Artificial Intelligence 2: Bayesian Networks I – Introduction and Representation

Shan He

School of Computer Science
University of Birmingham

Outline of Topics

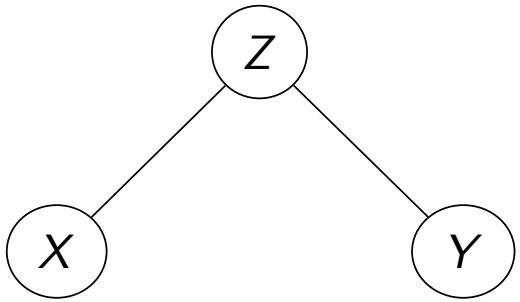
1. Introduction to Bayesian Networks
2. Bayesian Networks: Representation
3. BN structures as probabilistic relationships
   - Conditional Independence

- In today's lecture, we will learn Bayesian networks.
- Bayesian networks is a kind of directed graphical model. They are so called Bayesian networks because they use Bayes' theorem we learned for probabilistic inference, as we explain below.
- Bayesian networks have found many applications in biomedicine, cyber security, robotics, etc.
- In this lecture, we will learn what are Bayesian networks, why they are useful, what they look like, and how to use them to represent probabilistic relationships.

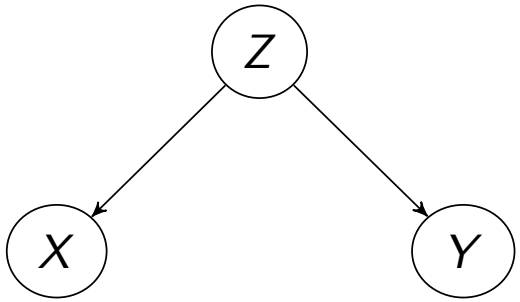
Probabilistic Graphical Models

Probabilistic graphical model: graphs which nodes represent random variables, and the edges (aka. links or arcs) represent **conditional independence** assumptions (explained later). According to the edges, we have:

- **Undirected graphical models**: Markov Random Fields (MRFs, or aka Markov networks)



- **Directed graphical models**: Bayesian Networks ✓



- So, what are Bayesian networks? Bayesian networks are also known as Bayesian Belief Networks, or Belief networks. They are a kind of probabilistic graphical model
- So, what is a probabilistic graphical model?
- Probabilistic Graphical models are a marriage between probability theory and graph theory. They provide a natural tool to deal with two problems that occur throughout AI and machine learning – uncertainty and complexity.
- Formally, a probabilistic graphical model is a graphs which nodes represent random variables, and the edges (aka. links or arcs) represent **conditional independence** assumptions. We learned this **conditional independence** in AI1 in your first year, but we will revisit and go deeper into this concept later.
- According to the edge, that is whether the edges are directional or non-directional, we can classify probabilistic graphical models into two types, Undirected graphical models, which is mainly the Markov Random Fields, also known as Markov networks.
- The directed graphical models are Bayesian networks, which are the main topic of this lecture.

Bayesian Networks

Bayesian Networks: a kind of probabilistic graphic models that uses the direction of edges to represent the cause-effect relationship and Bayes' theorem for probabilistic inference.

Real-world applications of Bayesian Networks:

- Legal Tech: Bayesian networks for settling legal disputes, see ArbiLex, A Harvard Law School Legal Tech Startup, Uses AI To Settle Arbitrations
- Chemistry: prediction of optimal reaction rate and energy, see Explainable and trustworthy artificial intelligence for correctable modeling in chemical sciences
- Cyber security – Darktrace, a British cyber security AI company that reach a valuation of $1.65 billion. Everything you need to know about Darktrace and their patent

- Bayesian networks are a kind of probabilistic graphic models that uses the direction of edges to represent the cause-effect relationship and Bayes' theorem for probabilistic inference.
- So, why should we learn Bayesian networks?
- The reason is that it is one of the widely used probabilistic graphical models. There are many real-world applications.
- You can follow the links to read more about the applications of Bayesian networks, some of them have generate billions of dollars value.
- For example, this cyber security AI company Darktrace, which was born at Cambridge University in 2013, uses Bayesian networks to cyber-threats.
- Their product is called Darktrace Enterprise Immune System, which can "autonomously detect and take action against cyber-threats across all diverse digital environments, including cloud and virtual environments, Internet of Things, and industrial control systems"
- The algorithms behind this product are not detailed, probably were kept as a business secret, but from some articles and this google patent record, we learned that they are probably Bayesian networks.

Bayesian Networks: pros and cons

Advantages:

- Graphical representation: Provide a visual representation of joint probability distributions of different random variables – **interpretability**
- Powerful: can capture complex relationships between random variables
- Combine data and prior knowledge: Prior knowledge can be incorporated and updated with statistically significant information from data – better approximation of true knowledge.
- **Generative approach**: generate new data similar to existing data

Disadvantages:

- Require prior knowledge of many probabilities.
- Sometimes computationally intractable.

