  $1 - P(E_1) = 1 - \frac{1}{6} = \frac{5}{6}$

Also,  $P(A|E_1)$  = Probability that man reports four and it is actually a four =  $\frac{2}{3}$

$P(A|E_2)$  = Probability that man reports four and it is not a four =  $\frac{1}{3}$

By using Bayes' theorem, probability that number obtained is actually a four,

$$P(E_1|A) = \frac{P(E_1)P(A|E_1)}{P(E_1)P(A|E_1) + P(E_2)P(A|E_2)} = \frac{\frac{1}{6} \times \frac{2}{3}}{\frac{1}{6} \times \frac{2}{3} + \frac{5}{6} \times \frac{1}{3}} = \frac{2}{7}$$

# Practice Problems

- Solve the following problems using Bayes Theorem.
1. A bag contains 5 red and 5 black balls. A ball is drawn at random, its colour is noted, and again the ball is returned to the bag. Also, 2 additional balls of the colour drawn are put in the bag. After that, the ball is drawn at random from the bag. What is the probability that the second ball drawn from the bag is red?
  2. Of the students in the college, 60% of the students reside in the hostel and 40% of the students are day scholars. Previous year results report that 30% of all students who stay in the hostel scored A Grade and 20% of day scholars scored A grade. At the end of the year, one student is chosen at random and found that he/she has an A grade. What is the probability that the student is a hosteler?
  3. From the pack of 52 cards, one card is lost. From the remaining cards of a pack, two cards are drawn and both are found to be diamond cards. What is the probability that the lost card is a diamond?



# What is Naïve Bayes Classifier in Machine Learning

- Naïve Bayes theorem is also a supervised algorithm, which is based on Bayes theorem and used to solve classification problems.
- It is one of the most simple and effective classification algorithms in Machine Learning which enables us to build various ML models for quick predictions.
- It is a probabilistic classifier that means it predicts on the basis of probability of an object. Some popular Naïve Bayes algorithms are **spam filtration, Sentimental analysis, and classifying articles.**

- Naive Bayes uses the Bayes' Theorem and assumes that **all predictors are independent**.
- In other words, this classifier assumes that the presence of one particular feature in a class doesn't affect the presence of another one.
- **Here's an example:** you'd consider fruit to be orange if it is round, orange, and is of around 3.5 inches in diameter. Now, even if these features require each other to exist, they all contribute independently to your assumption that this particular fruit is orange. That's why this algorithm has '**Naive**' in its name.
- Building the Naive Bayes model is quite simple and helps you in working with vast datasets. Moreover, this equation is popular for beating many advanced classification techniques in terms of performance.

# Naive Bayes Classifier

- a supervised learning algorithm for classification so the task is to find the class of observation (data point) given the values of features.
- Naive bayes classifier calculates the probability of a class given a set of feature values (i.e.  $p(y_i | x_1, x_2, \dots, x_n)$ ).
- Input this into Bayes' theorem:

$$p(y_i | x_1, x_2, \dots, x_n) = \frac{p(x_1, x_2, \dots, x_n | y_i) \cdot p(y_i)}{p(x_1, x_2, \dots, x_n)}$$

Apart from considering the independence of every feature, Naive Bayes also assumes that they contribute equally. This is an important point to remember.

