Machine Learning

Section 4: Bayesian networks

贝叶斯网络

Stefan Harmeling

13. October 2021

Computational difficulties of probability theory

概率论的计算困难

Computational difficulties of probability theory

概率论的计算困难

The problem:

- The joint distribution of propositional variables $A, B, \ldots, Z$ has many free parameters.

命题变量A, B, ...的联合分布。, Z有许多自由参数。

$$
\begin{aligned}
&[1] && p(A, B, \ldots, Z) = \ldots \\
&[2] && p(\neg A, B, \ldots, Z) = \ldots \\
&[3] && p(A, \neg B, \ldots, Z) = \ldots \\
& \vdots && \\
&[67108863] && p(\neg A, \neg B, \ldots, Z) = \ldots \\
&[67108864] && p(\neg A, \neg B, \ldots, \neg Z) = 1 - \sum p(\ldots)
\end{aligned}
$$

- Requires a large memory and calculating $p(A)$ requires a lot of time. 需要很大的内存，计算p(A)需要很多时间。
- How can we specify the joint distribution with fewer numbers?
- Can we restrict how variables are relevant to each other.

我们如何用更少的数字来指定联合分布？
我们能否限制变量之间的相关方式。

An important note about notation

So far:

A represents a formula (or event):

到目前为止。

A代表一个公式（或事件）。

$p(A)$ = probability that formula A is true

$p(\neg A)$ = probability that formula $\neg A$ is true

A是一个（命题）变量，其值在{0，1}，即p(A)是两个可能的输入值A=1和A=0的函数，也就是说，用的是稍微不寻常的符号。

From now on:

A is a (propositional) variable with values in $\{0, 1\}$, i.e. $p(A)$ is a function of two possible input values $A{=}1$ and $A{=}0$, i.e. with slightly unusual notation:

$p(A{=}1)$ = probability that proposition A is true

$p(A{=}0)$ = probability that proposition A is false

Stating that $p(A, B) = p(A)\, p(B)$ means:

$$p(A{=}1, B{=}1) = p(A{=}1)\, p(B{=}1)$$
$$p(A{=}1, B{=}0) = p(A{=}1)\, p(B{=}0)$$
$$p(A{=}0, B{=}1) = p(A{=}0)\, p(B{=}1)$$
$$p(A{=}0, B{=}0) = p(A{=}0)\, p(B{=}0)$$

Tracy, Jack and the wet grass (1) — joint prob.

from Barber 2012, 3.1.1

联合概率。

T = Tracey's grass is wet

R = it rained last night

S = Tracey's sprinkler was on last night

J = grass of Tracey's neighbor Jack is wet

T = 特蕾西的草是湿的
R = 昨晚下雨了
S = 特蕾西的洒水车昨晚开了
J = 特蕾西的邻居杰克的草是湿的

Joint probability

$$\begin{aligned} p(T, J, R, S) &= p(T, J, R|S)\, p(S) \\ &= p(T, J|R, S)\, p(R|S)\, p(S) \\ &= p(T|J, R, S)\, p(J|R, S)\, p(R|S)\, p(S) \end{aligned}$$

- apply three times product rule $p(A, B) = p(A|B)\, p(B)$

应用三次乘法则

Tracy, Jack and the wet grass (2) — parameter counting

from Barber 2012, 3.1.1

$$
\begin{aligned}
T &= \text{Tracey's grass is wet} \\
R &= \text{it rained last night} \\
S &= \text{Tracey's sprinkler was on last night} \\
J &= \text{grass of Tracey's neighbor Jack is wet}
\end{aligned}
$$

Number of parameters of joint probability 联合概率的参数数量

$$p(T, R, S, J) = p(T|J, R, S)\ p(J|R, S)\ p(R|S)\ p(S)$$

- $p(T, R, S, J)$ requires 15 parameters.
- rewritten with product rule requires $8 + 4 + 2 + 1$ parameters.

