MLAPP 读书笔记 - 10 离散图模型(Directed graphical models)(贝叶斯网络(Bayes nets))

> A Chinese Notes of MLAPP,MLAPP 中文笔记项目
> https://zhuanlan.zhihu.com/python-kivy

记笔记的人：cycleuser

2018年7月1日18:59:12

10.1 概论

> 以简单方式对待复杂系统的原则,我基本知道两个:首先就是模块化原则,其次就是抽象原则.我是机器学习中计算概率的辩护者,因为我相信概率论对这两种原则都有深刻又有趣的实现方式,分别通过可分解性和平均.在我看来,尽可能充分利用这两种机制,是机器学习的前进方向. Michael Jordan, 1997 (转引自 (Frey 1998)).

假如我们观测多组相关变量,比如文档中的词汇,或者图像中的像素,再或基因片段上的基因.怎么能简洁地表示联合分布$p(x|\theta)$呢?利用这个分布,给定其他变量情况下,怎么能以合理规模的计算时间来推导一系列的变量呢?怎么通过适当规模的数据来学习得到这个分布的参数呢?这些问题就是概率建模(probabilistic modeling),推导(inference)和学习(learning)的核心,也是本章主题了.

10.1.1 链式规则(Chain rule)

通过概率论的链式规则,就可以讲一个联合分布写成下面的形式,使用任意次序的变量都可以:

$$p(x_{1:V}) = p(x_1)p(x_2|x_1)p(x_3|x_2,x_1)p(x_4|x_1,x_2,x_3)...p(x_V|x_{1:V-1}) \quad (10.1)$$

其中的V是变量数目,MATLAB风格的记号$1:V$表示集合{1,2,...,V},为了简洁,上面狮子中去掉了对固定参数θ的条件.这个表达式的问题在于随着T变大,要表示条件分布$p(x_t|x_{1:t-1})$就越来越复杂了.

例如,设所有变量都有K个状态(states).然后可以将$p(x_1)$表示为$O(K)$个数字的表格,表示了一个离散分布(实际上只有K-1个自由参数,因为总和为一的约束条件,不过这里为了写起来简单就写成O(K)了).然后还可以将$p(x_2|x_1)$写成一个$O(K^2)$个数值的表格,写出$p(x_2=j|x_1=i)=T_{ij}$;T叫做随机矩阵(stochastic matrix)对$0 \leq T+ij \leq 1$的所有条目以及所有列i满足约束条件

$\sum_j T_{ij} = 1$.与此类似,还可以将$p(x_3|x_1, x_2)$表示成一个有$O(K^3)$个数值的三维表格.这些表格就叫做条件概率表格(conditional probability tables,缩写为CPT).然后在我们这个模型里面就有$O(K^V)$个参数.要学习得到这么多参数需要的数据量就要多得可怕了.

要解决这个问题可以将每个条件概率表(CPT)替换成条件概率分布(conditional probability distribution,缩写为CPD),比如多想逻辑回归,也就是$p(x_t = k|x_{1:t-1}) = S(W_t x_{1:t-1})_k$.这样全部参数就只有$O(K^2V^2)$个了,就得到了一个紧凑密度模型(compact density model)了(Neal 1992; Frey 1998).如果我们要评估一个全面观测向量$x_{1:t-1}$的概率,这已经足够了.比如可以使用这个模型定义一个类条件密度(class-conditional density)$p(x|y) = c$,然后建立一个生成分类器(generative classifier,Bengio and Bengio 2000).不过这个模型不能用于其他预测任务,因为这个模型里面每个变量都依赖之前观测的全部变量.所以其他问题要另寻办法.

10.1.2 条件独立性(Conditional independence)

高效率表征一个大规模联合分布的关键就是对条件独立性(Conditional independence,缩写为CI)进行假设.回忆本书2.2.4当中,在给定Z的情况下,X和Y的条件独立记作$X \perp Y|Z$,当且仅当条件联合分布可以写成条件边缘分布乘积的时候才成立,如下所示:

$$X \perp Y|Z \iff p(X, Y|Z) = p(X|Z)p(Y|Z) \tag{10.2}$$

这有啥用呢?假如有$x_{t+1} \perp x_{1:t-1}|x_t$,也就是说在给定当前值的条件下,未来值和过去值独立.这就叫(一阶(first order))马尔科夫假设(Markov assumption).利用这个假设,加上链式规则,就可以写出联合分布形式如下所示:

$$p(x_{1:V}) = p(x_1)\prod_{t=1}^{V} p(x_t|x_{t-1}) \tag{10.3}$$

这就叫做一个(一阶(first order))马尔科夫链(Markov chain).可以通过在状态上的初始分布(initial distribution)$p(x_1 = i)$来表示,另外加上一个状态转换矩阵(state transition matrix) $p(x_t = j|x_{t-1} = i)$.更多细节参考本书17.2.

10.1.3 图模型

虽然一阶马尔科夫假设对于定义一维序列分布很有用,但对于二维图像.或者三维视频,或者更通用的任意维度的变量集(比如生物通路上的基因归属等等),要怎么定义呢?这时候就需要图模型了.

图模型(Graphical models,缩写为GM)是通过设置条件独立性假设(CI assumption)来表示一个联合分布(joint distribution).具体来说就是图上的节点表示随机变量,而(缺乏的)边缘表示条件独立性假设(CI assumption)(对这类模型的更好命名应该是独立性图(independence diagrams),不过图模型这个叫法已经根深蒂固了.)有几种不同类型的图模型,取决于图是有向(directed)/无向(undirected)/或者两者结合.在本章只说有向图(directed graphs).到第19章再说无向图(undirected graphs).

此处参考原书图10.1

10.1.4 图模型术语(Graph terminology)

在继续讨论之前,先要定义一些基本术语概念,大部分都很好理解.

一个图(graph)$G = (V, \mathrm{E})$包括了一系列节点(node)或者顶点(vertices),$V = \{1, \ldots, V\}$,还有一系列的边(deges)$\mathrm{E} = \{(s, t) : s, t \in V\}$.可以使用邻近矩阵(adjacency matrix)来表示这个图,其中用$G(s, t)$来表示$(s, t) \in \mathrm{E}$,也就是$s \to t$是凸中的一个遍.如果当且仅当$G(t, s) = 1$的时候$G(s, t) = 1$,就说这个图是无向的(undirected),否则就是有向的(directed).一般假设$G(s, s) = 0$,意思是没有自我闭环(self loops).