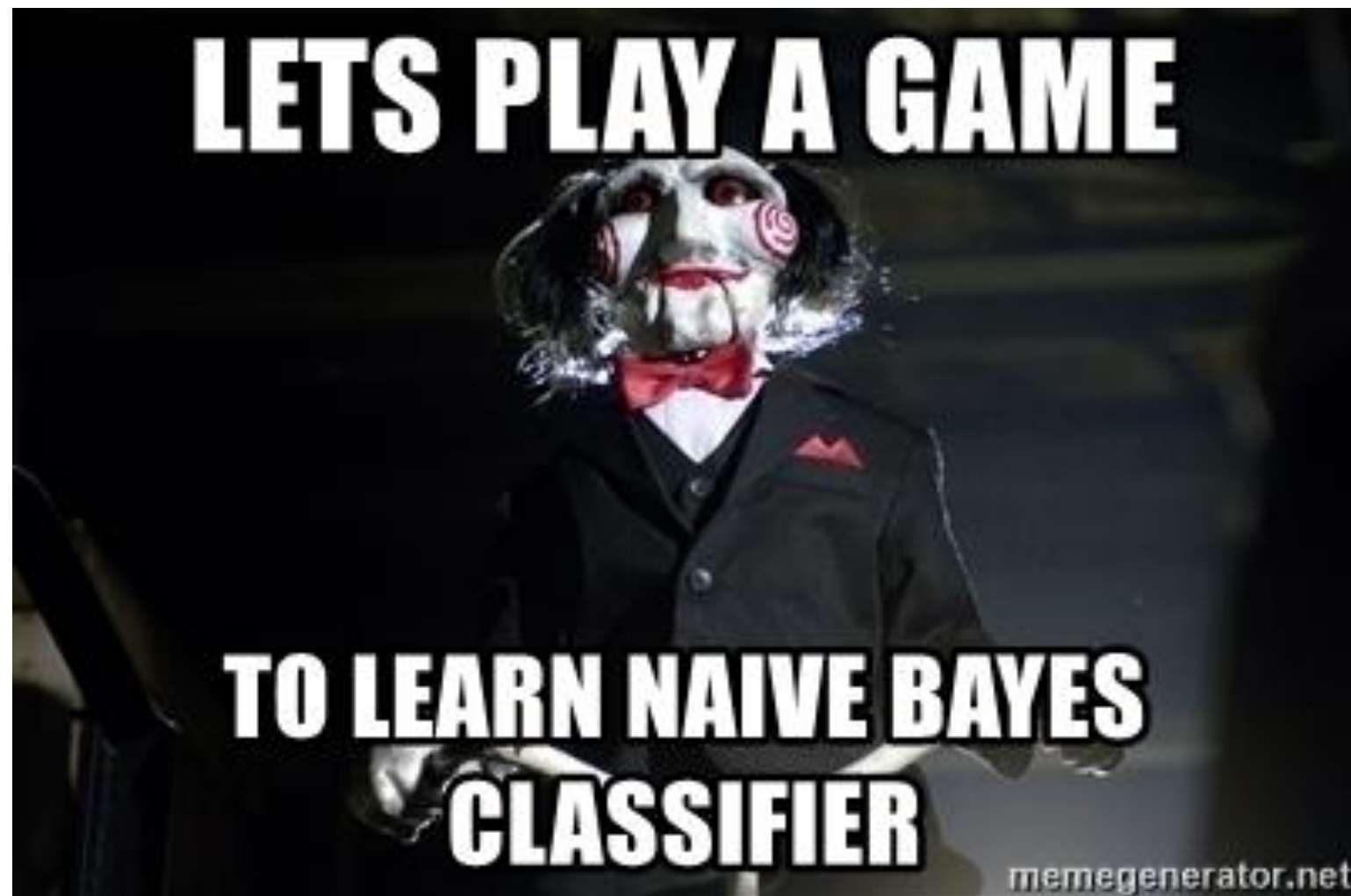
- $p(x_1, x_2, \dots, x_n | y_i)$  means the probability of a specific combination of features given a class label.
- To be able to calculate this, we need extremely large datasets to have an estimate on the probability distribution for all different combinations of feature values.
- To overcome this issue, **naive bayes algorithm assumes that all features are independent of each other.**
- Furthermore, denominator ( $p(x_1, x_2, \dots, x_n)$ ) can be removed to simplify the equation because it only normalizes the value of conditional probability of a class given an observation ( $p(y_i | x_1, x_2, \dots, x_n)$ ).

$$p(y_i) = \frac{\text{number of observations with class } y_i}{\text{number of all observations}}$$

- Under the assumption of features being independent,  $p(x_1, x_2, \dots, x_n | y_i)$  can be written as:

$$p(x_1, x_2, \dots, x_n | y_i) = p(x_1 | y_i) \cdot p(x_2 | y_i) \cdot \dots \cdot p(x_n | y_i)$$

- The conditional probability for a single feature given the class label (i.e.  $p(x_1 | y_i)$ ) can be more easily estimated from the data.
- The algorithm needs to store probability distributions of features for each class independently.
- For example, if there are 5 classes and 10 features, 50 different probability distributions need to be stored. The type of distributions depend on the characteristics of features:
  - For binary features (Y/N, True/False, 0/1): Bernoulli distribution
  - For discrete features (i.e. word counts): Multinomial distribution
  - For continuous features: Gaussian (Normal) distribution
  - It is common to name the naive bayes with the distribution of features (i.e. Gaussian naive bayes classifier). For mixed type datasets, a different type of distribution may be required for different features.



Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

- The posterior probability can be calculated by first, constructing a frequency table for each attribute against the target.
- Then, transforming the frequency tables to likelihood tables and finally use the Naive Bayesian equation to calculate the posterior probability for each class.
- The class with the highest posterior probability is the outcome of prediction.