Leave out irrelevant conditions (use domain knowledge)

撇开不相关的条件（使用领域知识）。

$$p(T, J, R, S) = p(T|R, S)\ p(J|R)\ p(R)\ p(S)$$

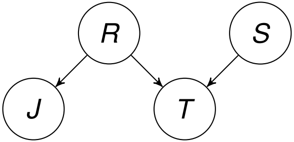
- only $4 + 2 + 1 + 1 = 8$ parameters!

Tracy, Jack and the wet grass (2) — representation

from Barber 2012, 3.1.1

$$p(T, J, R, S) = p(T|R, S)\; p(J|R)\; p(R)\; p(S)$$

Graphical representation



Conditional probability tables (CPTs)

$$p(R{=}1) = 0.2 \qquad p(S{=}1) = 0.1$$

$$p(J{=}1|R{=}1) = 1 \qquad p(J{=}1|R{=}0) = 0.2$$

$$p(T{=}1|R{=}1, S{=}0) = 1 \qquad p(T{=}1|R{=}1, S{=}1) = 1$$

$$p(T{=}1|R{=}0, S{=}1) = 0.9 \qquad p(T{=}1|R{=}0, S{=}0) = 0$$

Tracy, Jack and the wet grass (3) — inference

from Barber 2012, 3.1.1

Inference 鉴于我们观察到特蕾西的草地是湿的，那么洒水车开的概率是多少？

- What is the probability that the sprinkler was on given that we observe that Tracey's grass is wet?

$$p(S{=}1|T{=}1) = \frac{p(S{=}1, T{=}1)}{p(T{=}1)} = \frac{\sum_{J,R} p(T{=}1, J, R, S{=}1)}{\sum_{J,R,S} p(T{=}1, J, R, S)}$$

$$= \ldots = 0.3382$$

考虑到我们的情况，洒水车开启的概率是多少？

- What is the probability that the sprinkler was on given that we observe that Tracey's and Jack's grass is wet?

$$p(S{=}1|T{=}1, J{=}1) = \frac{p(S{=}1, T{=}1, J{=}1)}{p(T{=}1, J{=}1)} = \frac{\sum_{R} p(T{=}1, J{=}1, R, S{=}1)}{\sum_{R,S} p(T{=}1, J{=}1, R, S)}$$

$$= \ldots = 0.1604$$

杰克的湿草是在解释洒水车是特蕾西的湿草的原因。

Jack's wet grass is *explaining away* the sprinkler as a reason for the wet grass of Tracey. Note: $S \perp\!\!\!\perp J$ but $S \not\!\perp\!\!\!\perp J \mid T$.

What is probabilistic reasoning?

Barber 2012, 1.2

什么是概率推理?

1. identify all relevant variables, e.g. T, J, R, S
2. define joint probability $p(T, J, R, S)$
3. *evidence* fixes the values of certain variables, e.g. $T{=}1$
4. *inference* of the distribution of certain variables requires integrating out the rest, e.g. to calculate $p(S{=}1 | T{=}1)$

1.确定所有相关变量，如T , J, R, S
2. 定义联合概率p(T , J, R, S)
3.证据固定了某些变量的值，如T=1
4. 推断某些变量的分布需要整合其他变量,
例如计算p(S=1ST=1)

Bayesian networks aka Bayes nets, belief networks (1)

Typical definition from Barber 2012, 3.3 Belief networks; see also Pearl, 1988

Definition 4.1 (Bayesian network (version w/o explicit graph))

A Bayesian network is a distribution that can be written as

$$p(X_1, X_2, \ldots, X_D) = \prod_{i=1}^{D} p(X_i|\mathrm{pa}(X_i))$$

Don't use this definition!

where $\mathrm{pa}(X)$ *are the parental variables of variable* X*. A Bayesian network can be represented as a Directed Acyclic Graph (DAG) with the propositional variables as nodes and arrows from parents to children.*

Problems of this definition:

- The graph is not unique! E.g.

$$p(X_1, X_2) = p(X_1)p(X_2|X_1) = p(X_2)p(X_1|X_2)$$

In both case p is a Bayesian network.

