ICT600 Computational Mathematics

Probability distributions and Bayesian Network

Composed by Nirattaya Khamsemanan, Ph.D

1 Probability Distribution

Probability theory is a way to deal with **uncertain information**. Probability represents the **degree of belief** not the degree of truth.

1.1 Random Variables and Domain

In probability theory, variables are called **random variables**. We use an uppercase letter to begin their names, e.g. $Total, Die_1, Cavity$. Each random variable has a **domain** i.e. the set of possible values the variable can take on. For example, the domain of $Cavity$ is $\{true, false\}$; the domain of $Total$ is $\{2, \ldots, 12\}$.

Conventionally, $A = true$ is abbreviated as a, and $A = false$ is abbreviated as $\neg a$.

1.2 Probability Distribution

A **probability distribution** shows the probabilities of the values in the domain.

Example 1. *A random variable Weather has four possible values* $\langle sunny, rain, cloudy, snow\rangle$. *We may have the probabilities as follow:*

$$p(Weather = sunny) = 0.6$$
$$p(Weather = rain) = 0.1$$
$$p(Weather = cloudy) = 0.29$$
$$p(Weather = snow) = 0.01$$

or

$$\mathbf{P}(Weather) = \langle 0.6, 0.1, 0.29, 0.01\rangle$$

$\mathbf{P}(X)$ gives the values of $p(X = x_i) = p(x_i)$ for each possible i, and $\mathbf{P}(X \mid Y)$ gives the values of $p(X = x_i \mid Y = y_j) = p(x_i, y_j)$ for each possible i, j pair.

$\mathbf{P}(Weather, Cavity)$ represents a **joint probability distribution**. It is a 4×2 tables of probabilities.

$$\mathbf{P}(Weather, Cavity) = \mathbf{P}(Weather \mid Cavity)\mathbf{P}(Cavity)$$

represents 8 equations ($W = \mathit{Weather}$ and $C = \mathit{Cavity}$):

$$\begin{aligned}
p(W = \mathit{sunny} \wedge C = \mathit{true}) &= p(W = \mathit{sunny} \mid C = \mathit{true})p(C = \mathit{true}) \\
p(W = \mathit{rain} \wedge C = \mathit{true}) &= p(W = \mathit{rain} \mid C = \mathit{true})p(C = \mathit{true}) \\
p(W = \mathit{cloudy} \wedge C = \mathit{true}) &= p(W = \mathit{cloudy} \mid C = \mathit{true})p(C = \mathit{true}) \\
p(W = \mathit{snow} \wedge C = \mathit{true}) &= p(W = \mathit{snow} \mid C = \mathit{true})p(C = \mathit{true}) \\
p(W = \mathit{sunny} \wedge C = \mathit{false}) &= p(W = \mathit{sunny} \mid C = \mathit{false})p(C = \mathit{false}) \\
p(W = \mathit{rain} \wedge C = \mathit{false}) &= p(W = \mathit{rain} \mid C = \mathit{false})p(C = \mathit{false}) \\
p(W = \mathit{cloudy} \wedge C = \mathit{false}) &= p(W = \mathit{cloudy} \mid C = \mathit{false})p(C = \mathit{false}) \\
p(W = \mathit{snow} \wedge C = \mathit{false}) &= p(W = \mathit{snow} \mid C = \mathit{false})p(C = \mathit{false})
\end{aligned}$$

1.3 Independence

Recall that independence between a and b can be written as

$$p(a|b) = p(a) \text{ or } p(b|a) = p(b) \text{ or } p(a, b) = p(a)p(b)$$

Similarly, the independence between variable X and Y can be written as

$$\mathbf{P}(X|Y) = \mathbf{P}(X) \text{ or } \mathbf{P}(Y|X) = \mathbf{P}(Y) \text{ or } \mathbf{P}(X, Y) = \mathbf{P}(X)\mathbf{P}(Y)$$

1.4 Bayes' Rule

The more general case of Bayes' rule for multivalued variables can be written in the probability distribution notation as

$$\mathbf{P}(Y|X) = \frac{\mathbf{P}(X|Y)\mathbf{P}(Y)}{\mathbf{P}(X)}$$

1.5 Inference using Full Joint Distribution

Here, we explain a technique called **probabilistic inference**.