- Now, let's wear our critical thinking hat to discuss the pros and cons of Bayesian networks.
- The first and most obvious advantage is that they are a visual representation of joint probability distributions of different random variables. This characteristic gives Bayesian networks good interpretability.
- The second advantage is that they are powerful. You can use a Bayesian network to capture complex relationships between random variables, which is not possible for many machine learning algorithms.
- The third advantage is that you can use Bayesian networks to combine data and prior knowledge. You first use prior knowledge to construct a Bayesian network, then you updated it with statistically significant information from data, by doing so, you can get a better approximation of true knowledge.
- Finally, Bayesian networks is a kind of generative models, which model the the joint probability distribution of random variables. Therefore, once the Bayesian network is obtained, you can generate new data similar to existing data. GAN is just one type of generative models.
- There are a few disadvantages. The first one is it is very difficult to construct a Bayesian model, which requires you have a good prior knowledge about the problem at hand such that you know what to model and how to calculate the probabilities. They are also very computationally demanding, and sometimes not tractable.

Example 1: Naïve Bayes as Bayesian Networks

The simplest Bayesian network – Naïve Bayes Classifier

- Dependent variable Y (Class label)
- Independent variables $\boldsymbol{X} = \{X_1, X_2, \ldots, X_n\}$ (Evidence)
- Main assumption: independent variables are conditional independent given $Y = y$

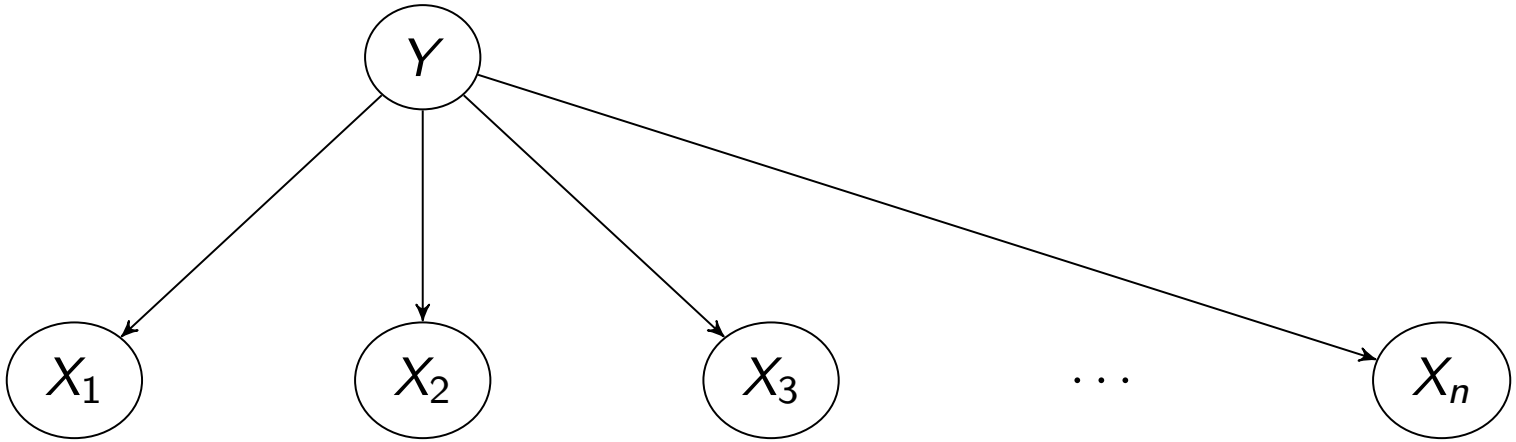


Figure 1: Graphical representation of a Naive Bayes Classifier

- Let's take a few examples of Bayesian networks.
- The first one is actually what you learned in your year 1's AI 1 module, Naive Bayes classifier, which is the simplest Bayesian network models.
- I will not go into the details of this Naive Bayes, but I just listed a few important information especially the main assumption behind this Naive Bayes classifier, that is, we assume the independent variables are independent conditional on the value of the dependent variable Y is known.
- This is also the meaning of a simple generative model, that is, we can interpret the directed graph a recipe of how the data was generated: first, a label was chosen at random; then a subset of n possible random variables were sampled independently and at random.

Main problems in Bayesian Networks Wet grass

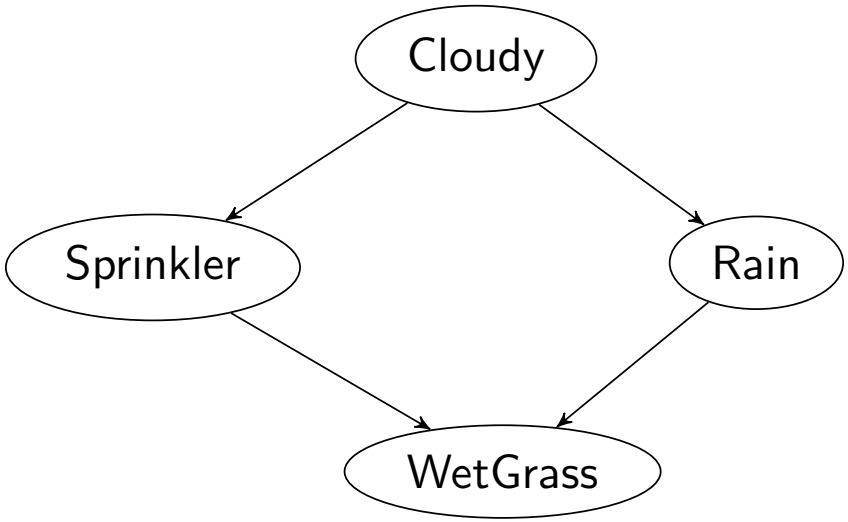


Figure 2: A famous example of Bayesian Networks

Three main problems with Bayesian Networks:

- Inference (from observations "it's cloudy" infer the probability of the wet grass) ✓
- Training the models ✓
- Determining the structure of the network (i.e. what is connected to what)

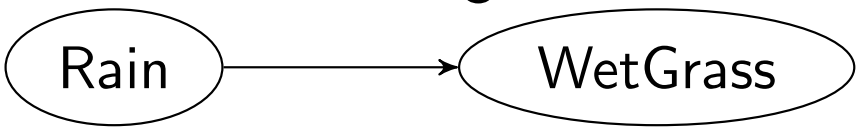
- Before discuss the main problems in Bayesian networks, let's take a look at a 'proper' Bayesian network. It is 'proper' because, rather than a tree structure used in a Naive Bayes classifier, here for this Bayesian network, we have an acyclic graph, that's why we call it network. However, it is **acyclic** because there are no loops, or more accurately, cycles, meaning starting from any one node, you can not go back to this node by the direction of the edges.
- In this particular example, all nodes are binary, i.e., have two possible values, which we will denote by 1 (true) and 0 (false).
- We see that the event "grass is wet" (W=1) has two possible causes: either the water sprinkler is on or it is raining. We will dissect this example as small parts to learn how to infer a Bayesian network in this lecture.
- So, given such a network, there are two main problems, inference and training. Inference means we infer the value or probability of some node. For example, from the observation, it is cloudy, we infer the probability of the grass is wet.
- We will also discuss how to train a Bayesian network with known structure and full observability, which is based on maximum likelihood estimate we learned.
- But if we do not know the network structure, then we need to determining the appropriate structure, which is an area of on-going research. We will not discuss this problem in our module.

Bayesian Networks: Representation

Problem: How to represent the joint probability distributions of random variables.

Solution: A Bayesian network – a directed, acyclic graph, which consists of

- a set of nodes: each represent a random variable
- a set of dircted edges that connect those nodes, for example:



, the directed edge represents "directed dependency" or "directed influences", also called "direct cause"

- A conditional distribution for each node given its parents:

$$P(X_i|\text{Parents}(X_i))$$

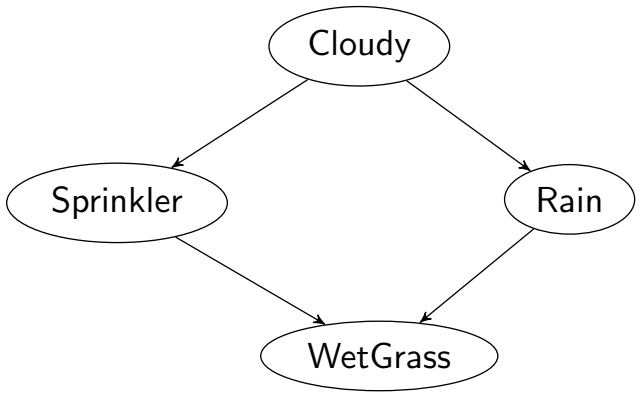
  - Discrete Random Variables: conditional distribution can be represented as a conditional probability table (CPT) – the distribution over X_i for each combination of parent values

- As discussed in our previous lecture, the aim of generative machine learning approach is to learn or represent the joint probability distributions of a set of random variables.
- Probabilistic Graphical Models use graphs to represent the joint distributions. Specifically, Bayesian networks use directed, acyclic graphs as the graph representation.
- In a Bayesian network, the directed, acyclic graph consists of a set of nodes, of which each represent a random variable.
- There are also a set of edges that connect those nodes. So what those edges represent? Basically a directed edge represents "directed dependency" or "directed influences", which means, the parent node will have direct influence on the child node, or the value of the child node directly dependents on the value of the parent node.
- Sometimes in the literature, you will find the edges are interpreted as the"cause - effect" relationship. In this lecture, we will use 'causes' and 'influence' interchangeably.
- To quantify the relationship, we use conditional probability distributions, or conditional dependence. For each node given its parents, the conditional distribution is defined as the distribution of $P(X_i)$ given all its parents.
- For Discrete Random Variables, we use a conditional probability table (CPT) to represent the conditional distributions, which lists the distribution over X_i for each combination of parent values. Let's take a