下面是其他一些要常用到的术语:

- 父节点(Parent)对一个有向图,一个节点的父节点就是所有节点所在的集合:$pa(s) \triangleq t : G(t, s) = 1$.
- 子节点(Child)对一个有向图,一个节点的子节点就是从这个节点辐射出去的所有节点的集合:$ch(s) \triangleq t : G(s, t) = 1$.
- 族(Family)对一个有向图,一个节点的族是该节点以及所有其父节点:$fam(s) = \{s\} \cup pa(s)$.
- 根(root)对一个有向图,根是无父节点的节点.
- 叶(leaf)对一个有向图,叶就是无子节点的节点.
- 祖先(Ancestors)对一个有向图,祖先包括一个节点的父节点/祖父节点等等.也就是说t的祖先是所有通过父子关系向下连接到t的节点:$anc(t) \triangleq \{s : s \rightsquigarrow t\}$.
- 后代(Descendants)对一个有向图,后代包括一个节点的子节点/次级子节点等等.也就是s的后代就是可以通过父子关系上溯到s的所有节点集合:$desc(t) \triangleq \{t : s \rightsquigarrow t\}$.
- 邻节点(Neighbors),对于任意图来说,所有直接连接的节点都叫做邻节点:$nbr(s) \triangleq \{t : G(s, t) = 1 \vee G(t, s) = 1\}$.对于无向图(undirected graph)来说,可以用$s \sim t$来表示s和t是邻节点(这样$(s, t) \in \mathrm{E}$就是图的边(edge)了).
- 度数(Degree)一个节点的度数是指该节点的邻节点个数.对有向图,又分为入度数(in-degree)和出度数(out-degree),指代的分别是某个节点的父节点和子节点的个数.
- 闭环(cycle/loop)顾名思义,只要沿着一系列节点能回到初始的位置,顺序为$s_1 \to s_2 \ldots \to s_n \to s_1, n \geq 2$,就称之为一个闭环.如果图是有向的,闭环也是有向的.图10.1(a)中没有有向的闭环,倒是有个无向闭环$1 \to 2 \to 4 \to 3 \to 1$.
- 有向无环图(directed acyclic graph,缩写为DAG)顾名思义,就是有向但没有有向闭环的,比如图10.1(a)就是一例.
- 拓扑排序(Topological ordering)对一个有向无环图(DAG),拓扑排序(topological ordering)或者也叫全排序(total ordering)是所有父节点比子节点数目少的节点的计数.比如在图10.1(a)中,就可以使用(1, 2, 3, 4, 5)或者(1, 3, 2, 5, 4)等.

- 路径(Path/trail)对于$s \rightsquigarrow t$来说路径就是一系列从s到t的有向边(directed edges).
- 树(Tree)无向树(undirected tree)就是没有闭环的无向图.有向树(directed tree)是没有有向闭环的有向无环图(DAG).如果一个节点可以有多个父节点,就称为超树(polytree),如果不能有多个父节点,就成为规范有向树(moral directed tree).
- 森林(Forest)就是树的集合.
- 子图(Subgraph)(包括节点的)子图G_A是使用A中的节点和对应的边(edges)创建的$G_A = (\mathrm{V}_A, \mathrm{E}_A)$.
- 团(clique)对一个无向图,团是一系列互为邻节点的节点的集合.在不损失团性质的情况下能达到的最大规模的团就叫做最大团(maximal clique).比如图10.1(b) 当中的{1,2}就是一个团,但不是最大团,因为把3加进去依然保持了团性质(clique property).图10.1(b)中的最大团:{1, 2, 3}, {2, 3, 4}, {3, 5}.

10.1.5 有向图模型

有向图模型(directed graphical model,缩写为DGM)是指整个图都是有向无环图(directed acyclic graph,缩写为DAG)的图模型.更广为人知的名字叫贝叶斯网络(Bayesian networks).不过实际上这个名字并不是说这个模型和贝叶斯方法有啥本质上的联系:只是定义概率分布的一种方法而已.这些模型也叫作信念网络(belief networks).这里的信念这个词(belief)指的是主观的概率.关于表征有向图模型(DGM)的概率分布的种类并没有什么本质上的主管判断.这些模型有时候也叫作因果网络(causal networks),因为有时候可以将有向箭头解释成因果关系.不过有向图模型本质上并没有因果关系(关于因果关系有向图模型的讨论参考本书26.6.1.)名字这么多这么乱,咱们就选择最中性的称呼,就叫它有向图模型(DGM).

有向无环图(DAG)的关键性之就是节点可以按照父节点在子节点之前来排序.这也叫做拓扑排序(topological ordering),在任何有向无环图中都可以建立这种排序.给定一个这样的排序,就定义了一个有序马尔科夫性质(ordered Markov property),也就是假设一个节点只取决于其直接父节点,而不受更早先辈节点的影响.也就是:

$x_s \perp x_{pred(s)/pa(s)} | x_{pa(s)}$ (10.4)

(注:上面公式中应该是反斜杠 "",但是我不知在LaTex里面怎么打出来.)

上式中的$pa(s)$表示的是节点s的父节点,而$pred(s)$表示在排序中s节点的先辈节点.这是对一阶马尔科夫性质(first-order Markov property)的自然扩展,目的是构成一个链条来泛华有向无环图(DAG).

此处参考原书图10.2

例如,在图10.1(a)中编码了下面的联合分布:

$$p(x_{1:5}) = p(x_1)p(x_2|x_1)p(x_3|x_1, \cancel{x_2})p(x_4|\cancel{x_1}, x_2, x_3)p(x_5|\cancel{x_1}, \cancel{x_2}, x_3, \cancel{x_4}) \quad (10$$
$$= p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2, x_3)p(x_5|x_3) \quad (10$$

其中每个$p(x_t|x_{pa(t)})$都是一个条件概率分布(CPD).将分布写成$p(x|G)$是要强调只有当有向无环图(DAG)G当中编码的条件独立性假设(CI assumption)是正确的时候登时才能成立.不过通常为了简单起见都会把这个条件忽略掉.每个节点都有$O(F)$个父节点以及K个状态,这个模型中参数总数就是$O(VK^F)$,这就比不做独立性假设的模型所需要的$O(K^V)$个要少多了.

10.2 样例

这一节展示一些可以用有向图模型(DGM)来表示的常见概率模型.

10.2.1 朴素贝叶斯分类器(Naive Bayes classifiers)

朴素贝叶斯分类器是在本书3.5就讲到了.这个模型中假设了给定类标签(calss label)条件下特征条件独立.此假设如图10.2(a)所示.这样就可以将联合分布写成下面的形式:

$p(y, x) = p(y)\prod_{j=1}^{D} p(x-j|y)$(10.8)

朴素贝叶斯加设很朴素,因为假设了所有特征都是条件独立的.使用图模型就是捕获变量之间相关性的一种方法.如果模型是属性的,就称此方法为树增强朴素贝叶斯分类器(tree-augmented naive Bayes classifiers,缩写为TAN,Friedman et al. 1997).如图10.2(b)所示.使用树形结构而不是一般的图模型的原因有两方面.首先是树形结构使用朱刘算法(Chow-Liu algorithm,1965年有朱永津和刘振宏提出)来优化,具体如本书26.3所示.另外就是树结构模型的缺失特征好处理,具体会在本书20.2有解释.

此处参考原书图10.3

此处参考原书图10.4

10.2.2 马尔科夫和隐形马尔科夫模型

图10.3(a)所示为将一阶马尔科夫连表示成了一个有向无环图(DAG).其中的假设是最邻近的上一个特征x_{t-1}包含了之前所有历史特征$x_{1:t-2}$的所有需要我们知道的内容,这个假设有点太强了.为了减弱一点这个假设,可以再添加一下从x_{t-2}到x_t的依赖;这样就成了一个二阶马尔科夫链(second order Markov chain),如图10.3(b)所示.对应的联合分布也就是:

$p(x_{1:T}) = p(x_1, x_2)p(x_3|x_1, x_2)p(x_4|x_2, x_3)\ldots = p(x_1, x_2)\prod_{t=3}^{T} p(x_t|x_{t-1}, x_{t-2})$
(10.9)

用类似方法还可以建立更高阶的马尔科夫链.关于马尔科夫模型的更多细节参考本书17.2.