图形不是唯一的! 例如:
在这两种情况下，p是一个贝叶斯网络。

Bayesian networks aka Bayes nets, belief networks (2)

Compare Peters, Def 6.32 of causal graphical model

Better definition:

是Xj的父母在G中的索引集。

Definition 4.2 (Bayesian network)

点

条件概率

A Bayesian network is a DAG $\mathcal{G}$ with vertices $X_1, \ldots, X_n$ and conditional probabilities $p(X_j | X_{\mathrm{pa}_j^{\mathcal{G}}})$ where $\mathrm{pa}_j^{\mathcal{G}}$ is the set of indices of the parents of X_j in $\mathcal{G}$ and $X_{\mathrm{pa}_j^{\mathcal{G}}}$ are the parent variables of X_j.
The $p(X_j | X_{\mathrm{pa}_j^{\mathcal{G}}})$ are also called conditional probability tables (CPTs).

条件概率表

Note that the conditional probabilities sum up to one in their first variable:

$$\sum_{X_j} p(X_j | X_{\mathrm{pa}_j^{\mathcal{G}}}) = 1$$

Note 4.3

A Bayesian network induces a joint distribution over $X_1, \ldots, X_n$:

$$p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_j | X_{\mathrm{pa}_j^{\mathcal{G}}})$$

Bayesian networks aka Bayes nets, belief networks (3)

Compare Peters, Def 6.32 of causal graphical model

Note 4.4

The product rule for n variables

$$p(x_1, \ldots, x_n) = \prod_{j=1}^{n} p(x_j | x_1, \ldots, x_{j-1})$$

creates a factorization of the joint distribution for any variable ordering/permutation π:

为任何变量排序/变异π创建一个联合分布的因子化。

$$p(x_1, \ldots, x_n) = \prod_{j=1}^{n} p(x_{\pi(j)} | x_{\pi(1)}, \ldots, x_{\pi(j-1)})$$

Thus any fully connected DAG together with any joint distribution forms a Bayesian network (which is not very interesting...).

E.g. ...

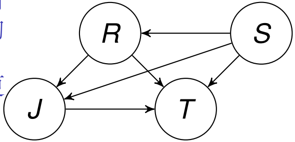
因此，任何完全连接的DAG与任何联合分布一起形成一个贝叶斯网络（这不是很有趣......）。

Bayesian networks aka Bayes nets, belief networks (3)

Without leaving out arrows it is also a Bayes net:

$$p(T, J, R, S) = p(T|J, R, S)\ p(J|R, S)\ p(R|S)\ p(S)$$

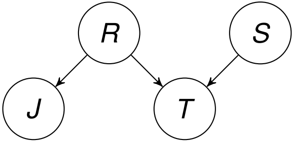
因此，对于任何变量排序，任何分布都可以写成一个完全连接的贝叶斯网。
然而，如果撇开箭头，效率会更高，但会带来限制。



Thus any distribution can be written as a fully connected Bayes net for any variable ordering.

However, leaving out arrows if more efficient, but imposes constraints:

$$p(T, J, R, S) = p(T|R, S)\ p(J|R)\ p(R)\ p(S)$$



How can we characterize those constraints?

Measuring relevance between variables (1)

衡量变量之间的相关性

Definition 4.5 (independence)

Two variables A and B are independent, if and only if their joint distributions factorizes into so-called marginal distributions, i.e.

两个变量A和B是独立的，当且仅当它们的联合分布被分解为所谓的边际分布，即

$$p(A, B) = p(A)\, p(B)$$

在这种情况下，p(A|B)=p(A)，这在直觉上也是合理的。符号。换句话说，关于B的信息并不能提供关于A的信息，反之亦然。

In that case $p(A|B) = p(A)$, which intuitively makes sense as well. Notation: $A \perp\!\!\!\perp B$. In words, information about B doesn't give information about A and vice versa.

Note that $p(R|S) = p(R)$ implies $p(R, S) = p(R)\, p(S)$.

