MLaPP Chapter 10 Bayes nets 贝叶斯网络

`MLaPP`

10.1 Introduction

书里开头就引用了迈克尔·乔丹对图模型的理解，他说处理复杂系统有两个原则，模块性

（modularity）个抽象性（abstraction），而概率论（probability theory）则通过因式分解（factorization）和求平均（averaging）深刻地实现了这两个原则。

概率图模型有三大任务：表征（representatino），推断（Inference），学习（Learning），表征就是怎样样用图模型表示概率分布，有有向图和无向图两种方法，推断就是怎么从已经学到的联合概率分布中推断出条件概率，学习就是怎么用数据学习出模型中的参数。具体内容分类如下，

- 表征，Representation
  - 有向图模型，又叫贝叶斯网络
    - Undirected Graphical Models, BN, Bayes Nets [chap10]
  - 无向图模型，又叫马尔可夫随机场
    - Directed Graphical Models, MRF, Markov random fields [chap19]
- 推断，Inference
  - 确切推断，Exact Inference [chap20]
  - 变分推断，Variational inference [chap21]
  - 更多变分推断，More Variational inference [chap22]
  - 蒙特卡洛推断，Monte Carlo inference [chap23]
  - 马尔可夫链蒙特卡洛推断，Markov chain Monte Carlo (MCMC) inference [chap24]
- 学习，Learning
  - EM algorithm [chap11]
  - Graphical model structure learning [chap26]
  - Latant variable models for discrete data [chap27]

概率图模型有什么好的参考资料？公开课，书，网页都可以。

10.1.1 Chain rule 链式法则

一般来说，我们处理的监督问题就是试图去拟合这样的一个函数，$\boldsymbol{f}:\mathbb{R}^n \rightarrow \mathbb{R}$ 也就是说输入 $\mathbf{x}$ 一般都是一个多维度的特征向量。在朴素贝叶斯模型中，我们曾假设不同的特征之间的相互独立的，如果不独立的话，可以用链式法则来计算一个序列的概率，

$$p(x_{1:V}) = p(x_1)p(x_2|x_1)p(x_3|x_2,x_1)\cdots p(x_V|x_{1:V-1})$$

注意上面的公式都是条件概率，为了表示的简介省略了参数 $\boldsymbol{\theta}$ 而已。链式法则可以用语言模

型来解释（language model），整个序列的概率就是某个句子的概率。可以发现上面的条件概率越往后越复杂。

我们来分析一下要计算这个联合概率的复杂度是多少。假设每个维度的变量都一样地拥有 K 个状态，那么 $p(x_1)$ 有 $K-1$ 个变量，复杂度是 $O(K)$；接着 $p(x_2|x_1)$ 要用大小为 $O(K^2)$ 的 **随机矩阵（stochastic matrix）** T 来表达，且矩阵元素满足

$$p(x_2=j|x_1=i)=T_{ij},\quad T_{ij}\in[0,1],\quad \sum_{ij}T_{ij}=1$$

同样地，$p(x_3|x_2,x_1)$ 要用 $O(K^3)$ 的 **条件概率表（conditional probability tables or CPTs）** 来表示。到最后的一个概率，需要用 $O(K^V)$ 个参数，这个复杂度是不可以接受的。

【TODO：这里是不是可以给出一个条件概率分布的具体例子，所以可以直观地感受一下】

一种解决方法是用 CPD，Condiational probability distribution 来替代 CPT，比如 multinomial logistic regression，

$$p(x_t=k|\mathbf{x}_{1:t-1})=\mathcal{S}(\mathbf{W}_t\mathbf{x}_{1:t-1})_k$$

参数复杂度为 $O(K^2V^2)$，为什么呢？考虑 $\mathbf{W}_t$ 和 $\mathbf{x}_{1:t-1}$ 之间做的是内积操作，参数数量应该是一样的，都是 $K(t-1)$，那么 x_t 一共要取 K 种状态，所以 $p(x_t|\mathbf{x}_{1:t-1})$ 有 $K(t-1)K$ 个参数。那么 $t=1,\cdots,V$ 的话，一共就是

