Probability

Data and probability are inseperable. Data is computational side of story whereas probability is the theoretical side of the story

Probability needs to address the practical problems. Probability address the prior knownledge of data and how to compute the confidence interval of an estimate.

Probability is a measure of size of a set. Probability telling us about the relative frequency of the data.

Example:

Probability of getting even number when rolling a die.

probability space:

- **Sample Space** Ω: The set of all possible outcomes from an experiment.
- **Event Space** $\mathcal{F}$: The collection of all possible events. An event E is a subset in Ω that defines an outcome or a combination of outcomes.
- **Probability Law** $\mathbb{P}$: A mapping from an event E to a number $\mathbb{P}[E]$ which, ideally, should measure the size of the event.

It is the collection of all possible states that can be drawn from an experiment. The probability law is the interface with the data analysis

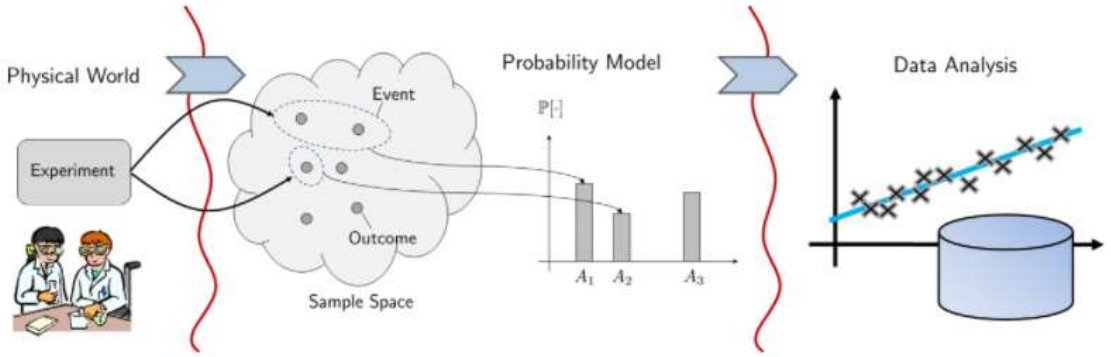


Figure 2.12: Given an experiment, we define the collection of all outcomes as the sample space. A subset in the sample space is called an event. The probability law is a mapping that maps an event to a number. This number denotes the size of the event.

Sample Space: Set of all possible outcomes from an experiment is a Sample space. A Sample space can contain discrete or continuous outcomes.

Example 1: (Discrete Outcomes)

- Coin flip: $\Omega = \{H,\ T\}$.
- Throw a dice: $\Omega = \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}$.
- Paper / scissor / stone: $\Omega = \{\text{paper}, \text{scissor}, \text{stone}\}$.
- Draw an even integer: $\Omega = \{2, 4, 6, 8, \ldots\}$.

In the last example, we see that a sample space can be infinite.

Example 2: (Continuous Outcomes)

- Waiting time for a bus in West Lafayette: $\Omega = \{t \mid 0 \le t \le 30 \text{ minutes}\}$.
- Phase angle of a voltage: $\Omega = \{\theta \mid 0 \le \theta \le 2\pi\}$.
- Frequency of a pitch: $\Omega = \{f \mid 0 \le f \le f_{\max}\}$.