Bayesian Networks: Wet grass example

$P(C=0)$	$P(C=1)$
0.4	0.6

C	$P(S=0)$	$P(S=1)$
0	0.5	0.5
1	0.9	0.1



C	$P(R=0)$	$P(R=1)$
0	0.8	0.2
1	0.2	0.8

S	R	$P(W=0)$	$P(W=1)$
0	0	1.0	0.0
1	0	0.1	0.9
0	1	0.1	0.9
1	1	0.01	0.99

- Here is an example of the wet grass Bayesian networks with all the Conditional Probability Tables for all the random variables.
- Here the first node is the root node, which does not have any parents, so it is the prior. We know in the UK, the probability of cloudy is half, which is represented in this CPT.
- Then given its cloudy or not, we can specify the conditional probability of sprinkler is on and whether it rains, which are in the these two tables. Notice that, each row sum up to 1, because of the Kolmogorov unit measure axiom, that is the probabilities of all possible outcome in the same sample space sum up to 1.
- However, unlike the joint PMF we learned in Lecture 5, the column values do not sum up to 1, that's because in each column, the values of different rows are in different sample space, e.g., one is cloudy and the other is not cloudy. Therefore, summing up the probabilities of a column does not make any sense.

Bayesian networks: the equation

Full joint distribution:

$$P(X_1, X_2, \ldots, X_n) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)\ldots P(X_n|X_1, X_2, \ldots, X_{n-1})$$
$$= \prod_{i=1}^{n} P(X_i|X_{i-1}, \cdots, X_1), \quad (1)$$

but for a Bayesian network, based on the edges, we define the full joint distribution as the product of the local conditional distributions:

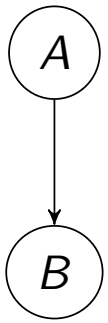
$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|\text{Parents}(X_i)) \quad (2)$$

Essence of a Bayesian network: a compact representation of a joint probability distribution in terms of conditional distribution

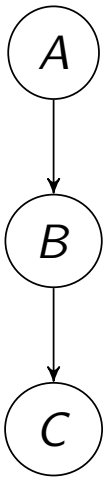
- To define a full joint distribution for n random variables, we need to use the Chain rule to derive the following long equations, which will grow exponentially with then number of n. Therefore, this full joint distribution is not practical for complex real-world problems which usually involve large number of random variables.
- In contrast, if we use a Bayesian network to model full joint distribution, after defining the structure, we use the product of the local conditional distribution as the following equation,
- This is the main equation to define a Bayesian network.
- So, from the this equation, we know that, the essence of a Bayesian network is a compact representation of a joint probability distribution in terms of conditional distribution.

Probabilistic relationship: standard structures

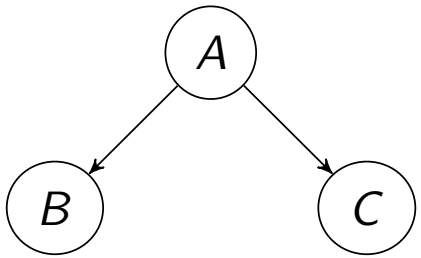
Direct Cause



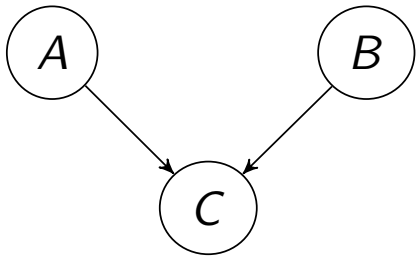
Indirect Cause



Common Cause



Common effect

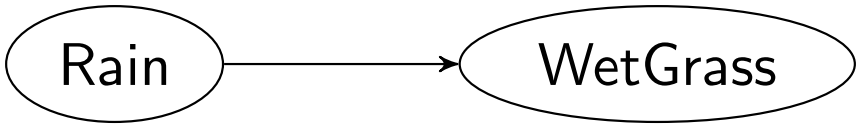


- As we mentioned, the Bayesian network use edges to represent directed influence using conditional probability distribution, or more generally, conditional independence.
- This conditional independence leads to the following four probabilistic relationships. They are direct cause, indirect cause, common cause and common effect, which are represented by the following four standard structures in Bayesian networks.
- We will discuss each of these four standard structures using our web grass example.

Edge: direct cause

Example: We notice the grass in our garden is wet, which might be caused by rain. Construct a Bayesian network to represent their probabilistic relationship.

Solution: The probabilistic graphic model is



Use binary random variables:

- W where $R_W = \{0, 1\}$ to represent whether the grass is wet (1) or not (0)
- R where $R_r = \{0, 1\}$ to represent whether it has been raining (1) or not (0)