$$K^2(0+1+2+\cdots+(V-1))=K^2V(V-1)/2$$

个参数，所以参数的复杂度为 $O(K^2V^2)$。

然而这种模型对有些预测问题不是很适用，因为每个变量都要依靠前面所有的变量。（不是很理解这句话。）

10.1.2 Conditioanl independence 条件独立

可以从随机变量的独立推广到条件独立，比如 X 和 Y 在给定 Z 的条件下独立，可以写作

$$X\perp Y|Z\iff p(X,Y|Z)=p(X|Z)p(Y|Z)$$

我们可以利用条件独立，对前面讲过的链式法则做一些假设，比如令将来的状态在给定现在状态的条件下与过去的状态无关，又叫做是 **一阶马尔可夫假设（first order Markov assumption）**，写成表达式就是，

$$x_{t+1} \perp \mathbf{x}_{1:t-1} | x_t$$

那么联合概率转换成了，

$$p(x_{1:V}) = p(x_1) \prod_{t=2}^{V} p(x_t | x_{t-1})$$

这个就是 **一阶马尔可夫链（first order Markov chain）**，可以通过初始状态分布 $p(x_1 = i)$ 和状态转移矩阵 $p(x_t = j | x_{t-1} = i)$ 来确定。后面的小节还会提到二阶和更高阶的情况。

10.1.3 Graphical models 图模型

一阶的马尔可夫可以用来处理以为的一维的序列数据，也可以定义二维的模型处理图片，三维的模型处理视频数据。当推广到任意维的时候，图模型就诞生了。

图模型就是一种通过做条件独立假设（CI assumption）来表示联合概率分布的方法。这一章主要讲有向图模型，即贝叶斯网络，而无向图模型放在了第 19 章。

10.1.4 Graph terminology 图术语

图模型是图论和概率论的结合，先来回顾一下图论的一些知识。

- graph $G = (V, E)$, nodes or vertices, edges, adjacency matrix（邻接矩阵）
- parent, child, family, root ,leaf, ancestors, descendants, neighbors
- degree, in-degree, out-degree, cycle or loop（环）
- DAG, directed acyclic graph（有向无环图）, topological ordering（拓扑排序）
- path or trail（路径或迹）, tree, polytree, moral directed tree, forest
- subgraph, clique, maximal clique

树和有向无环图的关系

- 树是一种特殊的有向无环图，DAG 不一定是树
  - github 的版本管理就是通过 DAG 来做的。
- DAG 可以有多个根节点（无父节点），但是树只有唯一的一个 root
- DAG 的某个节点可以有多个父节点

- 也叫 polytree，否则叫 moral directed tree
- 造成了两个节点可以有多条通路
- 如果把有向图的箭头去掉，DAG 可能会存在环，树则没有

DAG 的拓扑排序（topological ordering）

拓扑排序的名词一定要跟 DAG 联系起来才有意义。就是说，在图的顶点构成的所有排列中，对于图中任意边 $A \to B$ 都满足 A 排在 B 前面的排列都是合法的拓扑排序。可以参考这里的 讲义 。

团（clique）的定义
团就是无限图的完全子图，极大团（maximal clique）就是无法通过添加点和边仍然满足团定义的完全子图，最大团（maximum clique）就是最大的那个极大团，可能有多个。参考这里

10.1.5 Directed graphical models 有向图模型

有向图模型，字面上的理解就是图是 DAG 的图模型。有三个名字，

- 贝叶斯网络，Bayesian Network
- 置信网络，Belief Network
- 因果网络，Causal Network

然而这几个名字的意思和这个模型都没啥直接关系，所以作者决定就叫它有向图模型好了。

DAG 一定可以有拓扑排序，即父节点一定会在孩子节点前面，我们可以定义 ordered Markov property 假设，即每个节点在给定父节点的条件下，和其他所有先驱节点条件无关，表示如下，

$$x_s \perp \mathbf{x}_{\text{pred}(s) \setminus \text{pa}(s)} | \mathbf{x}_{\text{pa}(s)}$$

其中 $\text{pred}(s)$ 表示 s 的所有先驱节点（predecessors），$\text{pa}(s)$ 是 s 的父节点。图 10.1 的联合概率就可以重写了。