Marginalization or **summing out** is a process that we sum up the probabilities for each possible value of the other variables.

$$\mathbf{P}(Y) = \sum_{z \in Z} \mathbf{P}(Y, z) \qquad \text{(marginalization rule)}$$

$$\mathbf{P}(Y) = \sum_{z \in Z} \mathbf{P}(Y \mid z)\mathbf{P}(z) \qquad \text{(conditioning rule)}$$

Example 2. *Given a full joint distribution for the* $Toothache, Cavity, Catch$*, find the following probabilities values:*

	$toothache$		$\neg toothache$	
	$catch$	$\neg catch$	$catch$	$\neg catch$
$cavity$	0.108	0.012	0.072	0.008
$\neg cavity$	0.016	0.064	0.144	0.576

1. $p(cavity \vee toothache) =$

2. $p(cavity) =$

3. $p(catch \wedge \neg cavity) =$

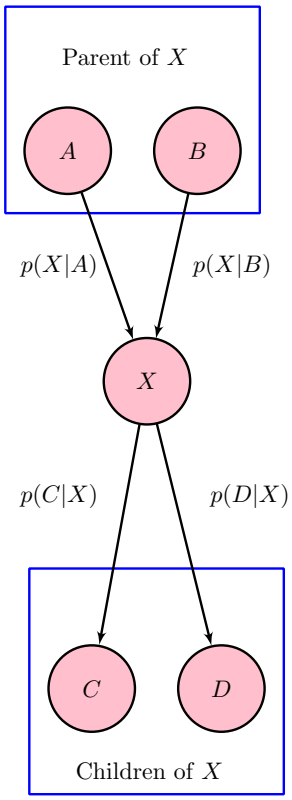
4. $p(cavity \mid toothache) =$

5. $p(catch \mid toothache \wedge cavity) =$

2 Bayesian Network

Definition 1. A Bayesian network is a directed acyclic graph that represents a probabilistic graphical models. In particular a Bayesian network is defined as a pair $B = (G, \Theta)$ where G is a directed graph with no cycles whose vertices $X_1, ..., X_n$ represents random variables, whose edges represent the direct dependencies between these variables. The symbol Θ represents the set of parameters specifies the probability distributions associated with each variable.

An edge from vertex X_i to X_j represents a statistical dependence between the corresponding variables. Vertex X_i is referred to as a parent of X_j and X_j is called the child of X_i. An extension of these genealogical terms is often used to define the sets of descendants which is the set of al vertices that can be reached on a direct path from the vertex. Ancestor vertices is vertices from which the vertex can be reached on a direct path.



Definition 2. The two variables X and Y are conditionally independent given Z if and only if

$$\mathbf{P}(X, Y|Z) = \mathbf{P}(X|Z)\mathbf{P}(Y|Z)$$

which is equivalent to

$$\mathbf{P}(X|Y, Z) = \mathbf{P}(X|Z) \text{ or } \mathbf{P}(Y|X, Z) = \mathbf{P}(Y|Z)$$

Recall the joint probability of $X_1, .., X_n$:

$$p(X_1, ..., X_n) = \prod_{i=1}^{n} p(X_i | X_1, ..., X_{i-1})$$

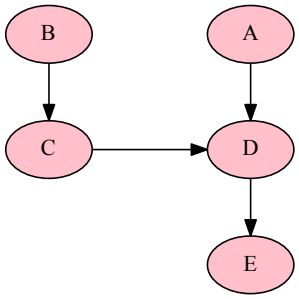
using the probability chain rule.

In Baysian network, the state X_i is conditionally independent of its ancestors given it parents. Therefore,

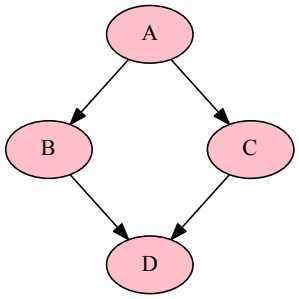
$$\mathbf{P}(X_1, ..., X_n) = \prod_{i=1}^{n} \mathbf{P}(X_i | \text{parents of } X_i)$$

Example 3. *Write the joint probability distribution of the following Bayesian networks.*

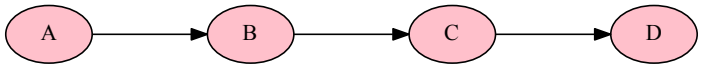
1. .



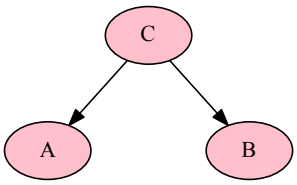
2. .



3. .



2.1 Common causes (tail to tail)

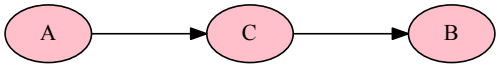


The vertex C is said to be tail-to-tail w.r.t. the path from vertex A to vertex B.

$$\begin{aligned} p(A,B,C) &= p(A|C)p(B|C)p(C) \\ p(A,B) &= \sum_{c\in C} p(A|c)p(B|c)p(c) \neq p(A)p(B) \\ p(A,B|C) &= \frac{p(A|C)p(B|C)p(C)}{p(C)} = p(A|C)p(B|C) \end{aligned}$$

A and B are not independent but they are conditionally independent, given C.

