



Machine Learning
ICT-4261

By-
Dr. Jesmin Akhter

Professor
Institute of Information Technology
Jahangirnagar University

Contents

The course will mainly cover the following topics:

- ✓ A Gentle Introduction to Machine Learning
- ✓ Logistic Regression
- ✓ Naive Bayes
- ✓ Support Vector Machines
- ✓ Decision Trees and Ensemble Learning
- ✓ Clustering Fundamentals
- ✓ Hierarchical Clustering
- ✓ Neural Networks and Deep Learning
- ✓ Unsupervised Learning

Outline

- ✓ Naive Bayes
 - Bayes' theorem

Bayes' theorem

What is conditional probability?

- A conditional probability is usually defined as the probability of one event given the occurrence of another event.
 - If A and B are two events, then the conditional probability may be designated as P(A given B) or P(A|B).
- ✓ Bayes's theorem is used for the calculation of a conditional probability where intuition often fails.
- Let's consider two probabilistic events A and B. We can correlate the marginal probabilities P(A) and P(B) with the conditional probabilities P(A|B) and P(B|A) using the product rule:

$$\begin{cases} P(A \cap B) = P(A|B)P(B) \\ P(B \cap A) = P(B|A)P(A) \end{cases}$$

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

Bayes' theorem

- ✓ Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes' theorem is stated mathematically as the following equation:

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

- ✓ where A and B are two probabilistic events and $P(B) \neq 0$.
- ✓ Basically, we are trying to find probability of event A, given the event B is true. Event B is also termed as **evidence**.
- ✓ $P(A)$ is the **priori** of A (the prior probability, i.e. Probability of event before evidence is seen). The evidence is an attribute value of an unknown instance(here, it is event B).
- ✓ $P(A|B)$ is a posteriori probability of B, i.e. probability of event after evidence is seen.

Bayes' theorem-Problem 1

- ✓ Imagine we want to implement a very simple spam filter and we've collected 100 emails. We know that 30 are spam and 70 are regular. So we can say that $P(\text{Spam}) = 0.3$.
- ✓ However, we'd like to evaluate using some criteria, for example, email text is shorter than 50 characters. Therefore, our query becomes:

$$P(\text{Spam}|\text{Text} < 50 \text{ chars}) = \frac{P(\text{Text} < 50 \text{ chars}|\text{Spam})P(\text{Spam})}{P(\text{Text} < 50 \text{ chars})}$$

- ✓ The first term is similar to $P(\text{Spam})$ because it's the probability of spam given a certain condition. For this reason, it's called a posteriori (in other words, it's a probability that we can estimate after knowing some additional elements). On the right-hand side, we need to calculate the missing values,
- ✓ Let's suppose that 35 emails have text shorter than 50 characters, so $P(\text{Text} < 50 \text{ chars}) = 0.35$. Looking only into our spam folder, we discover that **only 25 spam emails have short text, so that $P(\text{Text} < 50 \text{ chars}|\text{Spam}) = 25/30 = 0.83$** . The result is:

$$P(\text{Spam}|\text{Text} < 50 \text{ chars}) = \frac{0.83 \cdot 0.3}{0.35} = 0.71$$