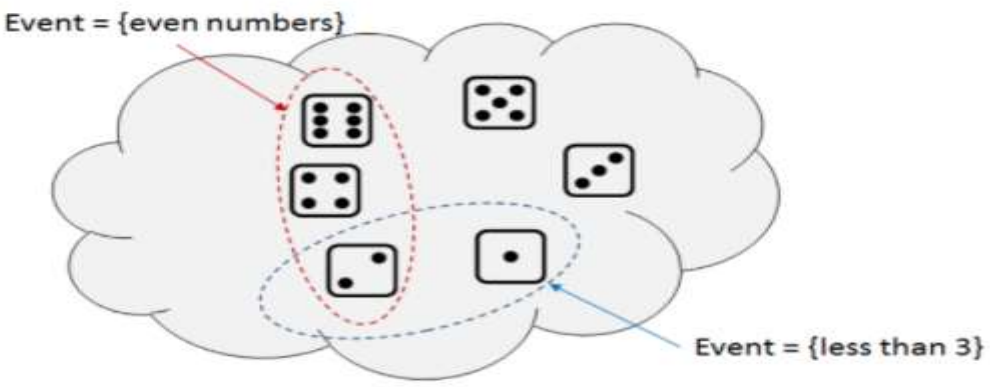
Practice Exercise. There are 8 processors on a computer. A computer job scheduler chooses one processor randomly. What is the sample space? If the computer job schedular can choose two processors at once, what is the sample space then?

Solution. The sample space of the first case is $\Omega = \{1, 2, 3, 4, 5, 6, 7, 8\}$. The sample space of the second case is $\Omega = \{(1,2), (1,3), (1,4), \ldots, (7,8)\}$.

Definition 2.14. *An **event** E is a subset in the sample space Ω. The set of all possible events is denoted as $\mathcal{F}$.*

Example 1. Throw a dice. Let $\Omega = \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}$. The followings are two possible events, as illustrated in Figure 2.14.

- $E_1 = \{\text{even numbers}\} = \{\boxdot, \boxdot, \boxdot\}$.
- $E_2 = \{\text{less than } 3\} = \{\boxdot, \boxdot\}$.



Definition 2.18. *A **probability law** is a function $\mathbb{P} : \mathcal{F} \to [0, 1]$ that maps an event E to a real number in $[0, 1]$.*

Example 1. Consider flipping a coin. The event space $\mathcal{F} = \{\emptyset, \{H\}, \{T\}, \Omega\}$. We can define the probability law as

$$\mathbb{P}[\emptyset] = 0, \;\; \mathbb{P}[\{H\}] = \frac{1}{2}, \;\; \mathbb{P}[\{T\}] = \frac{1}{2}, \;\; \mathbb{P}[\Omega] = 1,$$

as shown in Figure 2.16. This $\mathbb{P}$ is clearly consistent for all the events in $\mathcal{F}$.

Is it possible to construct an invalid $\mathbb{P}$? Certainly. Consider the following probability law:

$$\mathbb{P}[\emptyset] = 0, \;\; \mathbb{P}[\{H\}] = \frac{1}{3}, \;\; \mathbb{P}[\{T\}] = \frac{1}{3}, \;\; \mathbb{P}[\Omega] = 1.$$

This law is invalid because the individual events $\mathbb{P}[\{H\}] = \frac{1}{3}$ and $\mathbb{P}[\{T\}] = \frac{1}{3}$ but the union $\mathbb{P}[\Omega] = 1$. To fix this problem, one possible solution is to define the probability law as

$$\mathbb{P}[\emptyset] = 0, \;\; \mathbb{P}[\{H\}] = \frac{1}{3}, \;\; \mathbb{P}[\{T\}] = \frac{2}{3}, \;\; \mathbb{P}[\Omega] = 1.$$

Then, the probabilities for all the events are well defined and consistent.

What is a probability law $\mathbb{P}$?

- A probability law $\mathbb{P}$ is a **function**.
- It takes a subset (an element in $\mathcal{F}$) and maps it to a number between 0 and 1.
- $\mathbb{P}$ is a **measure**. It measures the size of a set.
- For $\mathbb{P}$ to make sense, it must satisfy the **axioms of probability**.

Definition 2.21. *A **probability law** is a function $\mathbb{P} : \mathcal{F} \to [0, 1]$ that maps an event A to a real number in $[0, 1]$. The function must satisfy the **axioms of probability**:*

I. Non-negativity: $\mathbb{P}[A] \geq 0$, for any $A \subseteq \Omega$.

II. Normalization: $\mathbb{P}[\Omega] = 1$.

III. Additivity: For any disjoint sets $\{A_1, A_2, \ldots\}$, it holds that

$$\mathbb{P}\left[\bigcup_{i=1}^{\infty} A_i\right] = \sum_{i=1}^{\infty} \mathbb{P}[A_i].$$

Why these three axioms?

