第三讲 条件独立性和贝叶斯网络

主讲：王　星
办公电话:86-10-82500167
电子邮箱:wangxingwisdom@126.com

第四节　朴素Bayes分类模型

- Need to know:
  - Priors: $P(Y=y_i)$
  - Conditionals: $P(X=x_k \mid Y=y_i)$
- $P(Y=y_i)$ are easily estimated from data.
  - If n_i of the examples in D are in y_i then $P(Y=y_i) = n_i / |D|$

联合概率：$P(X_1, X_2, \ldots, X_N)$

二值，则有2^N可能的值，其中2^{N-1}个独立。不是二值的话，NP问题？

如果相互独立：

$$P(X_1, X_2, \ldots, X_N) = P(X_1)\, P(X_2) \ldots P(X_N)$$

条件概率：

$$P(X_1, X_2, \ldots, X_N) = P(X_1|X_2, \ldots, X_N)\, P(X_2, \ldots, X_N)$$

迭代表示：

$$P(X_1, X_2, \ldots, X_N) = P(X_1)\, P(X_2| X_1)\, P(X_3| X_2X_1)\ldots P(X_N|X_{N-1}, \ldots, X_1)$$
$$= P(X_N)\, P(X_{N-1}| X_N)\, P(X_{N-2}| X_{N-1}X_N)\ldots P(X_1|X_2, \ldots, X_N)$$

Too many possible instances (e.g. 2^n for binary features) to estimate all $P(X=x_1,\ldots x_k \mid Y=y_i)$.

Still need to make some sort of independence assumptions about the features to make learning tractable.

- 独立：如果X与Y相互独立，则

  $P(X,Y) = P(X)P(Y)$

  $P(X|Y) = P(X)$

- 条件独立：如果在给定Z的条件下，X与Y相互独立，则

  $P(X|Y, Z) = P(X|Z)$

实际中，条件独立比完全独立更普遍

Conditional Independence

Definition: X is <u>conditionally independent</u> of Y given Z, if the probability distribution governing X is independent of the value of Y, given the value of Z

$$(\forall i, j, k) P(X = x_i | Y = y_j, Z = z_k) = P(X = x_i | Z = z_k)$$

Which we often write

$$P(X|Y,Z) = P(X|Z)$$

E.g.,

$$P(Thunder|Rain, Lightning) = P(Thunder|Lightning)$$

Naïve Bayes uses assumption that the X_i are conditionally independent, given Y

then:

$$P(X_1, X_2|Y) = P(X_1|X_2, Y)P(X_2|Y)$$
$$= P(X_1|Y)P(X_2|Y)$$

$$P(X_1 \ldots X_n|Y) = \prod_i P(X_i|Y)$$

How many parameters needed now for P(X|Y)? P(Y)?

$$\theta_{ij} \equiv P(X = x_i|Y = y_j) \qquad \pi_j \equiv P(Y = y_j)$$

Naïve Bayes assumes

$X = \langle X_1, \ldots, X_n \rangle$, Y discrete-valued

$$P(X_1 \ldots X_n|Y) = \prod_i P(X_i|Y)$$

i.e., that X_i and X_j are conditionally independent given Y, for all i≠j

Maximum likelihood estimates:

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\}}{|D|}$$

$$\hat{\theta}_{ijk} = \hat{P}(X_i = x_{ij}|Y = y_k) = \frac{\#D\{X_i = x_{ij} \wedge Y = y_k\}}{\#D\{Y = y_k\}}$$

MAP estimates (uniform Dirichlet priors):

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\} + l}{|D| + lR}$$

$$\hat{\theta}_{ijk} = \hat{P}(X_i = x_{ij}|Y = y_k) = \frac{\#D\{X_i = x_{ij} \wedge Y = y_k\} + l}{\#D\{Y = y_k\} + lM}$$

- 平滑参数l，R表示Y的类别数，M表示的X变量的取值数To account for estimation from small samples, probability estimates are adjusted or *smoothed*.
- Laplace smoothing using an *m*-estimate assumes that each feature is given a prior probability, *p*, that is assumed to have been previously observed in a “virtual” sample of size *m*.
- For binary features, *p* is simply assumed to be 0.5.