Example:

- Two coins.

$$A = \text{coin 1 shows heads}$$
$$B = \text{coin 2 shows heads}$$

A = 硬币1显示正面
B = 硬币2显示正面

Then $A \perp\!\!\!\perp B$. 不相关事件

Measuring relevance between variables (2)

条件独立性
在给定变量C的情况下，两个变量A和B是有条件独立的，当且仅当它们的条件分布因子化。

Definition 4.6 (conditional independence)

Two variables A and B are conditionally independent given variable C, if and only if their conditional distribution factorizes,

$$p(A, B|C) = p(A|C)\, p(B|C)$$

In that case we have $p(A|B, C) = p(A|C)$, i.e. in light of information C, B doesn't tell us about A. Notation: $A \perp\!\!\!\perp B \mid C$

根据信息C，B并没有告诉我们关于A的信息

Example:

- Two coins and a bell.

$$A = \text{coin 1 shows heads}$$
$$B = \text{coin 2 shows heads}$$
$$C = \text{bell rings if both coins show the same result}$$

A = 硬币1显示为正面
B = 硬币2显示为正面
C = 如果两个硬币都显示相同的结果，则铃声响起

Then $A \perp\!\!\!\perp B$ and $A \perp\!\!\!\perp C$ and $B \perp\!\!\!\perp C$,
but $A \not\perp\!\!\!\perp B \mid C$ and $A \not\perp\!\!\!\perp C \mid B$ and $B \not\perp\!\!\!\perp C \mid A$.

Measuring relevance between variables (3)

定义4.7（条件独立性 **Definition 4.7 (conditional independence)**

在给定一组变量C的情况下，两组变量A和B是条件独立的，当且仅当它们的条件分布因数化

Two sets of variables $\mathcal{A}$ and $\mathcal{B}$ are conditionally independent given a set of variables $\mathcal{C}$, if and only if their conditional distribution factorizes,

$$p(\mathcal{A}, \mathcal{B} | \mathcal{C}) = p(\mathcal{A}|\mathcal{C})\ p(\mathcal{B}|\mathcal{C})$$

where for $\mathcal{A} = \{A_1, A_2, \ldots, A_n\}$, we define $p(\mathcal{A}) := p(A_1, A_2, \ldots, A_n)$. We write $\mathcal{A} \perp\!\!\!\perp \mathcal{B} \mid \mathcal{C}$.

Note:

- The two previous definitions are special cases of the latter:

$$\begin{aligned} A \perp\!\!\!\perp B \quad &\text{iff} \quad \{A\} \perp\!\!\!\perp \{B\} \\ A \perp\!\!\!\perp B \mid C \quad &\text{iff} \quad \{A\} \perp\!\!\!\perp \{B\} \mid \{C\} \end{aligned}$$

Tracy, Jack and the wet grass — representation

from Barber 2012, 3.1.1

$$p(T, J, R, S) = p(T|R, S)\, p(J|R)\, p(R)\, p(S)$$

Conditional probability tables (CPTs)

$$p(R{=}1) = 0.2 \qquad p(S{=}1) = 0.1$$

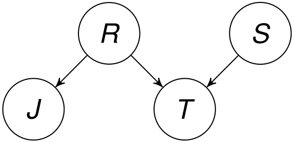
$$p(J{=}1|R{=}1) = 1 \qquad p(J{=}1|R{=}0) = 0.2$$

$$p(T{=}1|R{=}1, S{=}0) = 1 \qquad p(T{=}1|R{=}1, S{=}1) = 1$$

$$p(T{=}1|R{=}0, S{=}1) = 0.9 \qquad p(T{=}1|R{=}0, S{=}0) = 0$$

Graphical representation

图形表示法



我们可以仅从图中推断
出哪些独立因素？

What independencies can we infer only from the graph?

Conditional independencies in three variable networks

see also Barber 2012, 3.3.2

The four isolated paths in DAGs

DAGs中的四条孤立的路径

(i)	$A \to B \to C$	$p(A, B, C) = p(C \vert B)\, p(B \vert A)\, p(A)$
(ii)	$A \leftarrow B \leftarrow C$	$p(A, B, C) = p(A \vert B)\, p(B \vert C)\, p(C)$
(iii)	$A \leftarrow B \to C$	$p(A, B, C) = p(A \vert B)\, p(C \vert B)\, p(B)$
(iv)	$A \to B \leftarrow C$	$p(A, B, C) = p(B \vert A, C)\, p(A)\, p(C)$

…imply the following independencies (with elementary proofs):