一般地，贝叶斯网络可以写成这样的形式，

$$p(\mathbf{x}_{1:V}|G) = \prod_{t=1}^{V} p(x_t|\mathbf{x}_{\text{pa}(t)})$$

其中 $p(x_t|\mathbf{x}_{\text{pa}(t)})$ 是 CPD，Conditional Probablistic Distribution。假设每个点都只有 $O(F)$ 个父节点，那么整个模型的参数复杂度就是 $O(VK^F)$，比原来的 $O(K^V)$ 大大降低了。

10.2 Examples 例子

这一小节讲述 PGM 的应用。

10.2.1 Naive Bayes classifiers 朴素贝叶斯分类器

回忆一下朴素贝叶斯分类器，该模型假设了特征维度之间都是相互独立的，那么有

$$p(y, \mathbf{x}) = p(y)p(x_{1:D}|y) = p(y)\prod_{j=1}^{D} p(x_j|y)$$

如果想要抓住不同特征之间的相关性（correlation），可以考虑一种图模型，叫做 **tree-augmented naive Bayes classifier**，or **TAN**，此模型是 two-fold，可以很容易地用 Chow-Liu 算法（见26.3小节）找到最优的树结构，而且处理丢失特征（见20.2小节）也很简单。

10.2.2 Markov and hidden Markov models 马尔可夫和隐马尔可夫模型

我们可以做更高阶的马尔可夫假设，比如二阶的情况，正如图 10.3 所示，可以这样表示，

$$p(\mathbf{x}_{1:T}) = p(x_1, x_2)\prod_{t=3}^{T} p(x_t|x_{t-1}, x_{t-2})$$

不幸的是，二阶的马尔可夫假设对于长范围关系是不充分的，而更高阶的假设会带来参数复杂度的爆炸，因此有另一种常见的模型，是假设在观察变量中，有另一个隐藏变量在背后，这种叫做隐马尔可夫模型（HMM, Hidden Markov Model）。如果用 z_i 表示隐变量，x_i 表示观察变量的话，可以做两个一阶马尔可夫假设，

- **齐次马尔可夫假设**

    - 当前的隐变量只和前一个隐变量有关，得到转移模型（transition model）
    - 即 $p(z_t|z_{t-1}) = p(z_t|z_{1:T}, \mathbf{x}_{1:T})$ ，可以对 CPD $p(z_t|z_{t-1})$ 进行建模
- **观测独立性假设**
    - 当前的观察变量只和当前的隐变量有关，得到观察模型（observation model）
    - 即 $p(\mathbf{x}_t|z_t) = p(\mathbf{x}_t|z_{1:T}, \mathbf{x}_{1:T})$ ，可以对 CPD $p(\mathbf{x}_t|z_t)$ 进行建模

学了随机过程（stochastic process）的同学可能会了解的多一点，隐马尔可夫模型是一种最简单的动态贝叶斯网络（DBN, Dynamic Bayesian Network），是一种时间序列上的模型方法，是处理时间序列问题的一种应用最广泛的模型之一（其他的还有 CRF，RNN 等）。第 17 章会细讲这个模型，在这里举几个例子来简单说明一下其用途。

- **标注问题**，见李航的《统计学习方法》第 1.9 小节，输入观测序列，输出状态序列
    - 词性标注（part of speech tagging）
    - 分词，见 结巴分词
- **语音识别**，ASR, Automatic Speech Recognition
    - 令 $\mathbf{x}_t$ 表示语音信号提取的特征，令 z_t 表示说的单词（字），转移模型 $p(z_t|z_{t-1})$ 表示语言模型（language model），观察模型 $p(\mathbf{x}_t|z_t)$ 表示声学模型（acoustic model）。
- Video Activity Recognition
- Optical Character Recognition
- Protein Sequence alignment

10.2.3 Medical diagnosis 医学诊断

10.2.4 Genetic linkage analysis * 基因连锁分析

10.2.5 Directed Gaussian graphical models * 有向高斯图模型

10.3 Inference 推断

定义概率图模型和联合概率分布，主要的用途是拿来做概率推断（Probabilistic Inference）。先举个例子，在 HMM 中，用 $\mathbf{x}_v$ 表示可见变量，用 $\mathbf{x}_h$ 表示隐藏变量。我们假设已经得到了联合概率分布 $p(\mathbf{x}_h, \mathbf{x}_v|\boldsymbol{\theta})$，目标则是推断出隐变量的条件概率分布，

$$p(\mathbf{x}_h|\mathbf{x}_v,\boldsymbol{\theta}) = \frac{p(\mathbf{x}_h,\mathbf{x}_v|\boldsymbol{\theta})}{p(\mathbf{x}_v|\boldsymbol{\theta})} = \frac{p(\mathbf{x}_h,\mathbf{x}_v|\boldsymbol{\theta})}{\sum_{\mathbf{x}'_h} p(\mathbf{x}'_h,\mathbf{x}_v|\boldsymbol{\theta})}$$