然而很不幸,就算使用二阶马尔科夫假设,面对观测有长程相关性的情况也可能会不适用.也不能持续无限提高阶数,因为这样参数规模就太大了.另外一种方法就是假设存在潜在的隐藏过程,而这个过程可以用一个一阶马尔科夫链来建模,但数据是这个过程中的有噪音观测.这样就得到了隐形马尔科夫模型(hidden Markov model,缩写为HMM),如图10.4所示.其中的z_t是在"次数(time)"t上的隐含变量(hidden variable),而x_t是观测变量.(这个模型可以用于任意的有序数据,比如基因序列或者语言文字等等,所以t指代的可以使位置而不一定只是观测时间.)条件概率分布(CPD)$p(z_t|z_{t-1})$是转换模型(transition model),而条件概率分布(CPD)$p(x_t|z_t)$是观测模型(observation model).

此处参考原书表10.1

这些隐含变量往往是有用变量,比如某人讲话过程中某个词汇的作用.观测变量就是我们观测到的值,比如声音波形.我们需要做的就是在给定数据的情况下估计隐含状态,也就是计算$p(z_t|x_{1:t},\theta)$.这就叫做状态估计(state estimation).也是一种概率推导的形式.关于隐形马尔科夫模型的更多细节会在本书17章讲解.

10.2.3 医学诊断

接下来考虑这样一个场景,医院重症监护室(intensive care unit,缩写为ICU)中衡量各种人体参数变量,比如心率/呼吸频率/血压等等,对这些变量之间的关系进行建模.图10.5(a)所示的报警网络(alarm network)就是表征这些变量的相关性或者无关性的一种方法(Beinlich et al. 1989).这个模型有37个变量,504个参数.

这个模型是人工构建的,通过了一个叫做知识工程(knowledge engineering)的过程,这个系统也就叫做概率专家系统(probabilistic expert system).在本书10.4会讲到在假设图结构未知的情况下,如何从数据中学习得到有向图模型的参数.然后在本书26章会讲到如何学习图结构本身.

还有另外一种医疗诊断用的网络模型,叫做快速医疗指南(quick medical reference,缩写为QMR)网络(Shwe et al. 1991),如图10.5(b)所示.这个模型通常用于对传染病的建模.这种QMR模型往往是二分图结构(bipartite graph structure),疾病种类或者致病因素在顶层,而症状或者发病表现在底层.所有的节点都是二值化的(binary).可以将这个分布写成虾米的形式:

$$p(v,h)=\prod_s p(h_s)\prod_t p(v_t|h_{pa(t)}) \tag{10.10}$$

上式中的h_s就是隐藏节点(hidden nodes)(疾病),而v_t表示的是可见节点(visible nodes)(症状).

根节点(root nodes)的条件概率分布(CPD)就是伯努利分布,表示了对应疾病的先验概率.对叶节点(leaves)(症状)使用条件概率表格(CPT)来表示条件概率分布(CPD)会需要太多参数,因为很多叶节点(leaf node)都有特别多的父节点数.一个很自然的替代方法就是使用逻辑回归来对条件概率分布建模,$p(v_t=1|h_{pa(t)})=sigm(w_t^T h_{pa(t)})$.(条件概率分布为逻辑回归分布的有向图模型也叫作S形信念网络(sigmoid belief net,Neal 1992).)上面这个模型的参数是认为创建的,还有另外一个可用条件概率分布(CPD)叫做"噪音或模型"(noisy-OR model,这里的**或**是指**或门,or gate**).

此处参考原书图10.5
此处参考原书表10.2

噪音或模型(noisy-OR model)假设如果父节点为开(on),子节点通常也为开(on)(因为这是一个或门or gate),但有时候从父节点到子节点的连接也可能会失败,这种情况随机独立发生.这种情况下畸变父节点为开(on),子节点也可能为关(off).要对这类情况进行更确切地建模,可以设$s \rightarrow t$连接失败的概率为$\theta_{st} = 1 - q_{st}$,所以$q_{st} = 1 - \theta_{st} = p(v_t = 1|h_s = 1, h_{-s} = 0)$为s可单独激活t的概率(也就是因果律,causal power").而子节点为关(off)则只会是所有从父节点的链接都随机独立失败.也就是:

$$p(v_t = 0|h) = \prod_{s \in pa(t)} \theta_{st}^{I(h_s=1)} \text{(10.11)}$$

很明显$p(v_t = 1|h) = 1 - p(v_t = 0|h)$.

如果我们观测到$v_t = 1$,而素有父节点都为关(off)那么这就违背(contradicts)了这个模型.这样的一个数据案例会在这个模型下得到零概率.这就有问题了,因为很可能某个人并没有患上任何特定疾病,但也可能表现出某个症状.要处理这种情况,就要添加一个傻傻的遗漏节点(dummy leak node)h_0,这个节点总处于开状态(on);这就代表着所有其他原因.参数q_{0t}表示的是背景泄露本身导致这个效果的概率.修改过的条件概率分布(CPD)就成了$p(v_t = 0|h) = \theta_{0t} \prod_{s \in pa(t)} \theta_{st}^{h_s}$.数值样本参考表10.1.

如果定义$w_{st} \triangleq \log(\theta_{st})$,就可以将条件概率分布(CPD)重写成:

$$p(v_t = 1|h) = 1 - \exp(w_{ot} + \sum_s h_s w_{st}) \text{(10.12)}$$

可见这和逻辑回归模型就很相似了.

带有噪声或门的二分模型(Bipartite models with noisy-OR)的条件概率分布(CPD)也简写作BN2O模型.通常借助相关领域的专业经验来人为设置θ_{st}参数都比较简单.不过也可以从数据中学习来进行设置参数(Neal 1992; Meek and Heckerman 1997).有噪音或条件概率分布(Noisy-OR CPDs)在对人类的因果学习建模的时候也很有用(Griffiths and Tenenbaum 2005),另外也适用于通用二值化分类背景(Yuille and Zheng 2009).

10.2.4 基因链接分析(Genetic linkage analysis)*

对有向图模型的应用还有另外一个很重要也很悠久的领域,就是基因链接分析(genetic linkage analysis).先从谱系系树(pedigree graph)开始,这是一种表示了父子节点关系的有向无环图模型(DAG)如图10.6(a)所示.然后将其转换成有向图模型(DGM),接下来就会讲解这个过程.最后在得到的模型中进行概率推导.

此处参考原书图10.6

更详细说,对于每个人或者动物和沿着基因上的位置j都建立三个节点:观测标记(observed marker) X_{ij}(可以使血型啊或者一个能衡量的DNA片段),以及另外的两个隐藏的等位基因(hidden alleles) G_{ij}^m和G_{ij}^p,一个是来自i的母本(母本等位基因,maternal allele),另一个来自i的父本(父本等位基因,paternal allele)结合在一起就形成了有序基因对$G_{ij} = (G_{ij}^m, G_{ij}^p)$,这就是i在位置j上的隐藏基因型(hidden genotype).