- Axiom I (Non-negativity) ensures that probability is never negative.
- Axiom II (Normalization) ensures that probability is never bigger than 1.
- Axiom III (Additivity) allows us to add probabilities when two events do not overlap.

In words, if A and B are disjoint, then the probability of observing either A or B is the sum of the two individual probabilities. Figure 2.22 illustrates the idea.

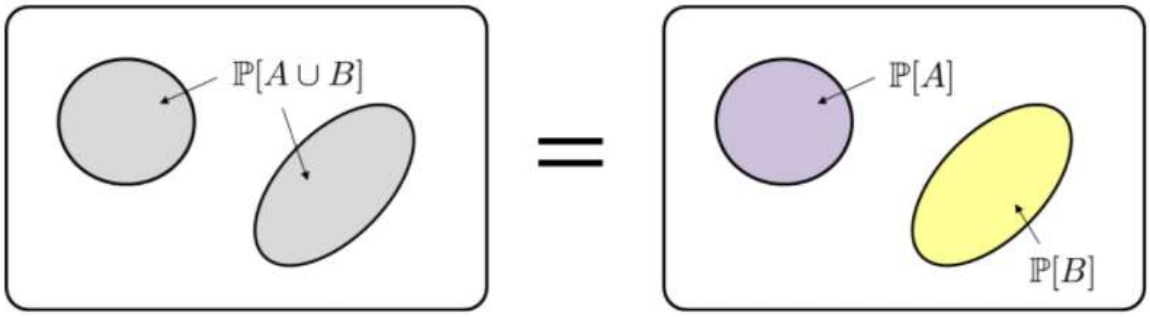


Figure 2.22: Axiom III says $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B]$ if $A \cap B = \emptyset$.

Example 1. Consider a sample space with $\Omega = \{\clubsuit, \heartsuit, \maltese\}$. The probability for each outcome is

$$\mathbb{P}[\{\clubsuit\}] = \frac{2}{6}, \quad \mathbb{P}[\{\heartsuit\}] = \frac{1}{6}, \quad \mathbb{P}[\{\maltese\}] = \frac{3}{6}.$$

Suppose we construct two disjoint events $E_1 = \{\clubsuit, \heartsuit\}$ and $E_2 = \{\maltese\}$. Then, Axiom 3 says

$$\mathbb{P}[E_1 \cup E_2] = \mathbb{P}[E_1] + \mathbb{P}[E_2] = \left(\frac{2}{6} + \frac{1}{6}\right) + \frac{3}{6} = 1.$$

Corollary 2.1. *Let $A \in \mathcal{F}$ be an event. Then,*

(a) $\mathbb{P}[A^c] = 1 - \mathbb{P}[A]$.

(b) $\mathbb{P}[A] \leq 1$.

(c) $\mathbb{P}[\emptyset] = 0$.

Corollary 2.2 (Unions of two Non-Disjoint Sets)**.** *For any A and B in $\mathcal{F}$,*

$$\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B].$$

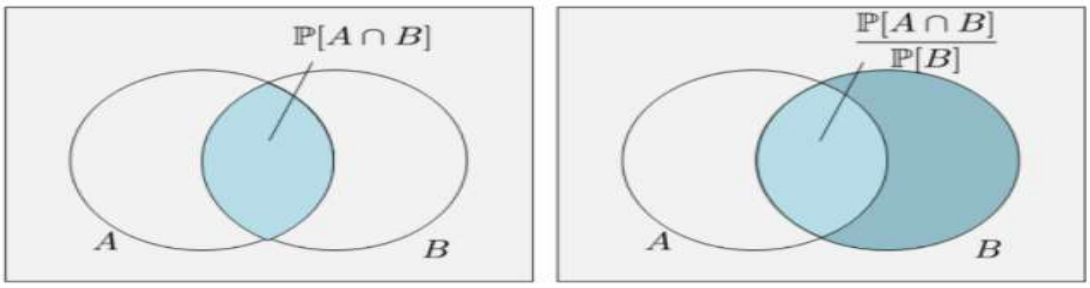
Definition 2.22. *Consider two events A and B. Assume $\mathbb{P}[B] \neq 0$. The* ***conditional probability*** *of A given B is*