注意下面的分母 $p(\mathbf{x}_v|\boldsymbol{\theta})$，其实就是训练集数据的似然概率，又叫做是 **probability of the evidence**，可以写成是边缘似然来求解。

还有另一种需求，就是可能只有部分的隐变量是我们感兴趣的，比如把隐变量划分成 **query variables** $\mathbf{x}_q$ 和 **nuisance variables** $\mathbf{x}_n$ 两部分。我们只关心 $\mathbf{x}_q$ 而不关心 $\mathbf{x}_n$，也可以通过 **marginalizing out** 来求边缘概率，

$$p(\mathbf{x}_q|\mathbf{x}_v,\boldsymbol{\theta}) = \sum_{\mathbf{x}_n} p(\mathbf{x}_q,\mathbf{x}_n|\mathbf{x}_v,\boldsymbol{\theta})$$

上面的方法都是确切推断（exact inference）。精确推断常见的方法有变量消去（variable elimination）和信念传播（Belief Propagation，又叫做和积算法，Sum-Product Algorithm）。

这种方法会消耗指数级别的复杂度，适用范围有限，因此有其他的近似推断（approximation inference）的方法。近似推断又主要有两大类，第一类是采样的方法，比如马尔科夫链蒙特卡洛采样（MCMC, Markov Chain Monte Carlo sampling）；第二类是使用确定性近似完成近似推断，比如变分推断（variational inference）。

10.4 Learning 学习

图模型中的学习（Learning）一般指的是学习模型中的参数，一般也是通过 MAP 来估计参数，

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \log p(\mathbf{x}_{i,v}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta})$$

其中 $\mathbf{x}_{i,v}$ 指的是隐变量中的第 i 个样例。若先验 $p(\boldsymbol{\theta}) \propto 1$，那么退化成 MLE。

一般的概率图模型文献中，都会把 Inference 和 Learning 区分开来，但是在贝叶斯学派的观点来看，参数和隐变量都是变量，两者没有区别。

10.4.1 Plate notation 盘子表示法

盘子表示法的示意图如图 10.7 所示。当一组变量 $\boldsymbol{X_1, \cdots, X_N}$ 是独立同分布的（iid, independently identically distribution），即都是从同一个分布 $\boldsymbol{\theta}$ 种产生的，那么可以用一个盘子（Plate）把这些变量圈起来，用一个循环表示出来，像书里那样子画的，相当于是一种语法糖（syntactic sugar）。那么把每个重复的画出来的方式就称作展开（unrolled）的模型。我们用图模型表示随机变量之间的联合概率，

$$p(\boldsymbol{\theta}, \mathcal{D}) = p(\boldsymbol{\theta}) \left[\prod_{i=1}^{N} p(\mathbf{x}_i | \boldsymbol{\theta}) \right]$$

另一个复杂一点的嵌套盘子（nested plate）表示法在 10.8 中画出来了。

我们先来回归一下朴素贝叶斯分类器，其假设为所有的特征之间都是相互条件独立的，若 class conditional density 是多努利分布，可以写成连乘的形式，

$$p(\mathbf{x}|y=c) = \prod_{j=1}^{D} \mathrm{Cat}(x_j|\boldsymbol{\theta}_{jc}) = \prod_{j=1}^{D} p(x_j|y=c, \boldsymbol{\theta}_{jc})$$

其中，

- $(\mathbf{x}_{ij})_{N \times D}$ 为训练集第 i 个样本的第 j 个特征维度，共 N 个样本，D 个特征维度
- $(\boldsymbol{\theta}_{jc})_{D \times C}$ 为模型学到的参数，表示第 c 类中出现特征 j 的概率
  - 如 bag-of-words 模型，表示第 c 类文档（比如广告邮件）出现第 j 个单词（比如“打折”）的概率。
  - NBC Model Fitting result: $\hat{\theta}_{jc} = N_{jc}/N_c$ （公式 3.58）
- $\boldsymbol{\pi}$ 是 class prior，类别的先验
  - NBC Model Fitting result: $\hat{\pi}_c = N_c/N$（公式 3.57）