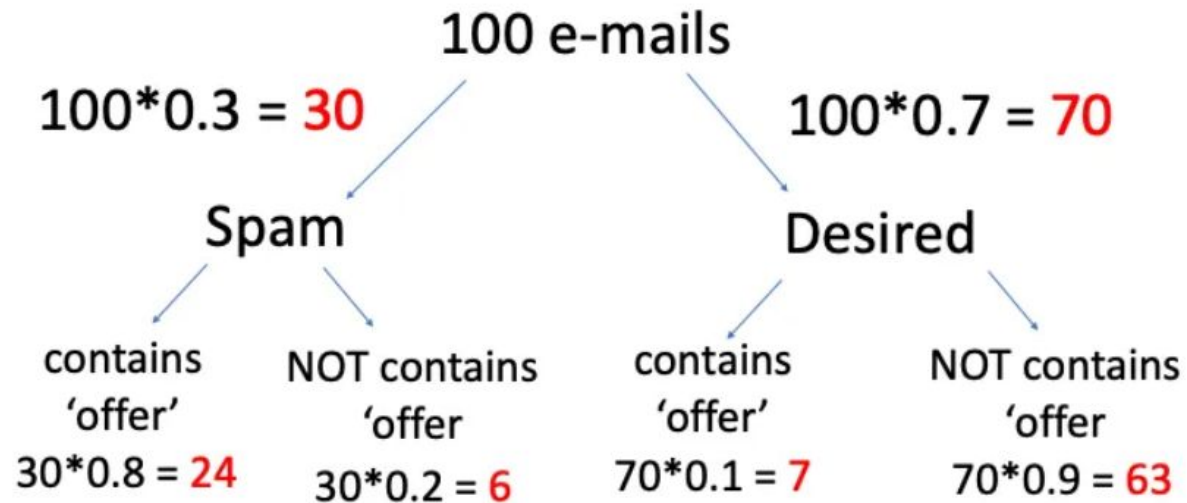
- ✓ So, after receiving a very short email, there is a 71% probability that it's spam.

Problem 2

- ✓ Assume that the word 'offer' occurs in 80% of the spam messages in your account. Also, let's assume 'offer' occurs in 10% of your desired e-mails. If 30% of the received e-mails are considered as a scam, and you will receive a new message which contains 'offer', what is the probability that it is spam?

Explanation without Bayes' Equation

- ✓ Assume total 100 e-mails.
- ✓ The percentage of spam in the whole e-mail is 30%.
- ✓ So, 30 spam e-mails and 70 desired e-mails in 100 e-mails.
- ✓ The percentage of the **word 'offer' that occurs in spam e-mails** is 80%.
 - It means 80% of 30 spam e-mail and it makes 24 spam email contain the word offer where 6 of them does not contain 'offer'.
- ✓ The percentage of the **word 'offer' that occurs in the desired e-mails** is 10%.
 - It means 7 of them (10% of 70 desired e-mails) contain the word 'offer' and 63 of them not.
- ✓ Now, we can see this logic in a simple chart.



- ✓ The question was what is the probability of spam where the mail contains the word 'offer':
- ✓ We need to find the total number of mails which contains 'offer' ;
- ✓ $24 + 7 = 31$ mail contain the word 'offer'
- ✓ 2. Find the probability of spam if the mail contains 'offer' ;
- ✓ In 31 mails 24 contains 'offer' means $77.4\% = 0.774$ (probability)

Solution with Bayes' Equation

- ✓ A = Spam
- ✓ B = Contains the word 'offer'

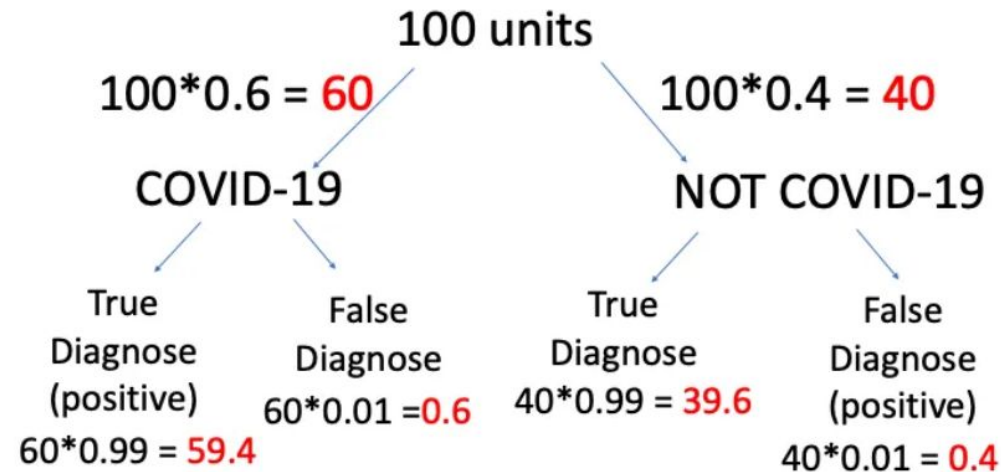
$$P(\text{spam}|\text{contains offer}) = \frac{P(\text{contains offer}|\text{spam}) * P(\text{spam})}{P(\text{contains offer})}$$