很明显必须添加$(G_{ij}^m \rightarrow X_{ij}$和$G_{ij}^p \rightarrow X_{ij}$来表示基因对性状(性状就是基因型的表现效应)的控制.条件概率分布(CPD)$p(X_{ij}|G_{ij}^m, G_{ij}^p)$就叫做外显率模型(penetrance model).比如$X_{ij} \in \{A, B, O, AB\}$表示了第i个人的观测学习,而$G_{ij}^m, G_{ij}^p \in \{A, B, O\}$表示的是他们的基因型(genotype).就可以使用表10.2所示的确定性条件概率分布(deterministic CPD)来表示其外显率模型.比如基因A的优先级高于O,所以一个人基因型如果为AO或者OA,他的血型就是A型.

另外还要在G_{ij}上加上i的父母,这反映了一个人从父母得到遗传物质的孟德尔遗传定律(Mendelian inheritance).具体来说就是设$m_i = k$为i的母本.然后G_{ij}^m就可以等于G_{kj}^m或者G_{kj}^p,也就是i的母本等位基因是其母本两个基因当中的一个的复制品.设Z_{ij}^m为隐藏变量确定了具体所选择的复制对象.然后可以用下面的条件概率分布对其进行建模,这个模型就是遗传模型(inheritance model):

$$p(G_{ij}^m|G_{kj}^m, G_{kj}^p, Z_{ij}^m) = \begin{cases} I(G_{ij}^m = G_{kj}^m) \text{ if } Z_{ij}^m = m \\ I(G_{ij}^m = G_{kj}^p) \text{ if } Z_{ij}^m = p \end{cases} \quad (10.13)$$

类似的方法还可以定义$p(G_{ij}^p|G_{kj}^m, G_{kj}^p, Z_{ij}^p)$,其中的$k = p_i$表示的是i的父本.$Z_{ij}$的值可以用来确定基因类型的相(phase).而$G_{ij}^p, G_{ij}^m, Z_{ij}^p, Z_{ij}^m$构成了第i个人在第j个位置上的单倍型(haplotype).

然后要指定根节点的先验$p(G_{ij}^m$和$p(G_{ij}^p)$.这就叫祖先模型(founder model),表示了在人口中不同基因型的总体比例.通常假设这些祖先基因在不同位置上相互独立.

最后要对控制遗传过程的转换变量指定先验.这些变量在空间上相关(spatially correlated),因为基因组上邻近的位置通常是一同遗传的,重新结合的情况很罕见.可以在Z上引入一个二阶马尔科夫链来对此进行建模,其中在位置j上的转换状态的概率由$\theta_j = \frac{1}{2}(1 - e^{-2d_j})$给出,其中的$d_j$是位置j和j+1之间的距离.这个模型就叫做重新组合模型(recombination model).

这样得到的涂抹些如图10.6(b)所示:一个同血缘的有向无环图(replicated pedigree DAG),通过转换Z变量增强(augmented),使用了马尔科夫链.(有个相关的模型叫做进化阴性马尔科夫模型(phylogenetic HMM,Siepel and Haussler 2003),模拟的是进化过程中的演化.)

为了简单,这里举例只看一个基因位置,也就是对应血型的片段.为了简单就去掉j索引了.假如观测到了$x_i = A$.那么就有三种可能的基因组G_i:(A, A), (A, O) ,(O, A).
这里会有多解性,因为基因组到基因表型的映射是多到一的.要将这个映射关系逆转,就要面对逆转问题(inverse problem).还好可以使用亲戚的学醒来降低甚至消除这些多解性.信息就会通过学院有

向无环树(pedigree DAG)从其他的x_i留到他们的G_i中,然后传递到第i个人的G_i上.因此可以结合局部证据(local evidence)$p(x_i|G_i)$和先验$p(G_i|X_{-i})$,以其他数据为条件,然后得到一个更低熵(less entropic)的局部后验$p(G_i|x) \propto p(x_i|G_i)p(G_i|x_{-i})$.
在实际应用中,这个模型一般用来根据给定疾病的致病基因来检测是否有疾病,也就是基因关联检测任务(genetic linkage analysis task).这种方法的工作流程如下.首先假设模型中所有参数,包括标记位置之间的距离,都是已知的.而唯一未知的是致病基因的位置.如果有L个标记位置,就构建L+1个模型:在模型l中,假定致病基因在标记l后出现,$0 < l < L + 1$.然后可以通过切换参数$\hat{\theta}_l$来估计马尔科夫模型,然后是致病基因和其最邻近位置之间的距离d_l.通过似然函数$p(D|\hat{\theta}_l)$来衡量模型质量.然后选出来最高似然率的模型(在均匀先验(uniform prior)下等价于最大后验分布模型(MAP model).

不过要记住,计算似然率的时候需要边缘化掉所有的隐藏的Z和G这些变量.关于这个模型的具体信息细节参考(Fishelson and Geiger 2002);这些是基于变量估计算法的模型,我们会在本书20.3讲到.不过很不幸,由于一些会在本书20.5讲到的原因,实际上如果个体规模或者基因位置太多的话,具体的模型在计算上可能会很困难.关于计算似然率的近似方法参考(Albers et al. 2006);这种方法是变分推导(variational inference)的一种形式,具体会在本书22.4.1讲到.

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

考虑一个有向图模型(DGM),其中全部变量都是实数值的,而所有条件概率分布(CPD)都有如下的形式:

$$p(x_t|x_{pa(t)}) = N(x_t|\mu_t + w_t^T x_{pa(t)}, \sigma_t^2) \quad (10.14)$$

这样的条件概率分布就叫做线性高斯条件概率分布(linear Gaussian CPD).将所有这些条件概率分布(CPD)乘到一起,就得到了一个大规模的连和高斯分布,形式为$p(x) = N(x|\mu, \Sigma)$.这就叫一个有向高斯图模型(缩写为 directed GGM),或者也叫做高斯贝叶斯网络(Gaussian Bayes net).

接下来解释一下如何从条件概率参数中推出μ和Σ,这部分的内容参考了(Shachter and Kenley 1989, App. B).为了简单,将条件概率分布写成下面的形式:
$$x_t = \mu_t + \sum_{s \in pa(t)} w_{ts}(x_s - \mu_s) + \sigma_t z_t \quad (10.15)$$

其中$z_t \sim N(0, 1)$,σ_t为给定亲本下x_t的条件标准偏差(conditional standard deviation),w_s是$s \to t$的强度,而μ_t是局部均值(local mean).

很明显局部均值的连接(concatenation)就是全局均值了,即$\mu = (\mu_1, \ldots, \mu_D)$.然后再推导全局协方差(global covariance)Σ.设$s \triangleq diag(\sigma)$是一个对角矩阵,包含了标准偏差.然后可以将等式10.15写成矩阵向量乘积的形式如下所示:
$$(x - \mu) = W(x - \mu) + Sz \quad (10.16)$$

然后设e为噪音项目的向量:

$e \triangleq Sz$(10.17)

重新整理就得到了:

$e = (I - W)(x - \mu)$(10.18)

因为W是下三角矩阵(因为如果在拓扑排序(topological ordering)中$t > s$,则$w_{ts} = 0$),所以则有$I - W$是一个对角线为1的下三角矩阵.因此:

$$\begin{pmatrix} e_1 \\ e_2 \\ . \\ . \\ . \\ e_d \end{pmatrix} = \begin{pmatrix} 1 & & & & \\ -w_{21} & 1 & & & \\ -w_{32} & -w_{31} & 1 & & \\ \cdots & & & & \\ -w_{d1} & -w_{d2} & \cdots & -w_{d,.d-1} & 1 \end{pmatrix} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \\ . \\ . \\ . \\ x_d - \mu_d \end{pmatrix} \quad (10.19)$$