Probabilistic relationship: An edge presents a cause-effect relationship, called direct cause, or conditional dependence between the parent node (cause) and the child node (effect):

$$P(W|R)$$

- We can construct the simplest Bayesain network with two nodes and one directe edge as the following, which represent the so-called direct dependency
- we then use random variable W where $R_W = \{0, 1\}$ to represent whether the grass is wet (1) or not (0), and random variable R where $R_r = \{0, 1\}$ to represent whether it has been raining (1) or not (0).
- We also know that the probability of raining through a day is

  $$P(R = 1) = 0.4 \Rightarrow P(R = 0) = 0.6$$

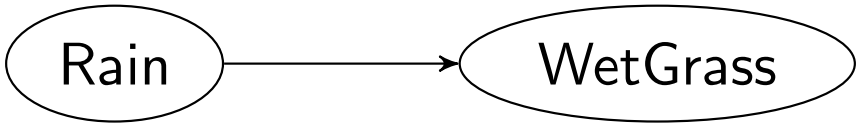
  and the probability that grass gets wet when it rains is 0.9, that is

  $$P(W = 1|R = 1) = 0.9 \Rightarrow P(W = 0|R = 1) = 0.1$$

- We can now formally represent their probabilistic relationship, which as we discussed, is the direct cause-effect relationship, called direct cause relationship, written as a conditional PMF $P(W)$ given R. So basically an edge capture a known conditional dependence between two random variables.

Probabilistic relationship: direct cause

Solution: The probabilistic graphic model is



Prior knowledge:

- The probability of raining through a day is
  $P_R(R = 1) = 0.4 \Rightarrow P_R(R = 0) = 0.6$
- The probability that grass gets wet when it rains is 0.9, that is
  $P(W = 1|R = 1) = 0.9 \Rightarrow P(W = 0|R = 1) = 0.1$
- The probability that grass get wet without raining, (e.g., when someone turns on the sprinkler):
  $P(W = 1|R = 0) = 0.2) \Rightarrow P(W = 0|R = 0) = 1 - 0.2 = 0.8$

- Here we need to specify probability distributions of prior knowledge.
- Let's assume we know that the probability of raining through a day is

$$P_R(R = 1) = 0.4,$$

  it is obvious the probability of not raining is $1 - 0.6 = 0.4$
- Since R, i.e., whether it rains or not, directly influence W, that is whether the grass is wet or not, we need to discuss the joint conditional probability distribution of R and W. The probability that grass gets wet when it rains is 0.9, that is

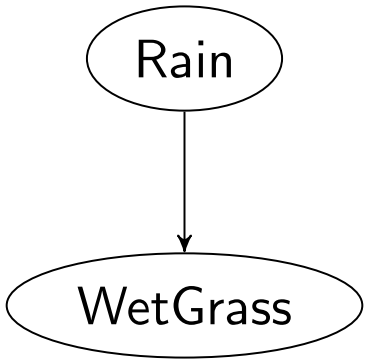
$$P(W = 1|R = 1) = 0.9$$

- We also know that, it is also possible that it rains, but not heavy enough, for example, just a drizzle or a very short rain, the grass is not wet enough to be notice. Let's say this happen with a probability of 0.1
- Don't forget, there is another possibility that the grass gets wet without raining, for example, we use sprinkler to water the grass. Let's assme this conditional probability is 0.2. We shall explicitly model this later. For now, we just acknowledge there are other causes to make the grass wet.

Probabilistic relationship: direct cause

Solution: The Bayesian network with CPTs:

$P(R=0)$	$P(R=1)$
0.6	0.4



R	$P(W=0)$	$P(W=1)$
0	0.8	0.2
1	0.1	0.9

- Now, we can construct a full Bayesian network representation that can be used for inference.
- Please note that a CPT represents conditional probability mass function (PMF) but not joint PMF. Therefore, you cannot derive marginal PMF from this table.

Direct cause: Inference

In general the random variables (nodes) fall into two groups:

- **Observed variables**: the ones we have knowledge about
- **Unobserved variables**: ones we do not know about and therefore have to infer the probability.

Question: you observed the grass is wet, what is the probability it rained?
Solution: Based on Bayes theorem, the probability that it rained given the grass is wet can be calculated as

$$P(R=1|W=1) = \frac{P(W=1|R=1)P(R=1)}{P(W=1)}$$

The marginal probability of wet grass can be computed by summing up the joint PMF over the possible values that its parent node can take:

$$P(W=1) = \sum_{r\in\{0,1\}} P(R=r, W=1) = \sum_{r\in\{0,1\}} P(W=1|R=r)P(R=r)$$
$$= 0.9 \times 0.4 + 0.2 \times 0.6 = 0.48$$

- Now, we can use this very simple Bayesian network to do some inference. But first, we need to now what random variables are observable and what are unobservable.
- Basically, the ones we have knowledge about are observable variables. For those we do not know about are called Unobserved variables. They are the ones we need to infer their probability.
- Let's say, you are working all day inside your house without noticing the weather outside, then in the evening, you observe the grass is wet, now what is the probability it rained? Here, the random variable W is the observed variable, and R is unobserved.
- If we observed that the grass is wet, the probability that it rained can be calculated as using the Bayes theorem to invert the dependencies and have a diagnosis as the following equation.
- The numerator, i.e., the product of the likelihood and the prior is known, however, we need to calculate the marginal likelihood of grass is wet, which is by summing up the joint PMF over the possible values that its parent node can take by the following equation.
- This marginal likelihood means that, not knowing whether it rained or not, what is the probability the grass is wet, that is 0.48. In contrast, if we knew for sure that it did not rain, the probability would be as low as 0.2 (See Page 13, CPT bottom, first row); If we knew that it rained, the probability of wet grass would be 0.9.