【此处该有图】

- 图①，针对所有的样本，每个维度都是条件独立的
- 图②，针对所有的样本的某个维度，比如 $\boldsymbol{x_{i1}}$，对所有的类别的贡献程度是不同的，比如 "discount" 这个词，对 “广告”，“正常”，“色情”等邮件类别的贡献概率是不同的，正相关。
- 图③，由图①把维度 $\boldsymbol{j:D}$ 放到盘子里得到的
- 图④，由图③把 $\boldsymbol{i:N, c:C}$ 放进盘子里得到的

可以看到嵌套的表示法更简洁。

10.4.2 Learning from complete data 从完整数据中学习

所谓的完整的数据（complete data），指的是没有丢失的数据或者没有隐变量（或者说隐变量都观察到了），那么训练集上的似然概率是，

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^{N} p(\mathbf{x}_i|\boldsymbol{\theta}) = \prod_{i=1}^{N}\prod_{t=1}^{V} p(x_{\mathrm{it}}|\mathbf{x}_{i,\mathrm{pa}(t)}, \boldsymbol{\theta}_t) = \prod_{t=1}^{V} p(\mathcal{D}_t|\boldsymbol{\theta}_t)$$

补充说明：中间的两个连乘项，就是概率图模型的因式分解，

$$p(\mathbf{x}|\boldsymbol{\theta}) = \prod_{t=1}^{V} p(x_t|\mathbf{x}_{\mathrm{pa}(t)}, \boldsymbol{\theta}_t)$$

考虑完整的贝叶斯推断，若也有可以因式分解的先验（factored prior），
$p(\boldsymbol{\theta}) = \prod_{t=1}^{V} p(\boldsymbol{\theta}_t)$，那么对应的后验也可以因式分解，

$$p(\boldsymbol{\theta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = \prod_{t=1}^{V} p(\mathcal{D}_t|\boldsymbol{\theta}_t)p(\boldsymbol{\theta}_t)$$

如果我们用表格（tabular）的方法表达所有的 CPD，那么表格里的每个项都是待学习的参数 θ，若用多努利分布来拟合模型，即 $x_t|\mathbf{x}_{\mathrm{pa}(t)} = c \sim \mathrm{Cat}(\boldsymbol{\theta}_{tc})$，可以定义

$$\theta_{tck} \triangleq p(x_t = k|\mathbf{x}_{\mathrm{pa}(t)} = c)$$

- $k = 1 : K_t$ 表示节点 t 的所有状态
- $c = 1 : C_t$ 其中 $C_t \triangleq \prod_{s\in \mathrm{pa}(t)} K_s$ 表示父节点组合的数目
- $t = 1 : T$ 表示节点（node）的个数

显然，有 $\sum_k \theta_{tck} = 1$ 即所有状态的累加和是1.

考虑在每个概率上加一个狄利克雷先验，$\boldsymbol{\theta}_{tc} \sim \mathrm{Dir}(\boldsymbol{\alpha}_{tc})$，那么后验就可以是
$\boldsymbol{\theta}_{tc}|\mathcal{D} \sim \mathrm{Dir}(\mathbf{N}_{tc} + \boldsymbol{\alpha}_{tc})$ 其中充分统计量定义如下，

$$N_{tck} \triangleq \sum_{i=1}^{N} \mathbb{I}(x_{i,t} = k, x_{i,\mathrm{pa}(t)} = c)$$

表示第 t 个节点在状态 k 且父节点是 c 的统计量。那么根据狄利克雷分布的均值公式可知，

$$\bar{\theta}_{tck} = \frac{N_{tck} + \alpha_{tck}}{\sum_{k'}(N_{tck'} + \alpha_{tck'})}$$

下面的例子讲的是，如何从数据集中应用上面的公式估计条件概率 $p(x_4|x_2, x_3)$（Really？）参数，就略过啦~

10.4.3 Learning with missing and/or latent variables

如果我们的数据有缺失（比如某个传感器故障，导致 $\mathbf{x}_i$ 的某个维度的数值无法测量到），或者含有隐变量，那么似然函数就无法因式分解，而且是非凸的。如果用 MLE 或者 MAP 只能得到局部最优解。