由于 I-W 总是可逆的(invertible),就可以写出:

$$x - \mu = (I - W)^{-1} e \triangleq Ue = USz \quad (10.20)$$

其中定义$U = (I - W)^{-1}$.另外回归权重(regression weights)对应着对Σ的柯列斯基分解(Cholesky decomposition),如下所示:

$$\begin{aligned} \Sigma = cov[x] = cov[x - \mu] & \quad (10.21) \\ = cov[USz] = UScov[z]SU^T = US^2U^T & \quad (10.22) \end{aligned}$$

10.3 推导(Inference)

上文表明图模型可以很简便地定义联合概率分布.那么给定了一个联合分布的时候,能怎么办呢?对这种联合分布的主要用途就是进行概率推导(probabilistic inference).这指的是从已知量来估计未知量的过程.比如在10.2.2中,降到了隐性马尔科夫模型,这个模型的一个目的就是要从观测(比如语音信号)中来估计隐藏状态(比如词汇).在10.2.4,降到了基因链接分析,目标之一就估计不同有向无环图下数据的似然函数,对应的是关于致病基因位置的不同假设.

一般来说,可以按照下面的步骤提出推到问题.加入有一系列相关的随机变量,联合分布为$p(x_{1:V}|\theta)$.(在本节,要假设参数θ是已知的.然后在10.4要讲如何学习参数.)然后讲这个向量分解成观测变量(visible variables)x_v,以及未观测到的隐藏变量(hidden variables)x_h.给定已知量之后计算未知量的后验分布的推导为:

$p(x_h|x_v, \theta) = \frac{p(x_h, x_v|\theta)}{p(x_v|\theta)} = \frac{p(x_h, x_v|\theta)}{\Sigma_{x'_h} p(x'_h, x_v|\theta)}$(10.23)

基本上可以将观测变量限定在观测值上来以数据为条件,然后归一化(normalizing),就从$p(x_h, x_v)$

得到了$p(x_h|x_v)$.归一化常数$p(x_v|\theta)$就是数据的似然函数,也叫做证据概率(probability of the evidence).

有时候隐藏变量中只有一部分是我们感兴趣的.所以可以对隐藏变量在进行区分,我们感兴趣的那部分称为查询变量(query variables)x_q,剩下那些不感兴趣的就叫扰嚷变量(nuisance variables)x_n.可以将扰嚷变量(nuisance variables)边缘化掉(marginalizing out)而只留下我们想知道的部分:
$p(x_q|x_v,\theta)=\sum_{x_n}p(x_q,x_n|x_v,\theta)$(10.24)

在本书4.3.1,一键看到了如果在$O(V^3)$时间内对一个多变量高斯分布如何进行所有运算,其中的V是变量数目.那么面对离散随机变量,比如K个状态的时候,会怎么样呢?如果联合分布可以表示成一个多维表格的形式,就总可以确切进行这些计算,但需要的时间为$O(K^V)$.在本书20章会解释如何利用图模型来分区解码在$O(VK^{w+1})$时间内进行计算,其中w是图的树宽(treewidth).这衡量了图像树的程度.如果图是一个树(或者一个链),则$w=1$,对于这些模型,推导需要的时间是节点数目的线性函数.不幸的是,对于更通用的图来说,确切的推导可能需要节点数目的指数级时间,具体原因会在本书20.5解释.在本书后面也会检验各种不同的估计推导策略.

10.4 学习

在图模型领域,通常都要区分推导(inference)和学习(learning).推到意味着计算$p(x_h|x_v,\theta$(或者其函数),其中的v是可见节点,h是隐藏节点,θ是模型参数,假设为已知.而学习过程通常意味着给定数据下计算参数的最大后验估计(MAP estimate):

$\hat{\theta}=\arg\max_\theta\sum_{i=1}^N\log p(x_{i,v}|\theta+\log p(\theta)$(10.25)

上式中的$x_{i,v}$表示的是情况i下的可见变量.如果使用一个均匀先验$p(\theta)\propto 1$,这就降低到最大似然估计(MLE).

如果选择贝叶斯角度来看,这些参数都是位置变量,也应该退到.所以对于一个贝叶斯主义者,在推到和学习之间并没有什么区别.实际上可以就添加一些参数作为图节点,以D为条件,然后推测所有节点的值(后面还会详细讲这个过程.)

这样来看,隐藏变量和参数的区别就是隐藏变量的规模随着训练数据的增多而增多(因为通常每个观测情景都有一系列的隐藏变量),而参数的规模通常都是固定的(至少在参数化模型是这样的).这就意味着我们必须要积分掉隐藏变量来避免过拟合,但也可以对参数使用点估计,毕竟参数数量少得多.

此处参考原书图10.7

10.4.1 板表示法(Plate notation)

在从数据推导参数的时候,通常假设数据位独立同分布(iid).可以使用一个图模型来表示这个假设,如图10.7(a)所示.这表示了每个数据情况都是独立生成的,但来自于同一个分部.要注意数据情况这时候只在参数θ为条件上独立;数据案例是在边界上独立的.不过也会发现在这个案例中,数据案例出现的顺序并不影响我们队参数θ的新年,因为所有排序都有同样的充分统计量.因此说这种数据是可交换的(exchangeable).

为了避免视觉上的混乱,通常使用一种语法糖叫做板(plates):将重复出现的变量画上一个小方框,当模型展开时框内节点就重复了.通常在方框的右下角写上复制或者重复次数.比如图10.7(b)为例.对应的联合分布的形式为:

$$p(\theta, D) = p(\theta)[\prod_{i=1}^{N} p(x_i|\theta)] \text{(10.26)}$$

这个有向图模型(DGM)表示了第五章我们所考虑模型背后的条件独立(CI)假设.

另外一个复杂点的例子如图10.8所示.其中图左边的是一个朴素贝叶斯分类器,展开有D个特征,但是用了一个板来表示在条件$i = 1:N$上的重复.右边的版本是对同一个模型使用网状板表示法(nested plate notation).当一个变量在两个班内的时候,就有两个分项指数(sub-indices).例如θ_{jc}表示了在类条件密度c中特征j的参数.注意这里的板可以使网状的或者交错的.

可以对更复杂参数绑定模型进行建模的记号设备(notational devices)也可以设计出来,比如(Heckerman et al. 2004),但用得不太广泛.图中不清楚的是当且仅当$y_i = c$的时候θ_{jc}用来生成x_{ij},否则就忽略掉.这就是背景特定独立性(context specific independence)的一个例子,因为条件独立关系$x_{ij} \perp \theta_{jc}$仅当$y_i \neq c$的时候才成立.

此处参考原书图10.8

10.4.2 从全部数据中学习

如果每个情况中的变量都全面观测到了,这样就没有错过的数据了,也就没有隐藏变量了,这也就说这个数据是完备的(complete).对于完备数据集的有向图模型(DGM),似然函数为:

$$p(D|\theta) = \prod_{i=1}^{N} p(x_i|\theta) = \prod_{i=1}^{N} \prod_{t=1}^{V} p(x_{it}|x_{i,pa(t)}, \theta_t) = \prod_{t=1}^{V} p(D_t|\theta_t) \text{(10.27)}$$

其中的D_t是与节点t和其母本相关的数据,也就是第t个族.这是一个项的乘积,每条件概率分布(CPD)一个.就说似然函数(likelihood)根据图结构来分解(decompose).