Direct cause: Inference

$$P(R=1|W=1) = \frac{P(W=1|R=1)P(R=1)}{P(W=1)}$$
$$= \frac{0.9 \times 0.4}{0.48} = 0.75$$

The above inference is called diagnosis, i.e., to obtain $P(\text{Cause}|\text{Effect})$
Interpretation: Knowing that the grass is wet increased the probability of it rained from 0.4 to 0.75.

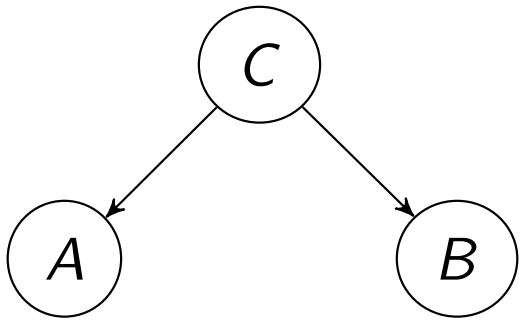
Conditional independence

Independence: Two sets of variables A and B are independent iff $P(A) = P(A|B)$ or equivalently $P(A, B) = P(A)P(B)$

Conditional Independence: Two random variables A and B are conditionally independent if they are independent given a third random variable C, written as:

$$(A \perp\!\!\!\perp B) \mid C \iff P(A, B|C) = P(A|C)P(B|C)$$

Graphically can be represented as

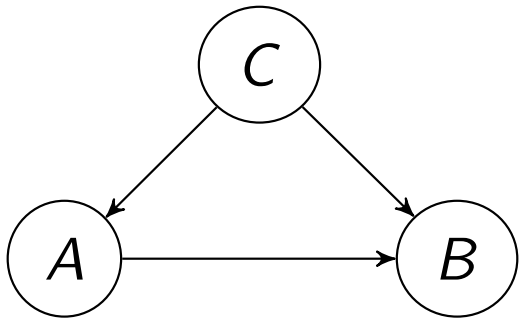


- As discussed, if there an directed edge between two nodes, there is an conditional dependence between the parent and the child.
- For two nodes without any edge between them, they are two independent random variables, formally defined as Two sets of variables A and B are independent iff $P(A) = P(A|B)$ or equivalently $P(A, B) = P(A)P(B)$
- The above dependent and the **direct cause** relationship, or the so-called conditional dependence is very simple, only involves two nodes. However, to model a real situation, the Bayesian network is more complex than a pair of nodes.
- To model a complex joint probability distribution but remain as mathematically tractable as possible, Bayesian network relies on the important concept in probability theory, Conditional Independence.
- Let's consider three random variables or three nodes in a Bayesian network. Informally, given two random variables A and B, they are conditionally independent given a third random variable C if and only if they are independent in their conditional probability distribution given C.
- This is formally written as the following equation. The above conditional independence concept can be visualised by the following Bayesian network with three nodes, A, B and C.
- The essence is that, the value of random variable A does not affect the distribution of B if C is known. Now, looking at this graph, which now turns out to be very simple, you might ask, what is the big deal about this conditional independence assumption in Bayesian networks??

Why conditional independence matters?

The full joint distribution of three random variables is:

$$P(A, B, C) = P(C)P(A|C)P(B|A, C)$$



Suppose A, B and C can take one of 2 values, $\{0, 1\}$. Modelling the complete joint distribution requires $1 + 2 + 4 = 7$ parameters. In contrast, using the conditional independence as specified in page 17, we need $1 + 2 + 2 = 5$ parameters.

- To understand the importance of this conditional independence assumption, let's take a look the situation without this assumption.
- If we abandon this assumption, then the following Bayesian network is possible, that is, C directly cause A, in addition, C and A acting together, directly cause B.
- This is the so-called full joint distribution of three random variables, which is given by this equation.
- If we use this full joint distribution, also suppose A, B and C can take one of 2 values, $\{0, 1\}$. Modelling the complete joint distribution requires 7 parameters. Please work out why yourself. In contrast, using the conditional independence in our previous slides requires 5 parameters. You can use the figure on Page 9 to derive this.

Markov Condition of Bayesian networks

Markov Condition: Each random variable X is conditionally independent of its non-descendants, given its parents.