Frequency Table		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3



Likelihood Table		Play Golf		
		Yes	No	
Outlook	Sunny	3/9	2/5	5/14
	Overcast	4/9	0/5	4/14
	Rainy	2/9	3/5	5/14
		9/14	5/14	

$$P(x | c) = P(\text{Sunny} | \text{Yes}) = 3 / 9 = 0.33$$



$$P(x) = P(\text{Sunny}) = 5 / 14 = 0.36$$



$$P(c) = P(\text{Yes}) = 9 / 14 = 0.64$$



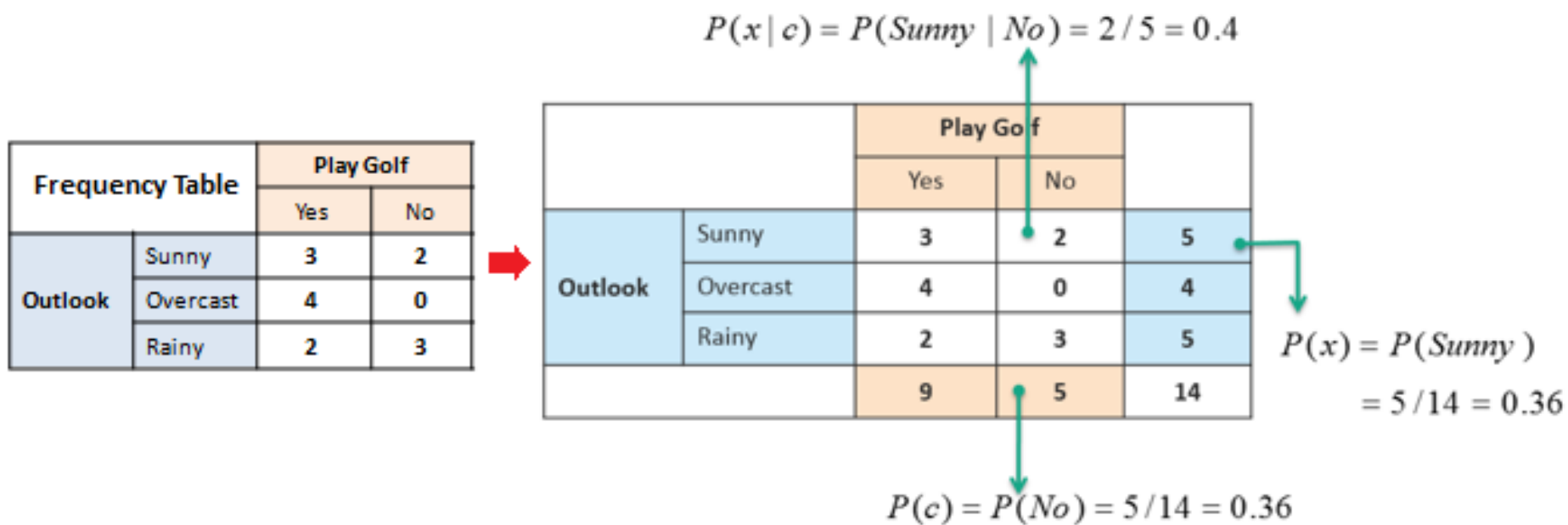
Posterior Probability:

$$P(c | x) = P(\text{Yes} | \text{Sunny}) = 0.33 \times 0.64 \div 0.36 = 0.60$$





■



Posterior Probability:

$$P(c | x) = P(\text{No} | \text{Sunny}) = 0.40 \times 0.36 \div 0.36 = 0.40$$

The likelihood tables for all four predictors.

Frequency Table

		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3



Likelihood Table

		Play Golf	
		Yes	No
Outlook	Sunny	3/9	2/5
	Overcast	4/9	0/5
	Rainy	2/9	3/5

		Play Golf	
		Yes	No
Humidity	High	3	4
	Normal	6	1



		Play Golf	
		Yes	No
Humidity	High	3/9	4/5
	Normal	6/9	1/5

		Play Golf	
		Yes	No
Temp.	Hot	2	2
	Mild	4	2
	Cool	3	1



		Play Golf	
		Yes	No
Temp.	Hot	2/9	2/5
	Mild	4/9	2/5
	Cool	3/9	1/5

		Play Golf	
		Yes	No
Windy	False	6	2
	True	3	3



		Play Golf	
		Yes	No
Windy	False	6/9	2/5
	True	3/9	3/5

Outlook	Temp	Humidity	Windy	Play
Rainy	Cool	High	True	?