然后假设先验也能同样分解:

$$p(\theta) = \prod_{t=1}^{V} p(\theta_t) \text{(10.28)}$$

然后很明显后验也是这么分:

$$p(\theta|D) \propto p(D\|theta)p(\theta) = \prod_{t=1}^{V} p(D_t|\theta_t)p(\theta_t) \text{(10.29)}$$

这就意味着可以分别计算每个条件概率分布(CPD)的后验.用文字来说就是:

分区先验加上分区似然函数就得到了分区的后验(10.30)

然后再举个例子,所有的条件概率分布都是表格化的(tabular),这样就可以扩展到本书3.5.1.2中的早期结论,其中讲的是贝叶斯朴素贝叶斯方法.对每个条件情况(conditioning case)比如每个父本值的组合如表格10.2所示,都有一个单独的列(row,也就是一个单独的多元伯努利分布).规范来说,就可以将第t个条件概率表个(CPT)写作$x_t|x_{pa(t)}=c \sim Cat(\theta_{tc})$,其中对$k=1:K_t, c=1:C_t, t=1:T$,有$\theta_{tck} \triangleq p(x_t=k|x_{pa(t)}=c)$.这里的$K_t$是节点t的状态书,$C_t \triangleq \prod_{s \in pa(t)} K_s$为父本组合数目,而T是节点数目.很明显对每个条件概率表(CPT)的每一列都有$\sum_k \theta_{tck}=1$.

然后对每个条件概率表(CPT)的没格列都使用一个单独的狄利克雷先验(a separate Dirichlet prior),也就是$\theta_{tc} \sim Dir(\alpha_{tc})$.然后以通过简单加起来伪计数到经验技术上来计算先验,就得到了$\theta_{tc}|D \sim Dir(N_{tc}+\alpha_{tc})$,其中$N_{tck}$是节点t在其父本为状态c的时候处于状态k的次数:
$N_{tck} \triangleq \sum_{i=1}^N I(x_{i,t}=k, x_{i,pa(t)}=c)$(10.31)

从等式2.77,可知这个分布的平均值为:
$\bar{\theta}_{tck}=\frac{N_{tck}+\alpha_{tck}}{\sum_{k'}(N_{tck'}+\alpha_{tck'})}$(10.32)

例如图10.1(a)中的有向图模型.假如训练集有下面5个状态:

x_1	x_2	x_3	x_4	x_5
0	0	1	0	0
0	1	1	1	1
1	1	0	1	0
0	1	1	0	0
0	1	1	1	0

接下来列出了所有充分统计量N_{tck},以及在$\alpha_{ick}=1$(对应的是加一光滑)的狄利克雷先验下的后验均值参数$\bar{\theta}_{ick}$,$t=4$个节点:
x_2 x_3	$N_{tck=1}$ $N_{tck=0}$	$\bar{\theta}_{tck=1}$ $\bar{\theta}_{tck=0}$
0 0	0 0	1/2 1/2
1 0	1 0	2/3 1/3
0 1	0 1	1/3 2/3
1 1	2 1	3/5 2/5

很明显最大似然估计(MLE)形式和等式10.32一样,除了没有α_{tck}的项之外,也就是:

$$\hat{\theta}_{tck} = \frac{N_{tck}}{\sum_{k'} N_{tck'}} \quad (10.33)$$

当然了,最大似然估计(MLE)会有在本书3.3.4.1讲过的零计数问题(zero-count problem),所以使用一个先验来正规化这个估计问题很重要.

10.4.3 在有缺失或者隐藏变量的情况下学习

如果有缺失掉的变量或者隐藏变量,似然函数就不能分解了,而且也确实不再是凸函数了,具体细节会在本书11.3讲到.这就意味着我们通常只能计算一个局部最优的最大似然估计或者最大厚颜轨迹.对参数进行贝叶斯推导就更难了.后面的章节会介绍适合这种情况的近似推导技术.

10.5 图模型(DGM)的条件独立性质

图模型的核心其实是一系列的条件独立(CI)假设.如果在一个图G中给定了CA独立于B,就写作$X_A \perp_G X_B | X_C$.设$I(G)$是这个图所编码的全部条件独立陈述(CI statements)的集合.

设G是一个对p的独立映射(independence map,缩写为I-map)或者可以说p是关于G的马尔可夫分布(Markov),当且仅当$I(G) \subseteq I(p)$成立,其中的$I(p)$是所有满足分布p的条件独立陈述(CI statements).也就是说如果这个模型没有做出任何不符合这个分布的条件独立假设,那就是一个独立映射(I-map).这样在推导p的条件独立性质的时候可以使用图来作为对p的安全代理.在设计用于处理大规模分类分布的时候这很有用,就可以无视具体的数值参数θ了.

要注意,完全连接图(fully connected graph)是一个对所有分布的独立映射(I-map)因为根本就没有条件独立声明(CI assertions)(因为全连接图没有缺失边(edges)).如果G是一个对p的独立映射(I-map),而又没有一个G的子集$G' \subseteq G$也是p的独立映射,就说G是p的最小独立映射(minimal I-map).

剩下就是要去判别$X_A \perp_G X_B | X_C$是否成立了.无向图的独立性推到很容易(参考本书19.2),但有向无环图(DAG)就有点复杂了,因为要顾及到有向边(directed edges)的方向.接下来详细说下这些.

10.5.1 有向分离(d-separation)和贝叶斯球算法(Bayes Ball algorithm)（全局马尔可夫性质(global Markov properties)）

首先引入一些定义.当且仅当下面的条件当中至少有一个满足的时候,就说无向路径(undirected path)P是由一系列节点E(包含证据(evidence))所有向分离的(d-separated):
1. P包含一个链,$s \to m \to t$或者$s \leftarrow m \leftarrow t$,其中$m \in E$
2. P包含一个帐篷或者叉子结构,$s \leftarrow m \to t$,其中$m \in E$
3. P包含一个V形结构,$s \to m \leftarrow t$,其中的m以及m的任意后代节点都不属于E

然后,给定了第三方观测集合E,当且仅当从A中每个节点$a \in A$到B中每个节点$b \in B$之间的无向路径(undirected path)被E有向分离,节点集合(set of nodes)A与一个不同的节点集合B是有向分离的.最后对于一个有向无环图的条件独立性质就可以如下定义:
$X_A \perp G X_B | X_C \iff$ 给定E条件下A和B有向分离(10.24)

贝叶斯球算法(Shachter 1998)就是基于上面的定义来判断给定E的情况下A是否和B有向分离的.基本思想是"遮住(shade)"E中的所有节点,表示他们被观测到了.然后在A终端每个节点放置"球",然后让他们通过某种规则"撞来撞去(bounce around)",然后去看是否有球能达到B的节点.三条主要规则如图10.9所示.这里要注意所有的求都可以沿着边的反方向运行.我们会发现一个球可以通过一个链传递,但如果中间遮住了就不行了.类似地,一个球可以通过一个叉子模型传递,但如果中间遮住了也不行.不过,对于V形结构来说,除非中间遮住了,否则不能传递球.