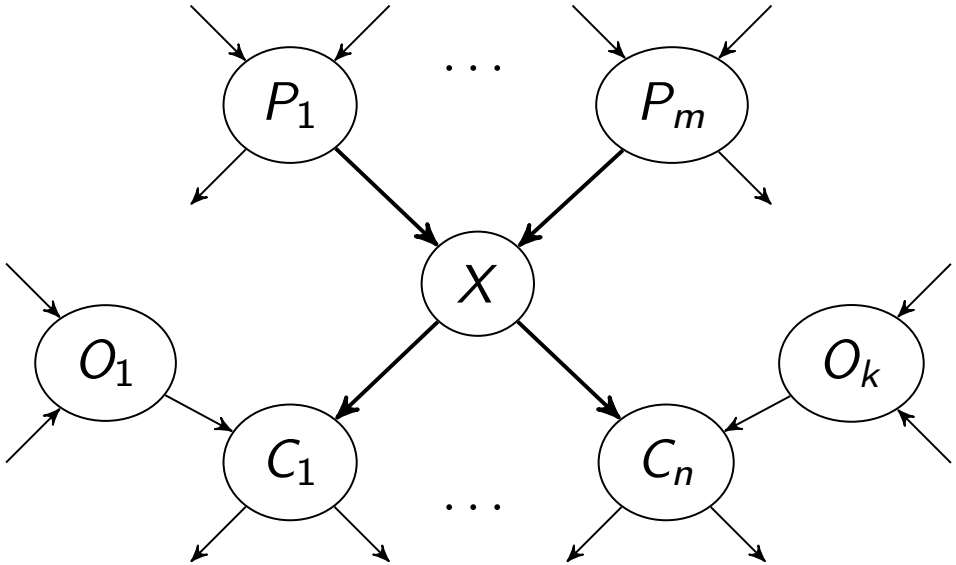
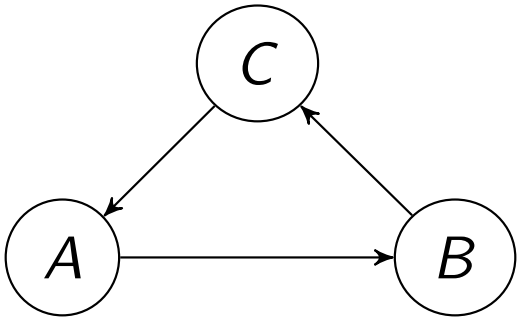


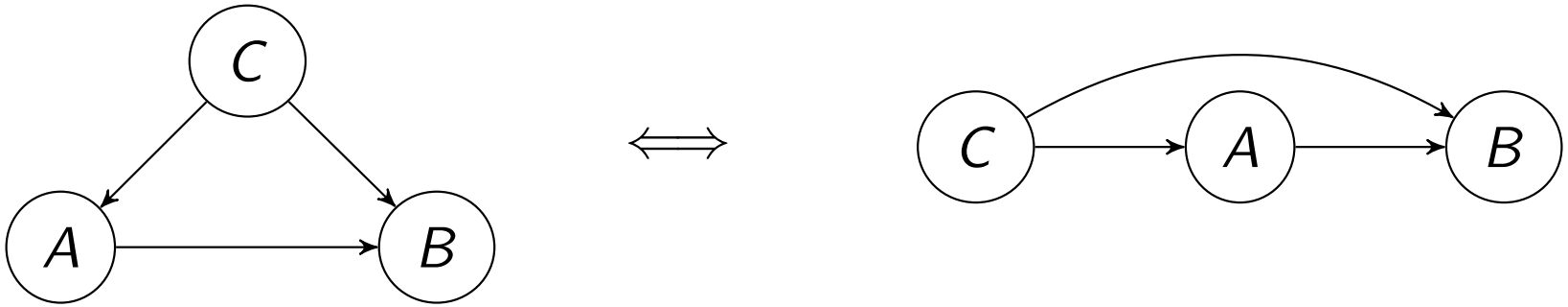
Figure 3: P_1, P_m are parents of node X, C_1 and C_n are the descendants or children of X. All other nodes, here, denoted as O_1 and O_k are neither parents or children. Therefore, given P_1 and P_m, X is conditionally independent of all other nodes O_1 and O_k. This figure also illustrates an Markov Blanket of X, which is defined as X's parents + children + children's parents. The Markov blanket is sufficient enough for inferring X.

- The conditional independence assumption also lead to an important property of Bayesian network, called Markov Condition, that is, every variable X is conditionally independent of all other variables that are neither parents nor children of X, given the parents of X.
- This can be illustrated by this figure. Here, P_1, P_m are parents of node X, C_1 and C_n are the descendants or children of X. All other nodes, here, denoted as O_1 and O_k are neither parents or children, therefore, X is conditionally independent to these other nodes O_1 and O_k

What do you mean?

A standard Bayesian network only allows the four standard structures in Figure 2, and the following structures are **NOT** allowed



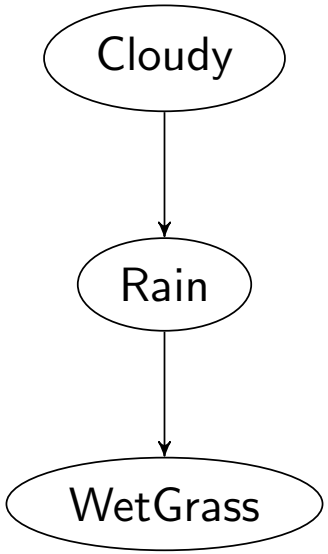


Take home message: a node is independent of its ancestors given its parents, where the ancestor/parent relationship is with respect to some fixed topological ordering of the nodes.