- ✓ $P(\text{contains offer}|\text{spam}) = 0.8$ (given in the question)
- ✓ $P(\text{spam}) = 0.3$ (given in the question)
- ✓ Now we will find the probability of e-mail with the word 'offer'. We can compute that by adding 'offer' in spam and desired e-mails. Such that;
- ✓ $P(\text{contains offer}) = 0.3*0.8 + 0.7*0.1 = 0.31$

$$P(\text{spam}|\text{contains offer}) = \frac{0.8 * 0.3}{0.31} = 0.774$$

Problem 3

- ✓ As you know, Covid-19 tests are common nowadays, but some results of tests are not true. Let's assume; a diagnostic test has 99% accuracy and 60% of all people have Covid-19. If a patient tests positive, what is the probability that they actually have the disease?



Solution with Bayes' Equation

- ✓ The total units which have positive results= $59.4 + 0.4 = 59.8$
- ✓ 59.4 units (true positive) is 59.8 units means $99.3\% = 0.993$ probability
- ✓ **With Bayes';**

$$P(\text{covid19}|\text{positive}) = \frac{P(\text{positive}|\text{covid19}) * P(\text{covid19})}{P(\text{positive})}$$

- ✓ $P(\text{positive}|\text{covid19}) = 0.99$
- ✓ $P(\text{covid19}) = 0.6$
- ✓ $P(\text{positive}) = 0.6*0.99+0.4*0.01=0.598$
- ✓

$$P(\text{covid19}|\text{positive}) = \frac{0.99 * 0.6}{0.598} = 0.993$$

Naive Bayes

- ✓ Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the “naive” assumption of conditional independence between every pair of features given the value of the class variable. Bayes' theorem states the following relationship, given class variable y and feature vector x_1 through x_n
- ✓ Given a dependent(class) variable y and n dependent features:-

$$P(y | x_1, \dots, x_n) = \frac{P(y)P(x_1, \dots, x_n | y)}{P(x_1, \dots, x_n)}$$

- ✓ Now using the naive independence assumption that all features are mutually independent represented by the equation:-

$$P(x_i | y, x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) = P(x_i | y),$$

- ✓ for all i , the relationship is simplified to the following equation:-

$$\begin{aligned} P(y | x_1, x_2, \dots, x_n) &= \frac{P(x_1, x_2, \dots, x_n | y)P(y)}{P(x_1, x_2, \dots, x_n)} \\ &= \frac{P(x_1 | y)P(x_2 | y) \dots P(x_n | y)P(y)}{P(x_1, x_2, \dots, x_n)} & P(y | x_1, \dots, x_n) &= \frac{P(y) \prod_{i=1}^n P(x_i | y)}{P(x_1, \dots, x_n)} \\ &\propto P(x_1 | y)P(x_2 | y) \dots P(x_n | y)P(y) \end{aligned}$$

Naive Bayes

- ✓ Now, we need to create a classifier model. For this, we find the probability of given set of inputs for all possible values of the class variable y and pick up the output with maximum probability. This can be expressed mathematically as:

Since $P(x_1, \dots, x_n)$ is constant given the input, we can use the following classification rule:

$$P(y \mid x_1, \dots, x_n) \propto P(y) \prod_{i=1}^n P(x_i \mid y)$$

\Downarrow

$$\hat{y} = \arg \max_y P(y) \prod_{i=1}^n P(x_i \mid y),$$

$P(y)$ is also called **class probability** and $P(x_i \mid y)$ is called **conditional probability**

Why is it called Naïve Bayes?