此处参考原书图10.9

下面说下三种贝叶斯球规则.首先考虑链式结构下$X \to Y \to Z$,编码了:
$p(x,y,z) = p(x)p(y|x)p(z|y)$(10.35)

如果以y为条件而x和z独立,就有:

$$p(x,z|y) = \frac{p(x)p(y|x)p(z|y)}{p(y)} = \frac{p(x,y)p(z|y)}{p(y)} = p(x|y)p(z|y)$$(10.36)

也就是$x \perp z|y$.因此在链式结构中观测中间节点就将其一分为二了(就跟马尔科夫链里面一样).

然后考虑叉子结构$X \leftarrow Y \to Z$.联合分布为:

$$p(x,y,z) = p(y)p(x|y)p(z|y)$$(10.37)

此处参考原书图10.10

如果以y为条件而x和z独立,就有:

$$p(x,z|y) = \frac{p(x,y,z)}{p(y)} = \frac{p(y)p(x|y)p(z|y)}{p(y)} = p(x|y)p(z|y)$$(10.38)

也就是$x \perp z|y$.所以观测一个根节点就会分离开这个根节点的子节点(就如朴素贝叶斯分类器一样,参考本书3.5).

最后考虑V形结构,$X \to Y \leftarrow Z$.联合分布为:
$p(x,y,z) = p(x)p(z)p(y|x,z)$(10.39)

如果以y为条件而x和z独立,就有:
$$p(x,z|y) = \frac{p(x)p(z)p(y|x,z)}{p(y)}$$(10.40)

也就是

$$x \not\perp z|y$$

.不过在非条件分布(unconditional distribution)中还是有:
$p(x,z) = p(x)p(z)$(10.41)

所以x和z还是边缘独立的(marginally independent).因此在V形结构底部的共有子节点取条件会使得父节点互相独立.这个重要的效应叫做explaining away, inter-causal reasoning,也叫做伯克森悖论(Berkson's paradox)(指两个通常独立的事物会在特定场合下关联起来,由此产生的相关性容易带来认知上的偏差).举个简单例子,抛两个硬币,设用二值化的0和1表示人头和字,然后观测总值.先验是两枚硬币互相独立,但一旦观测到了总和,这两个硬币就互相耦合(coupled)了,比如如果总和是1,而第一个硬币抛的结果是0,那就知道必然是第二个是1.

最后,贝叶斯求也需要"边界条件(boundary conditions)",如图10.10(a-b)所示.要理解这些规则的来源,可以参考图10.10(c).设Y'是Y的一个无噪音副本.然后如果我们观测Y',就也有效观测了Y,所以父节点X和Z必须竞争来对其进行解释(compete to explain this).如果发送了一个球
$X \to Y \to Y'$,就会沿着$Y' \to Y \to Z$"弹回来".不过如果Y以及所有子节点都是隐藏的,这球就不会弹回来了.

此处参考原书图10.11

例如在图10.11中,很明显$x_2 \perp x_6 | x_5$,因为路径$2 \to 5 \to 6$被观测到的x_5阻塞(blocked)了,而路径$2 \to 4 \to 7 \to 6$被x_7阻塞了(这个是隐藏的),路径$2 \to 1 \to 3 \to 6$则被x_1阻塞(也是隐藏的).另外也能看到

$$x_2 \not\perp x_6 | x_5, x_7$$

,因为这时候$2 \to 4 \to 7 \to 6$不再被x_7阻塞了(观测到了).练习10.2会给出判断图模型中条件独立关系的更多练习.

10.5.2 图模型的其他马尔科夫性质

根据有向分离准则(the d-separation criterion),可以得出:
$t \perp nd(t)/pa(t) | pa(t)$(10.42)

其中节点nd(t)的非子节点(non-descendants)就是指除了其子节点之外的其他所有节点,
$nd(t) = V/\{t \cup desc(t)\}$.等式10.42也叫做有向局域马尔科夫性质(directed local Markov property).例如在图10.11中,就有$nd(3) = \{2,4\}, pa(3) = 1$,所以$3 \perp 2,4|1$.

这个性质的一个特例是在根据某种拓扑排序查找一个节点的前辈节点(predecessors)的时候.这时候有:
$t \perp pred(t)/pa(t) | pa(t)$(10.43)

其中$pred(t) \subseteq nd(t)$.这叫做有序马尔科夫性质(ordered Markov property),证明了等式10.7.例如在图10.11中,如果排序为1,2,...,7,就发现$pred(3) = 1, 2, pa(3) = 1$,所以$3 \perp 2 | 1$.

这样就有了有向无环图(DAG)的三个马尔科夫性质:有向全局马尔科夫性质G如等式10.34所述,马尔科夫性质O如等式10.43所述,以及有向局域马尔科夫性质L如等式10.42所述.很明显$G \implies L \implies O$.不太明显但也真实成立的是$O \implies L \implies G$(证明参考Koller and Friedman 2009).因此所有这些性质都是等价的.

再进一步,任何关于G的马尔科夫分布p都可以如等式10.7那样因式分解;这也叫做因式分解性质F.很明显$O \implies F$,不过反过来也成立,具体证明还是参考Koller and Friedman 2009.

10.5.3 马尔可夫毯(Markov blanket)和全条件(full conditionals)

马尔可夫毯(Markov blanket)是指图模型中和t条件独立的所有其他节点的集合;记作$mb(t)$.很明显有向图模型(DGM)中一个节点的马尔科夫毯等价于其父节点/子节点/所有共有子节点的节点:

$$mb(t) \triangleq ch(t) \cup pa(t) \cup copa(t) \quad (10.44)$$

例如在图10.11中就有:

$$mb(5) = \{6, 7\} \cup \{2, 3\} \cup \{4\} = \{2, 3, 4, 6, 7\} \quad (10.45)$$

其中的4就和5有共有子节点7.

为啥这些共有子节点的父节点会在马尔科夫毯里面呢?当我们推导$p(x_t | x_{-t}) = p(x_t, x_{-t}) / p(x_{-t})$的时候,所有和$x_t$不相关的项目都会在分子分母之间约掉,所以剩下的就是一系列范围内(scope)包含了x_t的条件概率分布(CPD)的乘积(product).因此:

$$p(x_t | x_{-t}) \propto p(x_t | x_{pa(t)}) \prod_{s \in ch(t)} p(x_s | x_{pa(t)}) \quad (10.46)$$

例如在图10.11中既有:

$$p(x_5 | x_{-5}) \propto p(x_5 | x_2, x_3) p(x_6 | x_3, x_5) p(x_7 | x_4, x_5, x_6) \quad (10.47)$$

得到的这个表达式也叫做t的全条件(full conditional),在学习吉布斯采样(Gibbs sampling,本书24.2)的时候你就会发现这个很重要.

10.6 影响(决策)图解(influence(decision) diagram)*

决策图解(decision diagram)或者也叫做影响图解(influence diagram)是用来表示多阶段(贝叶斯)决策问题的图解(Howard and Matheson 1981;Kjaerulff and Madsen 2008).这种图解在有向图的基础上添加了决策节点(decision nodes)(也叫做行为节点(action nodes)),使用矩形表示;另外还添加

了效用节点(utility nodes)(也叫做值节点(value nodes)),用菱形表示.原来的随机变量就叫做机会节点(chance nodes),依然使用椭圆形表示.