- All the above discussion basically means in a Bayesian network only the standard structures in exist, and the following structures are **NOT** allowed.
- The first structure will never appear in a Bayesian network is this cycle, that's by definition, a Bayesian network is a directed acyclic network.
- The second structure is not allowed in Bayesian networks is this so-called feed-forward loop, which essentially is the conditional independence assumption, which means, give the parent of B, that is A, C should be independent of B.
- So, the take home message is: a node is independent of its ancestors given its parents, where the ancestor/parent relationship is with respect to some fixed topological ordering of the nodes.

Probabilistic relationship: Indirect cause

Consider Cloudy or not will influence rain, we have



Here since WetGrass is independent of Cloudy given Rain,
According to equation 2, we have the joint distribution:

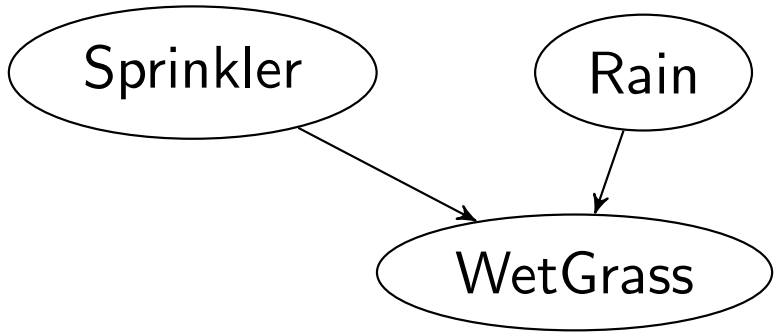
$$P(C, R, W) = P(C)P(R|C)P(W|R)$$

- Let's go back to our Web Grass example to investigate the four standard structures of the probabilistic relationships in Figure **??**. Consider the UK's weather which is always cloudy, it is natural to consider cloudy in the wet grass example.
- We know Cloudy or not will influence rain, and by common sense and also by the conditional independence assumption, we cannot have the direct cause or an edge connecting Cloudy to WetGrass. Therefore, we have this indirect cause relationship, that is, if rain is observed, then Cloudy and Rain are independent.
- Usiing equation 2, we can write the joint PMF as

$$P(C, R, W) = P(C)P(R|C)P(W|R)$$

Probabilistic relationship: Common effect

Suppose we identify another cause of wet grass, the sprinkler, we can then model the relationship as a converging connection (Common effects):



The joint probability distribution:

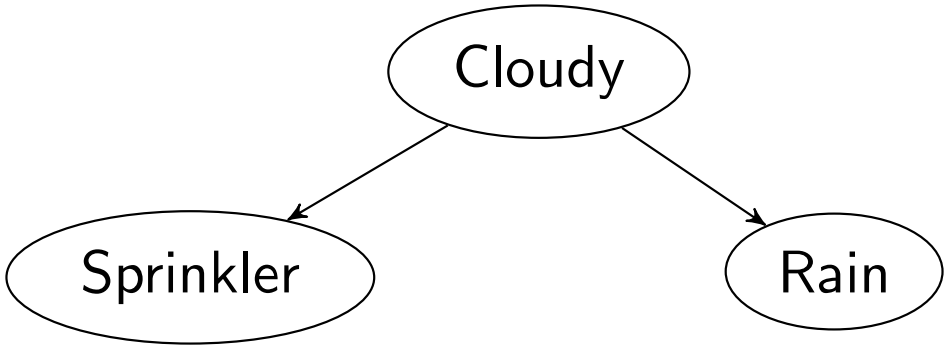
$$P(S, R, W) = P(S)P(R)P(W|S, R)$$

If neither W nor any of its descendants are observed, S and R are independent.

- In the above example, notice that the two causes "compete" to "explain" the observed data. Hence S and R become conditionally dependent given that their common child, W, is observed, even though they are marginally independent. For example, suppose the grass is wet, but that we also know that it is raining. Then the posterior probability that the sprinkler is on goes down:
- It refers to the phenomenon where knowing that one cause has occurred, reduces (but does not eliminate) the probability that the other cause(s) took place. We shall discuss this in detailed in my next lecture.

Probabilistic relationship: Common cause

It is natural to suppose the cloudy weather influence our decision to turn on the sprinkler or not and cloudy weather will also influence the chance of rain, we model this "common cause" probabilistic relationship using:



The joint probability distribution:

$$P(C, S, R) = P(C)P(S|C)P(R|C)$$

If C is observed, then S and R are independent.

- Finally, that's is the structure of common cause, that is exactly we used to illustrate the concept of conditional independence.
- Here in this example, cloudy weather influence our decision to turn on the sprinkler or not and cloudy weather will also influence the chance of rain, hence, Cloudy is the "common cause", we then model the probabilistic relationship using the following Bayesian network, and the joint PMF is defined as

$$P(C, S, R) = P(C)P(S|C)P(R|C)$$

- If C is observed, then S and R are independent.

Further reading

- Probabilistic Graphical Models: Chapters 3.1-3.3 (Please find the online version yourself.)
- Artificial Intelligence: Foundations of Computational Agents, Chapter 8