$$\mathbb{P}[A \mid B] \stackrel{def}{=} \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}. \tag{2.15}$$

- Section 2.4.1: **Conditional probability**. Conditional probability of A given B is $\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}$.
- Section 2.4.2: **Independence**. Two events are *independent* if the occurrence of one does not influence the occurrence of the other: $\mathbb{P}[A|B] = \mathbb{P}[A]$.
- Section 2.4.3: **Bayes theorem and law of total probability**. Bayes theorem allows us to switch the order of the conditioning: $\mathbb{P}[A|B]$ vs $\mathbb{P}[B|A]$, whereas the law of total probability allows us to decompose an event into smaller events.

Independent and Dependent Events (Important Topic)

Definition . Let E1 and E2 be any two events of a sample space. If the occurrence of E1 does not depend on the occurrence of E2 and the occurrence of E2 does not depend on the occurrence of E1 or in other words the occurrence of any one does not depend on the occurrence of other then E1 and E2 are called independent events otherwise they are called dependent events



e 2.26: Illustration of conditional probability and its comparison with $\mathbb{P}[A \cap B]$.

Definition 2.23. *Two events A and B are statistically* ***independent*** *if*

$$\mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B]$$

Why define independence in this way? Recall that $\mathbb{P}[A \mid B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}$. If A and B are independent, then $\mathbb{P}[A \cap B] = \mathbb{P}[A]\,\mathbb{P}[B]$ and so

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} = \frac{\mathbb{P}[A]\,\mathbb{P}[B]}{\mathbb{P}[B]} = \mathbb{P}[A]. \tag{2.18}$$

Example 2. Consider throwing a dice. Let

$$A = \{\text{Getting a 3}\} \quad \text{and} \quad B = \{\text{getting an odd number}\}.$$

Find $\mathbb{P}[A \mid B]$ and $\mathbb{P}[B \mid A]$.

Solution. The following probabilities are easy to calculate:

$$\mathbb{P}[A] = \mathbb{P}[\{\boxdot\}] = \frac{1}{6}, \qquad \text{and} \qquad \mathbb{P}[B] = \mathbb{P}[\{\boxdot, \boxdot, \boxdot\}] = \frac{3}{6}.$$

Also, the intersection is

$$\mathbb{P}[A \cap B] = \mathbb{P}[\{\boxdot\}] = \frac{1}{6}.$$

Given these values, the conditional probability of A given B can be calculated as

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} = \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{3}.$$

In words, if we know that we have an odd number, then the probability of obtaining a 3 has to be computed over $\{\boxdot, \boxdot, \boxdot\}$, which give us a probability $\frac{1}{3}$. If we do not know that we have an odd number, then the probability of obtaining a 3 has to be computed from the sample space $\{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}$ which will give us $\frac{1}{6}$.

The other conditional probability is

$$\mathbb{P}[B \mid A] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[A]} = 1.$$

Therefore, if we know that the dice is 3, then the probability for this number being an odd number is 1.