(i) $A \perp\!\!\!\perp C \mid B$ (ii) $A \perp\!\!\!\perp C \mid B$

(iii) $A \perp\!\!\!\perp C \mid B$ (iv) $A \perp\!\!\!\perp C$

However, they do not necessarily imply dependences, such as:

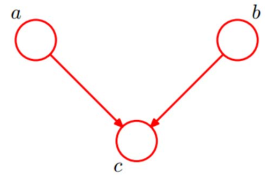
然而，它们不一定意味着依赖性，例如。

(i) $A \not\perp\!\!\!\perp C$ (ii) $A \not\perp\!\!\!\perp C$

(iii) $A \not\perp\!\!\!\perp C$ (iv) $A \not\perp\!\!\!\perp C \mid B$

Those might be true or wrong dependent on conditional probability tables.

贝叶斯网络的第一种结构形式如下图所示

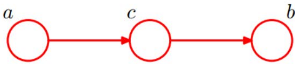


所以有：P(a,b,c) = P(a)*P(b)*P(c|a,b)成立，化简后可得：

$$\sum_{c} P(a,b,c)=\sum_{c} P(a)*P(b)*P(c\mid a,b)$$
$$\Rightarrow P(a,b)=P(a)*P(b)$$

2.3.3 形式3：head-to-tail

贝叶斯网络的第三种结构形式如下图所示：



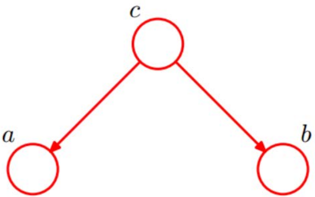
有：P(a,b,c)=P(a)*P(c|a)*P(b|c)。

化简后可得：

$$\begin{aligned}
& P(a,b\mid c) \\
=\ & P(a,b,c)/P(c) \\
=\ & P(a)*P(c\mid a)*P(b\mid c)\ /\ P(c) \\
=\ & P(a,c)*P(b\mid c)\ /\ P(c) \\
=\ & P(a\mid c)*P(b\mid c)
\end{aligned}$$

即：**在c给定的条件下，a，b被阻断(blocked)，是独立的**，称之为head-to-tail条件独立。

贝叶斯网络的第二种结构形式如下图所示



有P(a,b,c)=P(c)*P(a|c)*P(b|c)，则：P(a,b|c)=P(a,b,c)/P(c)，然后将P(a,b,c)=P(c)*P(a|c)*P(b|c)带入上式，得到：P(a,b|c)=P(a|c)*P(b|c)。

即**在c给定的条件下，a，b被阻断(blocked)，是独立的**，称之为tail-to-tail条件独立，对应本节

插一句：这个head-to-tail其实就是一个链式网络，如下图所示：

在xi给定的条件下，xi+1的分布和x1,x2...xi-1条件独立。即：xi+1的分布状态只和xi有关，和其他变量条件独立，这种顺次演变的随机过程，就叫做马尔科夫链（Markov chain）。且有：

$$P(X_{n+1}=x|X_0,X_1,X_2,\ldots,X_n)=P(X_{n+1}=x|X_n)$$

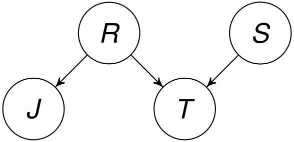
Tracy, Jack and the wet grass — cond. independencies

from Barber 2012, 3.1.1

Conditional independencies

(i) $A \rightarrow B \rightarrow C$	imply	$A \perp\!\!\!\perp C \mid B$
(ii) $A \leftarrow B \leftarrow C$	imply	$A \perp\!\!\!\perp C \mid B$
(iii) $A \leftarrow B \rightarrow C$	imply	$A \perp\!\!\!\perp C \mid B$
(iv) $A \rightarrow B \leftarrow C$	imply	$A \perp\!\!\!\perp C$

Graphical representation



我们可以仅从图中推断出哪些独立因素?

What independencies can we infer only from the graph?