一般可以用 EM, Expectation-Maximization 这种迭代算法。EM 算法的直观想法是，如果数据是完整的（complete），那么计算参数就会很容易，比如用 MLE；相反如果知道了模型的所有参数，那么补全完整的数据也很容易。（参考李航《统计学习方法》第9章和本书的第11章）

10.5 Conditional independence properties of DGMs

先来讲一下 **I-map** 的概念，这个东西有点绕，感觉是理论学家为了严谨而弄出的一些概念。图模型的一个很重要的核心就是做了很多条件独立性假设（Conditional Indepence Assumptions），比如 $\mathbf{x}_A \perp_G \mathbf{x}_B|\mathbf{x}_C$ 表示在图 G 中给定 C 后，A 和 B 是条件独立的。

所以 I-map 就是个集合，集合里的元素都是一个个条件独立语句，比如 $I(G)$ 表示图 G 中所有的条件独立语句，$I(p)$ 表示分布 p 所有的条件独立语句，那么当且仅当 $I(G) \subseteq I(p)$ 时，我们就可以说图 G 是分布 p 的一个 **I-map**，或者说 p 是关于图 G 的马尔可夫。（We say that G is an I-map (independent map) for p, or that p is Markov wrt G.）换句话说，就是图 G 中做的条件独立语句一定不能比分布 p 中多。

注意全连接的图（完全图？）一定是所有分布的 I-map，因为 I(G) 是空集，没有做任何的条件独立假设。因此有 minimal I-map 的概念，就是说当不存在子图 $G' \subseteq G$ 也是分布 p 的 I-map 时，图是分布 p 的 minimal I-map.

下面讨论图 G 中的所有 $\mathbf{x}_A \perp_G \mathbf{x}_B|\mathbf{x}_C$ 关系。即求解 $I(G)$。

10.5.1 d-separation and the Bayes Ball algorithm (global Markov properties)

在判断图 G 中的两个变量是否是条件独立时，常常用到一个很重要的概念，叫做 d-separation，可以有下面的定义。

当且仅当下面的三个条件满足时，我们会称无向路径 P 被包含证据的点集 E 给 d-分割了（undirected path P is d-separated by a set of nodes E, which contains the evidence），

1. 路径 P 包含链 $s \to m \to t$ 或者 $s \leftarrow m \leftarrow t$，其中 $m \in E$
2. 路径 P 包含像帐篷（tent）或者叉子（fork）一样的结构，即 $s \swarrow^{m} \searrow t$，其中 $m \in E$
3. 路径中包含对撞（collider）或者 v 型结构（v-structure），即 $s \searrow_{m} \swarrow t$，且 m 和其后代（descendant）全部不能出现在 E 中。

为什么要有 d-separate 这个概念呢？假设我们在已知观察变量，或者叫做 evidence 的点集 E 的前提下，想知道点集 A 和点集 B 是否关于 E 条件独立，那么我们就沿着从 A 到 B 的一条路径，如果这条路径被 d-separate，那么说明 A 和 B 关于 E 条件独立。表述如下，

$$\mathbf{x}_A \perp \mathbf{x}_B | \mathbf{x}_E \iff \mathrm{A\ is\ d\text{-}separated\ from\ B\ given\ E}$$

上面讲的就是 **全局马尔可夫性质（global Markov properties）**。书里有提到 Bayes ball algorithm，但是感觉没啥意义，就是前面讲过的思路。

上面的三种条件，我们再来推导一番，

- $X \to Y \to Z$
  - $p(x,y,z) = p(x)p(y|x)p(z|y)$
  - $x \perp z|y$
- $X \leftarrow Y \to Z$
  - $p(x,y,z) = p(y)p(x|y)p(z|y)$
  - $x \perp z|y$
- $X \to Y \leftarrow Z$
  - $p(x,y,z) = p(x)p(z)p(y|x,z)$
  - $x \not\perp z|y \quad \text{but} \quad x \perp z$