图10.12(a)就是一个例子,其中展示的是石油勘探问题(oil wild-catter problem).在这个问题重要对是否打井来做出决定.有两种行为:d=1意思是打井,d=0就是不打井.然后自然假设有三种可能的状态:o=0意思是枯井啥也没有,o=1意思是井里面有点石油,o=2意思就是井里面有很多石油.加入你的先验信念是$p(o) = [0.5, 0.3, 0.2]$.最终你必须确定效用函数(utility function)$U(d, o)$.由于状态和行为是离散的,所以可以使用一个表格来表示(就类似有向图模型(DGM)里面的条件概率表(CPT)).加入使用下面的数值,单位为美元:

	o=0	o=1	o=2
d=0	0	0	0
d=1	-70	50	200

从上表可见,如果你不打井,那就没啥开销,可是也不赚钱.如果你打了一个枯井,损失70美元;如果是含油量一般的井,赚50美元;如果是大储量的井,赚200美元.那么如果你打井的话,先验期望效用函数就是:

$$EU(d=1) = \sum_{o=0}^{2} p(o) \cup (d, o) = 0.5 \times (-70) + 0.3 \times 50 + 0.2 \times 200 = 20 \text{(10.48)}$$

此处参考原书图10.12

如果你不打井那么期望效用函数就是0.所以最大期望效用函数(maximum expected utility)为:

$$MEU = max\{EU(d=0), EU(d=1)\} = max\{0, 20\} = 20 \text{(10.49)}$$

因此最优行为还是去打井:

$$d^* = \arg\max\{EU(d=0), EU(d=1)\} = 1 \text{(10.50)}$$

然后对这个模型稍作扩展.加入要用声波估计油井状态.声波估计可能有三种状态:s=0意思是弥散反射模式(diffuse reflection pattern),表明没有石油;s=1是开放反射模式,表明有一些是有;s-2是闭合反射模式,说明有可多石油了.由于S是由O引起的,所以增加一条映射弧线(arc)$O \rightarrow S$进入到模型中.另外,假设声波测试的输出可以在决定钻井与否之前就得到;这样就增加了一个信息弧线(information arc)$S \rightarrow D$.如图10.12(b)所示.

然后使用下面的$p(s|o)$的条件分布来对这个传感器的现实情况进行建模吧:

	s=0	**s=1**	**s=2**
o=0	0.6	0.3	0.1
o=1	0.3	0.4	0.3
o=2	0.1	0.4	0.5

假如我们进行了声波检测然后观察得到的是s=0.那么对是有状态的后验就是:
$p(o|s=0)=[0.732,0.219,0.049]$(10.51)

现在对行为d的后验期望效用就是:
$EU(d|s=0)=\sum_{o=0}^{2}p(o|s=0)U(o,d)$(10.52)

如果d=1,则有:

$$EU(d=1|s=0)=0.732\times(-70)+0.219\times 50+0.049\times 200=-30.5$$(10.53)

可是,如果d=0,那么$EU(d=0|s=0)=0$,因为不打井就没开小了.所以如果观测到了s=0,最好就不要打井,这很好理解哈.

然后考虑这种情况,就是进行了声波测试,然后观测到了s=1.这时候$EU(d=1|s=1)=32.9$也还是比$EU(d=0|s=1)=0$高的.类似地,如果观测到了s=2,则有$EU(d=0|s=2)=87.5$,这就比$EU(d=0|s=2)=0$高多了.因此最有的策略$d^*(s)$应该是:如果s=0,就不打井,选择$d^*(0)=0$,不赚钱;如果s=1,就选择$d^*(1)=1$,能赚到32.9美元;如果s=2,就选择$d^*(2)=1$,赚87.5美元.

然后就可以计算你的期望利润(expected profit),也就是最大期望效用函数(maximum expected utility):
$MEU=\sum_s p(s)EU(d^*(s)|s)$(10.54)

上面这是给定声波测试的可能结果之后的期望效用函数,假设的是你根据给定的测试结果进行了最优行为.在测试结果上的先验边缘分布(prior marginal)为:
$p(s)=\sum_o p(o)p(s|o)=[0.41,0.35,0.24]$(10.55)

因此你的最大期望效用函数应该是:
$MEU=0.41\times 0+0.35\times 32.9+0.24\times 87.5=32.2$ (10.56)

然后假设你要选择是否去做测试了.这就可以使用图10.12(c)所示模型来建模,其中增加了一个新的测试节点T.如果T=1,就做测试,S就进入到三个状态之一,然后决定O,就跟上面所述一样.如果T=0,就不做声波测试,S就进入一个特定的未知状态.做声波测试毕竟也是需要花钱的.

是否值得花钱做测试呢?这要取决于如果知道测试结果之后最大期望效用函数(MEU)改变的幅度.如果不做测试,则根据等式10.49可知MEU=20.如果做了测试,参考等式10.56就得到了MEU=32.2.所以如果做了测试(然后根据结果反馈进行最优行为)得到的效用提升是12.2美元.这就叫做完美信息价值(value of perfect information,缩写为VPI).所以只要测试成本低于12.2美元,就应该做.

此处参考原书图10.13

用图模型的术语来说,一个变量T的完美信息价值(VPI)可以通过计算对于基本影响图解(base influence diagram)I的最大期望效用(MEU),然后在计算增加了从T到行为节点的信息弧之后的同样

影响图解下的最大期望效用(MEU),然后计算二者的差别.也就是:
$$VPI = MEU(I + T \to D) - MEU(I)$$(10.57)

其中的D是决策节点,T是要测量的变量.

可以修改变量消除算法(variable elimination algorithm)(参考本书20.3)来进行这种给定影像图条件下的最优策略计算.这些方法本质上都是从最后一步回溯运行的,假设所有后续行为都是最优选择的,计算每一步的最有决策.更多信息可以参考(Lauritzen and Nilsson 2001; Kjaerulff and Madsen 2008).

接下来可以多种方法来扩展这个模型.比如可以想想有一个IE动力系统,然后要测试观测输出,行为选择,接下来到另外一个油井,然后继续打钻(并且制造污染).实际上在机器人/商业/医疗/公共政策等等领域中的很多问题都可以表达成在时间上的影响图(Raiffa 1968; Lauritzen and Nilsson 2001; Kjaerulff and Madsen 2008).

这种形式的通用模型如图10.13(a)所示.这叫做部分观测马尔可夫决策过程(partially observed Markov decision process,缩写为POMDP).这是基于一个隐形马尔科夫模型(HMM,参考本书17.3),参数为行为和奖励节点.这可以用来去对感知行为循环(perception-action cycle)进行建模,这也是很多智能代理所用的方法,更多细节参考(Kaelbling et al. 1998).
部分观测马尔可夫决策过程(POMDP)的一个特例就是其中所有状态都完全观测到了,这就叫做马尔可夫决策过程(Markov decision process,缩写为MDP,如图10.13(b)所示.这就很好解了,因为只要计算一个从观测状态到行为的映射就可以了.这样就可以使用动态编程来解决(更多细节参考(Sutton and Barto 1998)).

在部分观测马尔可夫决策过程(POMDP)中,从x_t到a_t的信息弧并不足以单独决定最优行为,因为状态并没有完全观测到.要基于信念状态(belief state), $p(z_t|x_{1:t}, a_{1:t})$来选择行为.因为新年更新过程是动态的(参考本书17.4.2),可以计算一个信念状态的马尔可夫决策过程(belief state MDP).关于这类模型的计算方面的更多内容可以参考(Kaelbling et al. 1998; Spaan and Vlassis 2005).

此处参考原书图10.14

练习略