Definition 2.24. *Let A and B be two events such that $\mathbb{P}[A] > 0$ and $\mathbb{P}[B] > 0$. Then A and B are independent if*

$$\mathbb{P}[A \mid B] = \mathbb{P}[A] \quad \text{or} \quad \mathbb{P}[B \mid A] = \mathbb{P}[B]. \tag{2.19}$$

Theorem 2.8. ***(Law of Total Probability)*** *Let $\{A_1, \ldots, A_n\}$ be a partition of Ω, i.e., $A_1, \ldots, A_n$ are disjoint and $\Omega = A_1 \cup \ldots \cup A_n$. Then, for any $B \subseteq \Omega$,*

$$\mathbb{P}[B] = \sum_{i=1}^{n} \mathbb{P}[B \mid A_i]\, \mathbb{P}[A_i].$$

Corollary 2.4. *Let $\{A_1, A_2, \ldots, A_n\}$ be a partition of Ω, i.e., $A_1, \ldots, A_n$ are disjoint and $\Omega = A_1 \cup A_2 \cup \ldots \cup A_n$. Then, for any $B \subseteq \Omega$,*

$$\mathbb{P}[A_j \mid B] = \frac{\mathbb{P}[B \mid A_j]\, \mathbb{P}[A_j]}{\sum_{i=1}^{n} \mathbb{P}[B \mid A_i]\, \mathbb{P}[A_i]}. \tag{2.22}$$

Proof. We just need to apply Bayes Theorem and Law of Total Probability:

$$\begin{aligned} \mathbb{P}[A_j \mid B] &= \frac{\mathbb{P}[B \mid A_j]\, \mathbb{P}[A_j]}{\mathbb{P}[B]} \\ &= \frac{\mathbb{P}[B \mid A_j]\, \mathbb{P}[A_j]}{\sum_{i=1}^{n} \mathbb{P}[B \mid A_i]\, \mathbb{P}[A_i]}. \end{aligned}$$

Example 1. Consider a tennis tournament. There are three types of players A, B, and C. The percentage of these players are: A 50%, B 25%, and C 25%. Your chance of winning these players are different.

0.3 against player A.

0.4 against player B.

0.5 against player C.

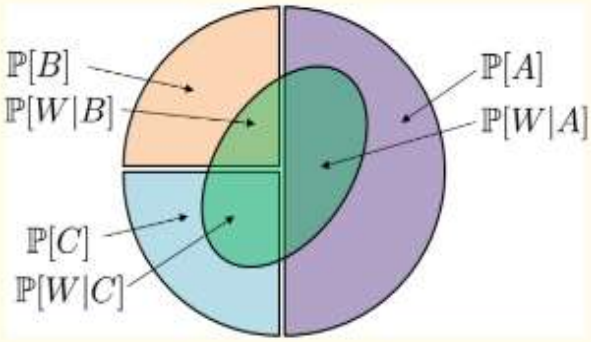
So now, if you enter the game, what is the probability of winning the game? Suppose that you have won the game, what is the probability that you played against player A?

Solution. The first thing to do in this problem is to list out all the available probabilities. We know from the percentage of players that

$$\mathbb{P}[A] = 0.5, \quad \mathbb{P}[B] = 0.25, \quad \mathbb{P}[C] = 0.25.$$

Now, let W be the event that you win the game. Then, the conditional probabilities are defined as follows:

$$\mathbb{P}[W|A] = 0.3, \quad \mathbb{P}[W|B] = 0.4, \quad \mathbb{P}[W|C] = 0.5.$$



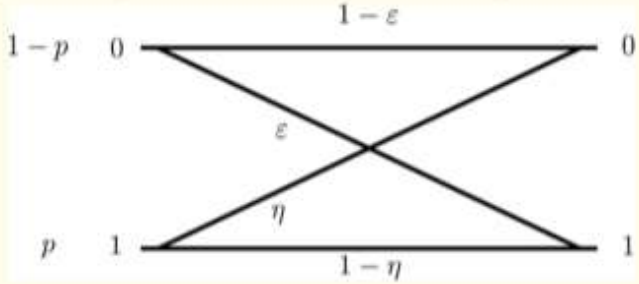
Therefore, by law of total probability, we can show that the probability of winning the game is

$$\begin{aligned}\mathbb{P}[W] &= \mathbb{P}[W \mid A]\,\mathbb{P}[A] + \mathbb{P}[W \mid B]\,\mathbb{P}[B] + \mathbb{P}[W \mid C]\,\mathbb{P}[C] \\ &= (0.3)(0.5) + (0.4)(0.25) + (0.5)(0.25) \\ &= 0.375.\end{aligned}$$