$$P(Yes | X) = P(Rainy | Yes) \times P(Cool | Yes) \times P(High | Yes) \times P(True | Yes) \times P(Yes)$$

$$P(Yes | X) = 2/9 \times 3/9 \times 3/9 \times 3/9 \times 9/14 = 0.00529 \rightarrow 0.2 = \frac{0.00529}{0.02057 + 0.00529}$$

$$P(No | X) = P(Rainy | No) \times P(Cool | No) \times P(High | No) \times P(True | No) \times P(No)$$

$$P(No | X) = 3/5 \times 1/5 \times 4/5 \times 3/5 \times 5/14 = 0.02057 \rightarrow 0.8 = \frac{0.02057}{0.02057 + 0.00529}$$



Type	Long	Not Long	Sweet	Not Sweet	Yellow	Not Yellow	Total
Banana	400	100	350	150	450	50	500
Orange	0	300	150	150	300	0	300
Other	100	100	150	50	50	150	200
Total	500	500	650	350	800	200	1000

- predict if a given fruit is a ‘Banana’ or ‘Orange’ or ‘Other’ when only the 3 features (long, sweet and yellow) are known.

- **Step 1: Compute the ‘Prior’ probabilities for each of the class of fruits.** That is, the proportion of each fruit class out of all the fruits from the population.
  - $P(Y=\text{Banana}) = 500 / 1000 = 0.50$
  - $P(Y=\text{Orange}) = 300 / 1000 = 0.30$
  - $P(Y=\text{Other}) = 200 / 1000 = 0.20$
- **Step 2: Compute the probability of evidence that goes in the denominator.**
  - $P(x1=\text{Long}) = 500 / 1000 = 0.50$
  - $P(x2=\text{Sweet}) = 650 / 1000 = 0.65$
  - $P(x3=\text{Yellow}) = 800 / 1000 = 0.80$

- **Step 3: Compute the probability of likelihood of evidences that goes in the numerator.**
- $P(\text{Long} \mid \text{Banana}) = 400/500 = 0.8$ . Here, I have done it for Banana alone.
- $P(x_1=\text{Long} \mid Y=\text{Banana}) = 400 / 500 = 0.80$
- $P(x_2=\text{Sweet} \mid Y=\text{Banana}) = 350 / 500 = 0.70$
- $P(x_3=\text{Yellow} \mid Y=\text{Banana}) = 450 / 500 = 0.90$ .
- So, the overall probability of Likelihood of evidence for Banana =  $0.8 * 0.7 * 0.9 = 0.504$

- Step 4: Substitute all the 3 equations into the Naive Bayes formula, to get the probability that it is a banana.

Step 4: If a fruit is 'Long', 'Sweet' and 'Yellow', what fruit is it?

$$P(\text{Banana} \mid \text{Long, Sweet and Yellow}) = \frac{P(\text{Long} \mid \text{Banana}) * P(\text{Sweet} \mid \text{Banana}) * P(\text{Yellow} \mid \text{Banana}) * P(\text{banana})}{P(\text{Long}) * P(\text{Sweet}) * P(\text{Yellow})}$$

$$= \frac{0.8 * 0.7 * 0.9 * 0.5}{P(\text{Evidence})} = 0.252 / P(\text{Evidence})$$

$$P(\text{Orange} \mid \text{Long, Sweet and Yellow}) = 0, \text{ because } P(\text{Long} \mid \text{Orange}) = 0$$

$$P(\text{Other Fruit} \mid \text{Long, Sweet and Yellow}) = 0.01875 / P(\text{Evidence})$$

Answer: Banana - Since it has highest probability amongst the 3 classes

# How does Naive Bayes Work?

## Example 3

- To understand how Naive Bayes works, we should discuss an example.
- Suppose we want to find stolen cars and have the following dataset:

Serial No.	Color	Type	Origin	Was it Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes



According to our dataset, we can understand that our algorithm makes the following assumptions:

- It assumes that every feature is independent. For example, the colour 'Yellow' of a car has nothing to do with its Origin or Type.
- It gives every feature the same level of importance. For example, knowing only the Color and Origin would predict the outcome correctly. That's why every feature is equally important and contributes equally to the result.

Now, with our dataset, we have to classify if thieves steal a car according to its features. Each row has individual entries, and the columns represent the features of every car. In the first row, we have a stolen Red Sports Car with Domestic Origin. We'll find out if thieves would steal a Red Domestic SUV or not (our dataset doesn't have an entry for a Red Domestic SUV).