Answer: $J \perp\!\!\!\perp T \mid R$ and $R \perp\!\!\!\perp S$. But also $J \perp\!\!\!\perp S \mid R$, $J \perp\!\!\!\perp S$, $J \perp\!\!\!\perp S|R, T$ with the d-separation criterion (stay tuned).

与d分离 标准（敬请关注）。

A sophisticated criterion on graphs 图形上的一个复杂标准

copied from Peters, Def 6.1

Definition 4.8 (Pearl's d-separation)

Given a DAG $\mathcal{G}$.

节点i1和im之间的路径被一个集合S阻断

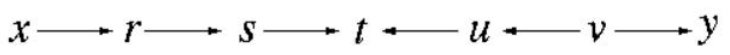
1. *A path between nodes i_1 and i_m is **blocked by a set** S (with $i_1 \notin S$ and $i_m \notin S$), whenever there is a node i_k, such that one of the following two possibilities holds:*
   - *$i_k \in S$ and*
     $$\begin{aligned} & i_{k-1} \to i_k \to i_{k+1} \\ \text{or } & i_{k-1} \leftarrow i_k \to i_{k+1} \\ \text{or } & i_{k-1} \leftarrow i_k \leftarrow i_{k+1} \end{aligned}$$
   - *neither i_k nor any of its descendents is in S and*

     ik和它的任何子孙都不在S中，并且
     $$i_{k-1} \to i_k \leftarrow i_{k+1}$$
2. *Two disjoint subsets of vertices A and B are **d-separated** by a third (also disjoint) subset S if every path between nodes in A and B is blocked by S. We write*
   $$A \perp\!\!\!\perp_{\mathcal{G}} B \mid S$$

首先，这里我们先说明一下path，我们说两个结点之间的path的时候是不管他们之间边的方向的。

没有条件集的独立性

规则1：*如果x到y的任一path(路径)都经过collider(碰撞点)，则x和y独立。注意，这里的路径是忽略边的方向的，而碰撞点是指有多个箭头指向的它的节点，即类似于下图的s->t<-u。*



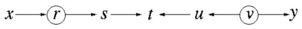
现在我们看看这个图变量之间的独立性是怎样的。先考虑x和t的路径：x-r-s-t。这条路径中并不存在碰撞点，所以x和t不独立，同样地，t和y、u、v都不独立。然而，对于变量x到y而言，x，y之间的路径必然会经过碰撞点t，所以x和y是独立，同样地，在碰撞点两侧的变量都是独立的，比如x和v，s和u，r和u也是独立的。

https://zhuanlan.zhihu.com/p/72011891

一般的条件独立

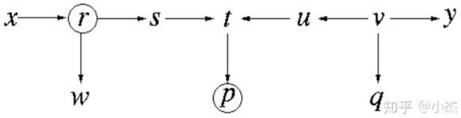
那么如果一条路径中没有碰撞点，怎么才能让他们独立呢？我们还能使用条件独立的性质，只要条件集Z能够将一条路径block掉，那么就可以让两个变量独立。注意的是，当碰撞点或碰撞点的子代出现在条件集的时候要小心，很有可能会导致不同的结果，这个问题我们留到规则3。现在我们先假设条件集中不存在碰撞点。

规则2：*当x到y的之间的任一路径都经过Z中的节点，且Z并不包含碰撞点或碰撞点的子代，则x和y独立。*



如图，设结点集Z={r, v}（图中画圈的节点），根据规则2，x和s是条件独立的，因为x和s之间的路径被block掉了，同样地，u和y也是条件独立的。

规则3：*当碰撞点或碰撞点的子孙节点为集合Z的成员时，该碰撞点不再截断路径。*

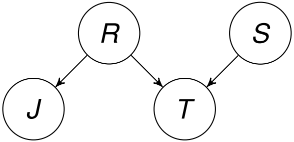


设结点集Z={r, p}（图中画圈的节点），根据规则3，在给定Z的情况下，s和y不独立，因为t的孩子节点p在集合Z中，所以t没有办法像规则1一样截断路径s-t-u-v-y，与此相反，条件集p使得s和u变得不独立了。然而在这里，x和u是独立的，虽然t不能截断它，但是r可以截断它（根据规则2）。

Tracy, Jack and the wet grass — cond. independencies

from Barber 2012, 3.1.1

Graphical representation



What independencies can we infer only from the graph?