2.2 Causal chain (head to tail)

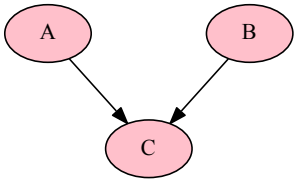


The vertex C is said to be head-to-tail w.r.t. the path from vertex A to vertex B.

$$\begin{aligned} p(A,B,C) &= p(A)p(C|A)p(B|C) \\ p(A,B) &= \sum_{c\in C} p(A)p(c|A)p(B|c) \neq p(A)p(B) \\ p(A,B|C) &= \frac{p(A)p(C|A)p(B|C)}{p(C)} = p(A|C)p(B|C) \end{aligned}$$

A and B are not independent but they are conditionally independent, given C.

2.3 Common effects (head to head)



The vertex C is said to be head-to-head w.r.t. the path from vertex A to vertex B.

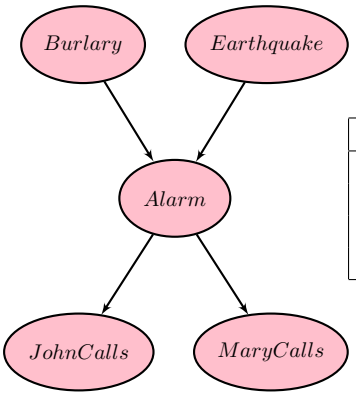
$$\begin{aligned} p(A,B,C) &= p(A)p(B)p(C|A,B) \\ p(A,B) &= \sum_{c\in C} p(A)p(B)p(c|A,B) = p(A)p(B)\sum_{c\in C} p(c|A,B) = p(A)p(B) \\ p(A,B|C) &= \frac{p(A)p(B)p(C|A,B)}{p(C)} \neq p(A|C)p(B|C) \end{aligned}$$

A, B are independent, but conditionally independent given C.

Example 4. *You have a new burglar alarm installed. It reliably detects burglary, but also responds to minor earthquakes. Two neighbors, John and Mary, promise to call the police when they hear the alarm. John always calls when he hears the alarm, but sometimes confuses the alarm with the phone ringing and calls then also. On the other hand, Mary likes loud music and sometimes doesn't hear the alarm. Given evidence about who has and hasn't called, you'd like to estimate the probability of a burglary. A BN representation is shown below (Pearl 1988)*

$p(b)$	$p(\neg b)$
0.001	0.999

$p(e)$	$p(\neg e)$
0.002	0.998



B	E	$p(a \mid B, E)$	$p(\neg a \mid B, E)$
b	e	0.95	0.05
b	$\neg e$	0.94	0.06
$\neg b$	e	0.29	0.71
$\neg b$	$\neg e$	0.001	0.999

A	$p(j \mid A)$	$p(\neg j \mid A)$
a	0.90	0.10
$\neg a$	0.05	0.95

A	$p(m \mid A)$	$p(\neg m \mid A)$
a	0.70	0.30
$\neg a$	0.01	0.99

1. What is the probability that the alarm has sounded but neither a burglary nor an earthquake has occurred, and both Mary and John call?

2. What is the probability that the alarm would sound, given that there was an earthquake?

3. What is the probability that the alarm will sound?

4. What is the probability that there was a burglary, given that John called?

3 Constructing Bayesian Network Structures

Given an application domain with a set of variables, we can construct a Bayesian network structure by the following algorithm:

1. Determine an order of the variables in the domain.
2. Start with an empty network. Select a variable from the order to be added to the network.
3. To add a variable X_i,
   (a) Determine a subset $Parents(X_i)$ of variables in the network such that
   $$\mathbf{P}(X_i|X_1, \ldots, X_{i-1}) = \mathbf{P}(X_i|Parents(X_i))$$
   Here, we need a domain knowledge to make the determination.
   (b) Add a directed edge from each of $Parents(X_i)$ to X_i.

The variable order that follows the *causal model* will result in a network with fewer edges.

Example 5. *A Lecturers Life. Dr. Ann Nicholson spends 60% of her work time in her office. The rest of her work time is spent elsewhere. When Ann is in her office, half the time her light is off (when she is trying to hide from students and get research done). When she is not in her office, she leaves her light on only 5% of the time. 80% of the time she is in her office, Ann is logged onto the computer. Because she sometimes logs onto the computer from home, 10% of the time she is not in her office, she is still logged onto the computer.*

1. *Construct a Bayesian network to represent the Lecturers Life scenario just described.*

2. *Suppose a student checks Dr. Nicholsons login status and sees that she is logged on. What effect does this have on the students belief that Dr. Nicholsons light is on?*