Frequency Table

		Stolen?	
		Yes	No
Color	Red	3	2
	Yellow	2	3



Likelihood Table

		Stolen?	
		P(Yes)	P(No)
Color	Red	3/5	2/5
	Yellow	2/5	3/5

Frequency Table

		Stolen?	
		Yes	No
Type	Sports	4	2
	SUV	1	3



Likelihood Table

		Stolen?	
		P(Yes)	P(No)
Type	Sports	4/5	2/5
	SUV	1/5	3/5

Frequency Table

		Stolen?	
		Yes	No
Origin	Domestic	2	3
	Imported	3	2



Likelihood Table

		Stolen?	
		P(Yes)	P(No)
Origin	Domestic	$2/5$	$3/5$
	Imported	$3/5$	$2/5$

Color	Type	Origin	Stolen
Red	SUV	Domestic	?

# Example-4

## Example 8.4

Class:

C1:buys\_computer = 'yes'

C2:buys\_computer = 'no'

Data instance

X = (age <=30,

Income = medium,

Student = yes

Credit\_rating = fair)

age	income	student	credit_rating	comp
<=30	high	no	fair	no
<=30	high	no	excellent	no
31...40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
>40	medium	no	excellent	no

# A Practice Example

- $P(C_i)$ :  $P(\text{buys\_computer} = \text{"yes"}) = 9/14 = 0.643$   
 $P(\text{buys\_computer} = \text{"no"}) = 5/14 = 0.357$
- Compute  $P(X|C_i)$  for each class  
 $P(\text{age} = \text{"<=30"} | \text{buys\_computer} = \text{"yes"}) = 2/9 = 0.222$   
 $P(\text{age} = \text{"<= 30"} | \text{buys\_computer} = \text{"no"}) = 3/5 = 0.6$   
 $P(\text{income} = \text{"medium"} | \text{buys\_computer} = \text{"yes"}) = 4/9 = 0.444$   
 $P(\text{income} = \text{"medium"} | \text{buys\_computer} = \text{"no"}) = 2/5 = 0.4$   
 $P(\text{student} = \text{"yes"} | \text{buys\_computer} = \text{"yes"}) = 6/9 = 0.667$   
 $P(\text{student} = \text{"yes"} | \text{buys\_computer} = \text{"no"}) = 1/5 = 0.2$   
 $P(\text{credit\_rating} = \text{"fair"} | \text{buys\_computer} = \text{"yes"}) = 6/9 = 0.667$   
 $P(\text{credit\_rating} = \text{"fair"} | \text{buys\_computer} = \text{"no"}) = 2/5 = 0.4$
- **$X = (\text{age} \leq 30, \text{income} = \text{medium}, \text{student} = \text{yes}, \text{credit\_rating} = \text{fair})$**   
  
 $P(X|C_i) : P(X|\text{buys\_computer} = \text{"yes"}) = 0.222 \times 0.444 \times 0.667 \times 0.667 = 0.044$   
 $P(X|\text{buys\_computer} = \text{"no"}) = 0.6 \times 0.4 \times 0.2 \times 0.4 = 0.019$   
  
 $P(X|C_i) * P(C_i) : P(X|\text{buys\_computer} = \text{"yes"}) * P(\text{buys\_computer} = \text{"yes"}) = 0.028$   
 $P(X|\text{buys\_computer} = \text{"no"}) * P(\text{buys\_computer} = \text{"no"}) = 0.007$   
  
**Therefore, X belongs to class ("buys\_computer = yes")**

# Advantages of Naïve Bayes Classifier in Machine Learning:

- It is one of the simplest and effective methods for calculating the conditional probability and text classification problems.
- A Naïve-Bayes classifier algorithm is better than all other models where assumption of independent predictors holds true.
- It is easy to implement than other models.
- It requires small amount of training data to estimate the test data which minimize the training time period.
- It can be used for Binary as well as Multi-class Classifications.

# Disadvantages of Naïve Bayes Classifier in Machine Learning:

- The main disadvantage of using Naïve Bayes classifier algorithms is, it limits the assumption of independent predictors because it implicitly assumes that all attributes are independent or unrelated but in real life it is not feasible to get mutually independent attributes.





- <https://www.upgrad.com/blog/bayes-theorem-in-machine-learning/#:~:text=Bayes%20Theorem%20is%20a%20method,contribute%20to%20more%20accurate%20results.>
- <https://machinelearningmastery.com/bayes-theorem-for-machine-learning/>
- <https://byjus.com/maths/bayes-theorem/>
- <https://www.javatpoint.com/bayes-theorem-in-machine-learning>
- <https://towardsdatascience.com/bayes-theorem-the-core-of-machine-learning-69f5703e511f>