- Learning (generative) classifiers based on Bayes rule
- Conditional independence
  - What it is
  - Why it’s important
- Naïve Bayes assumption and its consequences
- Naïve Bayes with discrete inputs, continuous inputs

Outlook	Temperature	Humidity	Windy	Class
sunny	hot	high	false	N
sunny	hot	high	true	N
overcast	hot	high	false	P
rain	mild	high	false	P
rain	cool	normal	false	P
rain	cool	normal	true	N
overcast	cool	normal	true	P
sunny	mild	high	false	N
sunny	cool	normal	false	P
rain	mild	normal	false	P
sunny	mild	normal	true	P
overcast	mild	high	true	P
overcast	hot	normal	false	P
rain	mild	high	true	N

P(p) = 9/14	
P(n) = 5/14	

outlook	
P(sunny|p) = 2/9	P(sunny|n) = 3/5
P(overcast|p) = 4/9	P(overcast|n) = 0
P(rain|p) = 3/9	P(rain|n) = 2/5
temperature	
P(hot|p) = 2/9	P(hot|n) = 2/5
P(mild|p) = 4/9	P(mild|n) = 2/5
P(cool|p) = 3/9	P(cool|n) = 1/5
humidity	
P(high|p) = 3/9	P(high|n) = 4/5
P(normal|p) = 6/9	P(normal|n) = 1/5
windy	
P(true|p) = 3/9	P(true|n) = 3/5
P(false|p) = 6/9	P(false|n) = 2/5

20201204作业1：使用朴素贝叶斯计算如下分类结果，
1. P(C=? | Outlook=Rain,TempCool, Hum=High,Windy=True)
2. P(C=? | Outlook=Overcast,TempCool, Hum=High,Windy=True)

朴素贝叶斯分类的优点和缺点

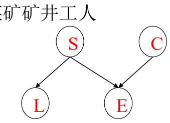
- 优点
  - 计算速度比较快
  - 规则清楚易懂
  - 独立事件假设，大多数问题不至于发生太大偏差
- 不足：
  - 只能用于类别变量
  - 只能用于分类
  - 假设变量之间独立互不影响，使用时需要谨慎分析变量之间的相关性。

第三节 Bayes 网络

- 1988年由Pearl提出, 贝叶斯网络（Bayes Network）成功地应用于知识发现领域，成为表示不确定性知识和推理的一种流行的方法。基于贝叶斯方法的贝叶斯网络是一种适应性很广的手段和工具，具有坚实的数学理论基础。在综合先验信息（领域知识）和数据样本信息的前提下，还可避免只使用先验信息可能带来的主观偏见。虽然很多贝叶斯网络涉及的学习问题是NP难解的。但是，由于已经有了一些成熟的近似解法，加上一些限制后计算可大为简化，很多问题可以利用近似解法求解。
- 研究变量和变量之间关系的重要方法。
- 是图论与概率论的结合。
- 贝叶斯网络又称信度网络，是Bayes方法的扩展，不确定知识表达和推理领域最有效的理论模型之一。

贝叶斯网络（因果关系网络）

假设：
- 命题S(smoker)：该患者是一个吸烟者
- 命题C(coal Miner)：该患者是一个煤矿矿井工人
- 命题L(lung Cancer)：他患了肺癌
- 命题E(emphysema)：他患了肺气肿
- 由专家给定的假设可知：
  - 命题S对命题L和命题E有因果影响，
  - 而C对E也有因果影响。
  - 命题之间的关系可以描绘成因果关系网。每一个节点代表一个证据，每一条弧代表一条规则（假设），连接结点的弧表达了有规则给出的节点间的直接因果关系。其中，节点S，C是节点L和E的父节点或称双亲节点，同时，L，E也称为是S和C的子节点或称后代节点。



贝叶斯网络的由来

- 全联合概率计算复杂性十分巨大
- 朴素贝叶斯太过简单
- 现实需要一种自然、有效的方式来捕捉和推理——不确定性知识
- 变量之间的独立性和条件独立性可大大减少为了定义全联合概率分布所需的概率数目