- ✓ The Naïve Bayes algorithm is comprised of two words Naïve and Bayes, Which can be described as:
- ✓ **Naïve:** It is called Naïve because it assumes that the occurrence of a **certain feature is independent of the occurrence of other features**. This means the **presence or absence of one feature doesn't influence the presence or absence of another feature**, considering the class is already known. In reality, this assumption isn't always true.
 - Such as if the fruit is identified on the bases of color, shape, and taste, then red, spherical, and sweet fruit is recognized as an apple. Hence each feature individually contributes to identify that it is an apple without depending on each other.
- ✓ **Bayes:** It is called Bayes because it depends on the principle of Bayes' Theorem.

Estimate Prior and Conditional probability

Obs	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
X1	A	A	A	A	A	B	B	B	B	B	C	C	C	C	C
X2	S	M	M	S	S	S	M	M	L	L	L	M	M	L	L
Y	0	0	1	1	0	0	0	1	1	1	1	1	1	1	0

✓ Use formula above to estimate prior and conditional probability, and we can get:

Prior Probability

$$P(Y=1) = 9/15 \quad P(Y=0)=6/15$$

Conditional Probability

$$P(X1=A|Y=1)=2/9 \quad P(X1=B|Y=1)=3/9 \quad P(X1=C|Y=1)=4/9$$

$$P(X2=S|Y=1)=1/9 \quad P(X2=M|Y=1)=4/9 \quad P(X2=L|Y=1)=4/9$$

$$P(X1=A|Y=0)=3/6 \quad P(X1=B|Y=0)=2/6 \quad P(X1=C|Y=0)=1/6$$

$$P(X2=S|Y=0)=3/6 \quad P(X2=M|Y=0)=2/6 \quad P(X2=L|Y=0)=1/6$$

$$P(Y=1)P(X1=B|Y=1)P(X2=S|Y=1)=1/45$$

$$P(Y=0)P(X1=B|Y=0)P(X2=S|Y=0)=1/15$$

$$P(Y=0)P(X1=B|Y=0)P(X2=S|Y=0) > P(Y=1)P(X1=B|Y=1)P(X2=S|Y=1), \text{ so } y=0.$$

Obs	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
X1	A	A	A	A	A	B	B	B	B	B	C	C	C	C	C
X2	S	M	M	S	S	S	M	M	L	L	L	M	M	L	L
Y	0	0	1	1	0	0	0	1	1	1	1	1	1	1	0

Tabular representation of an example dataset

- ✓ The dataset is divided into two parts, namely, **feature matrix** and the **response vector**.
- Feature matrix contains all the vectors(rows) of dataset. In the dataset, features are 'Outlook', 'Temperature', 'Humidity' and 'Windy'. Features are assumed to be **independent**.
 - Response vector contains the value of **class variable**(prediction or output) for each row of feature matrix. In the dataset, the class variable name is 'Play golf'

	Outlook	Temperature	Humidity	Windy	Play Golf
0	Rainy	Hot	High	False	No
1	Rainy	Hot	High	True	No
2	Overcast	Hot	High	False	Yes
3	Sunny	Mild	High	False	Yes
4	Sunny	Cool	Normal	False	Yes
5	Sunny	Cool	Normal	True	No
6	Overcast	Cool	Normal	True	Yes
7	Rainy	Mild	High	False	No
8	Rainy	Cool	Normal	False	Yes
9	Sunny	Mild	Normal	False	Yes
10	Rainy	Mild	Normal	True	Yes
11	Overcast	Mild	High	True	Yes
12	Overcast	Hot	Normal	False	Yes
13	Sunny	Mild	High	True	No

Now, with regards to our dataset, we can apply Bayes' theorem in following way:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

where, y is class variable and X is a dependent feature vector (of size n) where:

$$X = (x_1, x_2, x_3, \dots, x_n)$$

Just to clear, an example of a feature vector and corresponding class variable can be: (refer 1st row of dataset)

```
X = (Rainy, Hot, High, False)
```

```
y = No
```

So basically, $P(y|X)$ here means, the probability of “Not playing golf” given that the weather conditions are “Rainy outlook”, “Temperature is hot”, “high humidity” and “no wind”.