注意第三种情况，就是说本来 x 和 z 是独立的，但是看到 y 后，反而不独立了，这种情况就叫做 explaining away, inter-causal reasoning, or Berkson's paradox. 可以举个例子，如果 x, z 表示两次独立的扔硬币事件，y 表示两个硬币点数之和，明显知道了 y 以后，两个事件就不独立了，可以做一些推导。比如已知 $y=1, x=0$ ，那么明显 z 只能取 1 了，即

$$p(z=1|x=0)=0.5 \neq p(z=1|x=0,y=1)=1$$

对于 v-structure 的情况，还有一点要注意，即 boundary conditions，如图 10.10(c) 所示。就是 evidence 不一定要非得是 y，任意 y 的后代，比如 y' 都可以起到同样的作用。

10.5.2 Other Markov properties of DGMs

有了前面的 d-separation 的概念，我们可以对一个PGM中所有的点做一个结论，称作 **有向局部马尔可夫性质（directed local Markov property）**，

$$t \perp \mathbf{nd}(t) \backslash \mathbf{pa}(t) \mid \mathbf{pa}(t)$$

其中 $\mathbf{nd}(t)$ 表示节点 t 的所有非后代（descendant）节点，$\mathbf{pa}(t)$ 表示所有父节点。这个性质是说，给定节点父节点的条件下，该节点和其他所有除了父节点和后代节点的节点都是条件独立的。

其实很好理解，假设我们从该节点往父节点上面走，发现一定不是 v-structure，但是由于给定了所有父节点为 evidence，根据第 1,2 条规则，肯定是被 d-separated 了，也就是说条件独立成立。再往下走，肯定和孩子节点或者自己构成 v-structure 结构（不然还是孩子节点，可以画出来看看），那么由于孩子节点不是 evidence 的子集，根据第 3 条规则，此条路径必然也被 d-separated 了。所以条件独立。可以举书里 10.11 的例子来说明这个性质，略。

上面的性质有种特殊情况，

$$t \perp \mathbf{pred}(t) \backslash \mathbf{pa}(t) \mid \mathbf{pa}(t)$$

其中 $\mathbf{pred}(t) \subseteq \mathbf{nd}(t)$ 表示节点 t 的先继节点集合。这个性质叫做 **有序马尔可夫性质（ordered Markov property）**。Koller 大神的书里证明了这里讲过的三种性质 G, L, O 都是等价的。

10.5.3 Markov blanket and full conditionals

某节点 t 的 **Markov blanket** 可以表示为 $\mathbf{mb}(t)$，定义如下，

$$\mathbf{mb}(t) \triangleq \mathbf{ch}(t) \cup \mathbf{pa}(t) \cup \mathbf{copa}(t)$$

后面的三个符号分别表示节点 t 的 children, parent 和 co-parent，而 co-parent 表示和 t 共享某个孩子节点的节点集合。

Markov blanket 这个概念很有用，比如我们要计算这个条件概率，

$$p(x_t|\mathbf{x}_{-t}) = \frac{p(x_t, \mathbf{x}_{-t})}{p(\mathbf{x}_{-t})}$$

考虑分子（numerator）分母（denominator）中约去一些项，只剩包含 x_t 的那些 CPD，

$$p(x_t|\mathbf{x}_{-t}) \propto p(x_t|\mathbf{x}_{\mathbf{pa}(t)}) \prod_{s \in \mathbf{ch}(t)} p(x_s|\mathbf{x}_{\mathbf{pa}(t)})$$

后面的连乘项里，就包含了 co-parent 的节点。这个概率 $p(x_t|\mathbf{x}_{-t})$ 叫做 **full conditional**，分布就叫做 full conditional distribution，会在第 24 章的 MCMC 中的吉普斯采样（Gibbs Sampling）会用到。

10.6 Influence (decision) diagrams * 影响（决策）图

我们可以用影响图或者叫决策图来表示多阶段贝叶斯决策问题。这种图是在 PGM 的基础上增加了一些节点，

节点名称	作用	用什么表示
chance nodes	图模型节点，表示随机变量	椭圆，oval
decision nodes, action nodes	要决策的东西	矩形，rectangle
utility nodes, value nodes	Utility	菱形，diamond

例子这里就不提了，到时候如果要讲的话，可以把图贴到 PPT 里。还有前面三个小结的内容同理。课后的习题也要看一下。