Suppose you have won the game, then the probability of A given W is

$$\begin{aligned}\mathbb{P}[A|W] &= \frac{\mathbb{P}[W|A]\mathbb{P}[A]}{\mathbb{P}[W]} \\ &= \frac{(0.3)(0.5)}{0.375} = 0.4.\end{aligned}$$

Example 2. Consider a communication channel shown below. The probability of sending a 1 is p and the probability of sending a 0 is $1-p$. Given that 1 is sent, the probability of receiving 1 is $1-\eta$. Given that 0 is sent, the probability of receiving 0 is $1-\varepsilon$. Find the probability that a 1 has been correctly received.



Solution. Define the events

$$S_0 = \text{“0 is sent”}, \quad \text{and} \quad R_0 = \text{“0 is received”}.$$
$$S_1 = \text{“1 is sent”}, \quad \text{and} \quad R_1 = \text{“1 is received”}.$$

Then, the probability that 1 is received is $\mathbb{P}[R_1]$. However, $\mathbb{P}[R_1] \neq 1-\eta$ because $1-\eta$ is the conditional probability that 1 is received given 1 is sent. It is possible that we receive 1 as a result of an error when 0 is sent. Therefore, we need to consider the probabilities of having S_0 and S_1. Using Law of total probability we have

$$\begin{aligned}\mathbb{P}[R_1] &= \mathbb{P}[R_1 \mid S_1]\,\mathbb{P}[S_1] + \mathbb{P}[R_1 \mid S_0]\,\mathbb{P}[S_0] \\ &= (1-\eta)p + \varepsilon(1-p).\end{aligned}$$

Now, suppose that we have received 1. What is the probability that 1 was originally sent? This is asking the posterior probability $\mathbb{P}[S_1 \mid R_1]$, which can be found using Bayes Theorem

$$\mathbb{P}[S_1 \mid R_1] = \frac{\mathbb{P}[R_1 \mid S_1]\,\mathbb{P}[S_1]}{\mathbb{P}[R_1]} = \frac{(1-\eta)p}{(1-\eta)p + \varepsilon(1-p)}.$$

Bayes Theorem: (Very Important)

An event B can be explained by a set of exhaustive and mutually exclusive hypothesis A1,A2,….,An. Given ‘a priori‘ probabilities P(A1),P(A2),…,P(An) corresponding to a total absence of knowledge regarding the occurrence of B and conditional probabilities

P(B/A1),P(B/A2),…,P(B/An)

the ‘a posterior‘ probability P(A / B) j of some event Aj is given by

$$\mathrm{P}(\mathrm{A}_j / \mathrm{B}) = \frac{P(\mathrm{A}_j).\mathrm{P}(B / A_j)}{\sum_{i=1}^{n} \mathrm{P}(\mathrm{A}_i) P(B / A_i)}$$

Proof: since the event B can occur when either A1 occurs, or A2 occurs, or,…, An occurs i.e , B can occur in composition with either A1 or A2 consequently

$$B = BA_1 \cup BA_2 \cup BA_3 \cup .....BA_n$$

$$P(B) = P(BA_1 \cup BA_2 \cup BA_3 \cup .....BA_n)$$

Since A_1,A_2…An are mutually exclusive, hence BA_1,BA_2,…BA_n are mutually exclusive forms, therefore by total probability theorem , we have

$$P(B) = P(BA_1) + P(BA_2) + P(BA_3) + ... + P(BA_n)$$

$$= \sum_{i=1}^{n} P(BA_i) = \sum_{i=1}^{n} P(A_i)P(B/A_i)$$

Where $P(B/A_i)$ is the conditional probability of B when A_i has already occurred .