- ✓ With relation to our dataset, this concept can be understood as:
 - We assume that no pair of features are dependent. For example, the temperature being ‘Hot’ has nothing to do with the humidity or the outlook being ‘Rainy’ has no effect on the winds. Hence, the features are assumed to be **independent**.
 - Secondly, each feature is given the same weight(or importance). For example, knowing only temperature and humidity alone can't predict the outcome accurately. None of the attributes is irrelevant and assumed to be contributing **equally** to the outcome.

Naive assumption

Example 1

- ✓ 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	

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

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

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

Posterior Probability:

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

Naive assumption

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



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

$$P(x|c) = P(\text{Sunny} | \text{No}) = 2 / 5 = 0.4$$

$$P(c) = P(\text{No}) = 5 / 14 = 0.36$$

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

Posterior Probability:

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



Naive assumption

- ✓ 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

Naive assumption

Example 2

- ✓ In this example we have 4 inputs (predictors). The final posterior probabilities can be standardized between 0 and 1.

Outlook	Temp	Humidity	Windy	Play
Rainy	Cool	High	True	?

$$P(\text{Yes} \mid X) = P(\text{Rainy} \mid \text{Yes}) \times P(\text{Cool} \mid \text{Yes}) \times P(\text{High} \mid \text{Yes}) \times P(\text{True} \mid \text{Yes}) \times P(\text{Yes})$$

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

$$P(\text{No} \mid X) = P(\text{Rainy} \mid \text{No}) \times P(\text{Cool} \mid \text{No}) \times P(\text{High} \mid \text{No}) \times P(\text{True} \mid \text{No}) \times P(\text{No})$$

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

	Outlook	Temperature	Humidity	Windy	Play Golf
0	Rainy	Hot	High	False	No
1	Rainy	Hot	High	True	No
2	Overcast	Hot	High	False	Yes
3	Sunny	Mild	High	False	Yes
4	Sunny	Cool	Normal	False	Yes
5	Sunny	Cool	Normal	True	No
6	Overcast	Cool	Normal	True	Yes
7	Rainy	Mild	High	False	No
8	Rainy	Cool	Normal	False	Yes
9	Sunny	Mild	Normal	False	Yes
10	Rainy	Mild	Normal	True	Yes
11	Overcast	Mild	High	True	Yes
12	Overcast	Hot	Normal	False	Yes
13	Sunny	Mild	High	True	No

Naive assumption

Example 3

- ✓ Let us try to apply the Naive formula manually on our weather dataset.
- ✓ We need to find $P(x_i | y_j)$ for each x_i in X and y_j in y .
- ✓ So, in the figure, we have calculated $P(x_i | y_j)$ for each x_i in X and y_j in y manually in the tables 1-4. For example, probability of playing golf given that the temperature is cool, i.e $P(\text{temp.} = \text{cool} | \text{play golf} = \text{Yes}) = 3/9$.
- ✓ Also, we need to find class probabilities ($P(y)$) which has been calculated in the table 5. For example, $P(\text{play golf} = \text{Yes}) = 9/14$.

Outlook					Temperature				
	Yes	No	P(yes)	P(no)		Yes	No	P(yes)	P(no)
Sunny	2	3	2/9	3/5	Hot	2	2	2/9	2/5
Overcast	4	0	4/9	0/5	Mild	4	2	4/9	2/5
Rainy	3	2	3/9	2/5	Cool	3	1	3/9	1/5
Total	9	5	100%	100%	Total	9	5	100%	100%

Humidity					Wind				
	Yes	No	P(yes)	P(no)		Yes	No	P(yes)	P(no)
High	3	4	3/9	4/5	False	6	2	6/9	2/5
Normal	6	1	6/9	1/5	True	3	3	3/9	3/5
Total	9	5	100%	100%	Total	9	5	100%	100%

Play		P(Yes)/P(No)
Yes	9	9/14
No	5	5/14
Total	14	100%

Naive assumption

- ✓ So now, we are done with our pre-computations and the classifier is ready!
- ✓ Let us test it on a new set of features (let us call it today):
- ✓ today = (Sunny, Hot, Normal, False)
- ✓ So, probability of playing golf is given by:

$$P(Yes|today) = \frac{P(SunnyOutlook|Yes)P(HotTemperature|Yes)P(NormalHumidity|Yes)P(NoWind|Yes)P(Yes)}{P(today)}$$