我们可以仅从图中推断出哪些独立性？答案：

Answer:

- $J \perp\!\!\!\perp T \mid R$: because the path from J to T is d-separated by observing R, so all paths between them are d-separated
- $R \perp\!\!\!\perp S$: because the path from R to S is d-separated, if we do not observe T, so all paths ...
- $J \perp\!\!\!\perp S \mid R$: because the path from J to S is d-separated by observing R, so all paths ...
- $J \perp\!\!\!\perp S$, because the path from J to S is d-separated by not observing T, so all paths ...
- $J \perp\!\!\!\perp S|R, T$, because the path from J to S is d-separated by observing R, so all paths ...

Linking graphs and distributions

Peters, Def 6.21

Definition 4.9

2. 就DAG G而言，如果每个变量都是独立于其非后代的，那么本地马尔科夫属性是独立于其父代的，并且

Given a DAG $\mathcal{G}$, a joint distribution p satisfies

1. *the **global Markov property** wrt. the DAG $\mathcal{G}$ if*

   全局马尔科夫属性

   $$A \perp\!\!\!\perp_{\mathcal{G}} B \mid C \implies A \perp\!\!\!\perp B \mid C$$

   for all disjoint vertex sets A, B and C and where $A \perp\!\!\!\perp B \mid C$ describes cond. ind. wrt. p.

2. *the **local Markov property** wrt. the DAG $\mathcal{G}$ if each variable is independent of its non-descendants given its parents, and*
3. *the **Markov factorization property** wrt. the DAG $\mathcal{G}$ if*

   马尔科夫因子化属性

   $$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_j \mid \mathrm{pa}_j^{\mathcal{G}})$$

如果某个联合分布有一个密度p，那么所有马尔科夫属性(来自前面的定义)都是等价的。

Theorem 4.10 (Equivalence of Markov properties)

If some joint distribution has a density p then all Markov properties (from the previous def.) are equivalent.

arkov property for undirected graphs

- We say $\mu(\cdot)$ satisfy the **global Markov property** (G) w.r.t. a graph G if for any partition (A, B, C) such that B separates A from C,

$$\mu(x_A, x_C | x_B) = \mu(x_A | x_B)\, \mu(x_C | x_B)$$

- We say $\mu(\cdot)$ satisfy the **local Markov property** (L) w.r.t. a graph G if for any $i \in V$,

$$\mu(x_i, x_{\text{rest}} | x_{\partial i}) = \mu(x_i | x_{\partial i})\, \mu(x_{\text{rest}} | x_{\partial i})$$

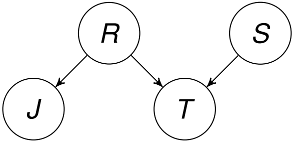
- We say $\mu(\cdot)$ satisfy the **pairwise Markov property** (P) w.r.t. a graph G if for any $i, j \in V$ that are not connected by an edge

$$\mu(x_i, x_j | x_{\text{rest}}) = \mu(x_i | x_{\text{rest}})\, \mu(x_j | x_{\text{rest}})$$

viously: (G) $\Rightarrow$ (L) $\Rightarrow$ (P)

Example

A distribution $p(R, S, T, J)$ is Markovian wrt to graph $\mathcal{G}$



if either (global Markov property)

$$J \perp\!\!\!\perp T \mid R$$
$$R \perp\!\!\!\perp S$$
$$J \perp\!\!\!\perp S \mid R$$
$$J \perp\!\!\!\perp S$$
$$J \perp\!\!\!\perp S|R, T$$

or if (Markov factorization property)

$$p(T, J, R, S) = p(T|R, S)\; p(J|R)\; p(R)\; p(S)$$

Summary

- A joint distribution, such as $p(A, B, C, \ldots, Z)$ requires lots of parameters, thus lots of memory.
- Exploit conditional independencies between variables.
- Factorize the joint distribution along a graph.
- There is a (somewhat complicated) criterion on graphs which corresponds to conditional independence

Main idea: combine probabilities and graphs