Now from the theorem of compound probability , we have

$$P(A_jB) = P(A_j)P(A_j/B)$$

$$P(A_j/B) = \frac{P(A_jB)}{P(B)} = \frac{P(A_j)P(A_j/B)}{P(B)} \qquad \text{...(2)}$$

From(1)and (2) we get

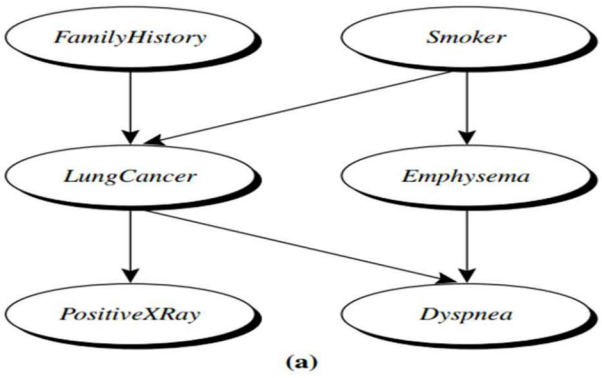
$$P(A_j/B) = \frac{P(A_i)P(B/A_i)}{\sum_{i=1}^{n} P(A_i)P(B/A_i)} \qquad \text{...(3)}$$

Bayesian Belief Networks (very important)

Bayesian belief networks—probabilistic graphical models, which unlike naïve Bayesian classifiers allow the representation of dependencies among subsets of attributes. Bayesian belief networks can be used for classification.

The naıve Bayesian classifier makes the assumption of class conditional independence, that is, given the class label of a tuple, the values of the attributes are assumed to be conditionally independent of one another. This simplifies computation. When the assumption holds true, then the naïve Bayesian classifier is the most accurate in comparison with all other classifiers. In practice, however, dependencies can exist between variables. Bayesian belief networks specify joint conditional probability distributions. They allow class conditional independencies to be defined between subsets of variables. They provide a graphical model of causal relationships, on which learning can be performed. Trained Bayesian belief networks can be used for classification. Bayesian belief networks are also known as belief networks, Bayesian networks, and probabilistic networks.

A belief network is defined by two components—a directed acyclic graph and a set of conditional probability tables. Each node in the directed acyclic graph represents a random variable. The variables may be discrete- or continuous-valued. They may correspond to actual attributes given in the data or to "hidden variables" believed to form a relationship (e.g., in the case of medical data, a hidden variable may indicate a syndrome, representing a number of symptoms that, together, characterize a specific disease). Each arc represents a probabilistic dependence. If an arc is drawn from a node Y to a node Z, then Y is a parent or immediate predecessor of Z, and Z is a descendant Simple Bayesian belief network. (a) A proposed causal model, represented by a directed acyclic graph. (b) The conditional probability table for the values of the variable LungCancer
(LC) showing each possible combination of the values of its parent nodes, FamilyHistory (FH) and Smoker (S).



(a)

	FH, S	FH, ~S	~FH, S	~FH, ~S
LC	0.8	0.5	0.7	0.1
~LC	0.2	0.5	0.3	0.9

(b)

A belief network has one conditional probability table (CPT) for each variable. The CPT for a variable Y specifies the conditional distribution P(Y/Parents(Y)), where Parents(Y) are the parents of Y. The conditional probability for each known value of LungCancer is given for each possible combination of the values of its parents. For instance, from the upper leftmost and bottom rightmost entries, respectively, we see that
P(LungCancer = yes/FamilyHistory = yes, Smoker = yes) = 0.8
P(LungCancer = no/FamilyHistory = no, Smoker = no) = 0.9.
Let X = (x1,……, xn) be a data tuple described by the variables or attributes Y1,……, Yn, respectively. Recall that each variable is conditionally independent of its nondescendants in the network graph, given its parents. This allows the network to provide a complete representation of the existing joint probability distribution with the following equation:

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | Parents(Y_i)),$$