- ✓ and probability to not play golf is given by:

$$P(No|today) = \frac{P(SunnyOutlook|No)P(HotTemperature|No)P(NormalHumidity|No)P(NoWind|No)P(No)}{P(today)}$$

Naive assumption

- ✓ Since, $P(\text{today})$ is common in both probabilities, we can ignore $P(\text{today})$ and find proportional probabilities as:

$$P(\text{Yes}|\text{today}) \propto \frac{2}{9} \cdot \frac{2}{9} \cdot \frac{6}{9} \cdot \frac{6}{9} \cdot \frac{9}{14} \approx 0.0141$$

and

$$P(\text{No}|\text{today}) \propto \frac{3}{5} \cdot \frac{2}{5} \cdot \frac{1}{5} \cdot \frac{2}{5} \cdot \frac{5}{14} \approx 0.0068$$

Now, since

$$P(\text{Yes}|\text{today}) + P(\text{No}|\text{today}) = 1$$

Naive assumption

- ✓ These numbers can be converted into a probability by making the sum equal to 1 (normalization):

$$P(Yes|today) = \frac{0.0141}{0.0141+0.0068} = 0.67$$

and

$$P(No|today) = \frac{0.0068}{0.0141+0.0068} = 0.33$$

Since

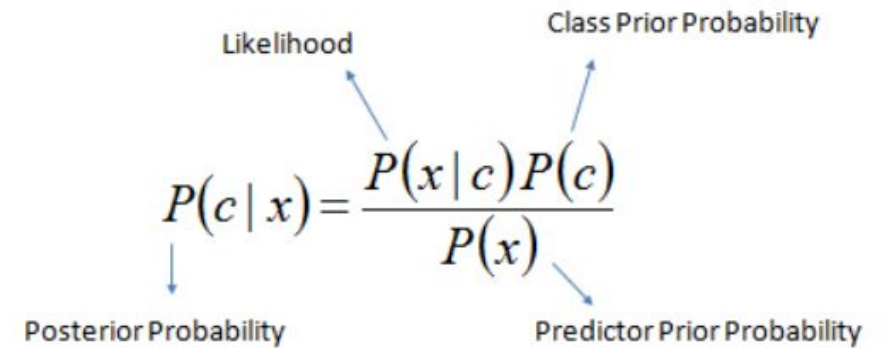
$$P(Yes|today) > P(No|today)$$

- ✓ So, prediction that golf would be played is 'Yes'.
- ✓ The method that we discussed above is applicable for discrete data. In case of continuous data, we need to make some assumptions regarding the distribution of values of each feature. The different naive Bayes classifiers differ mainly by the assumptions they make regarding the distribution of $P(x_i | y)$.

Thank You

Bayes' theorem

- ✓ The Naive Bayesian classifier is based on Bayes' theorem with the independence assumptions between predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.
- ✓ A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods.
- ✓ **Algorithm**
- ✓ Bayes theorem provides a way of calculating the posterior probability, $P(c|x)$, from $P(c)$, $P(x)$, and $P(x/c)$. Naive Bayes classifier assume that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence.



The diagram shows the Bayes' theorem formula $P(c|x) = \frac{P(x|c)P(c)}{P(x)}$ with four labels and arrows: 'Likelihood' points to $P(x|c)$, 'Class Prior Probability' points to $P(c)$, 'Posterior Probability' points to $P(c|x)$, and 'Predictor Prior Probability' points to $P(x)$.

$$P(c | X) = P(x_1 | c) \times P(x_2 | c) \times \dots \times P(x_n | c) \times P(c)$$

- $P(c|x)$ is the posterior probability of *class (target)* given *predictor (attribute)*.
- $P(c)$ is the prior probability of *class*.
- $P(x|c)$ is the likelihood which is the probability of *predictor* given *class*.
- $P(x)$ is the prior probability of *predictor*.
- In ZeroR model there is no predictor, in OneR model we find the single best predictor, naive Bayesian includes all predictors using Bayes' rule and the independence assumption between predictors.