贝叶斯网络的定义

一个 Bayes 网络是一个二元组 $B=\langle G,\Theta\rangle$,其中

1) $G=\{\Pi_1,\Pi_2,\cdots,\Pi_N\}$ 是一个有向无环图(DAG),其结点为

$$U=\{X_1,X_2,\cdots,X_N\},N\geqslant 1,$$

$\Pi_i\subset 2^U,i=1,\cdots,N$ 是结点 X_i 的父结点集合;

2) $\Theta=\{P(X_i\mid\Pi_i)\mid X_i\in U,i=1,\cdots,N\}$ 是一组条件概率的集合,称为网络参数.

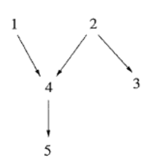
一个 Bayes 网络 $B=\langle G,\Theta\rangle$ 确定了一个概率空间 $\langle\Omega,F,P\rangle$, 其中

$$P(U)=\prod_{i=1}^{N}P(X_i\mid\Pi_i). \qquad (1)$$

根据 Bayes 网络的定义,从数据样本中学习 Bayes 网络,就是要学习网络模型的结构 G 和参数

- 有向无环图(Directed Acyclic Graph, DAG)
- 随机变量集组成网络节点，变量可离散或连续
- 一个连接节点对的有向边或箭头集合

- 条件独立：贝叶斯网络中的一个结点，如果父母结点已知，则它条件独立于它的所有非后代结点
- **每个节点*Xi*都有一个条件概率分布表：*P(Xi|Parents(Xi))*，每个节点受其他节点的影响都可以转化为只受其父节点对该节点的影响**

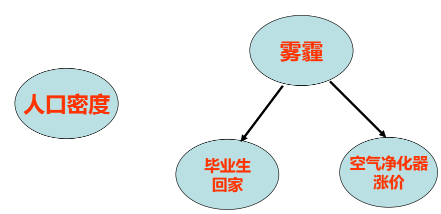


$$p(\mathbf{X})=p(X_5\mid X_4)p(X_4\mid X_1,X_2)p(X_3\mid X_2)p(X_2)p(X_1)$$

给出一个图 G，观测D的似然函数可以表示为:

$$P(D\mid G)=\prod_{i=1}^{n}p(X_i\mid parents(X_i))$$

独立和条件独立性



- 没有边的两个节点之间互相独立，例如： Population和其它3个变量相互独立
- 如果B隔离了A和C时，那么可以认为A与C是关于B条件独立的给定雾霾后，毕业生就业和销量条件独立

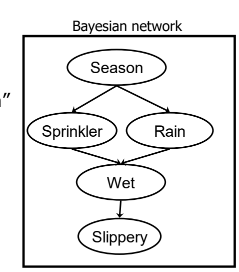
Bayesian Network的应用

We want to describe the causal relationship between the following events:

1) The season
2) Whether it is raining outside
3) The sprinkler is on
4) The sidewalk is wet
5) The sidewalk is slippery

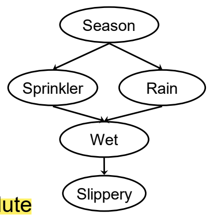
假设:

- “Sprinkler on” and “Rain” are determined by “Season”
- “Sidewalk wet” is determined by “Sprinkler on” and “Rain”
- “Sidewalk slippery” is determined by “Sidewalk wet”



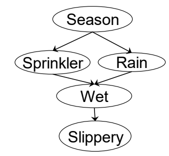
- 每个节点代表一个随机变量，这里这个随机变量是一个二值变量，所以代表一个随机事件是否发生。

Properties of Bayesian Networks



- Links are *not* absolute
  - If the sprinkler is on, this does not always mean that the sidewalk is wet
  - For example, the sprinkler may be aimed away from the sidewalk

Properties of Bayesian Networks



- Given that the sidewalk is wet, we can calculate the probability that the sprinkler is on:

  P(sprinkler on | sidewalk wet)

- Bayesian networks allow us to calculate such values from a small set of probabilities in a process called reasoning or Bayesian Inference

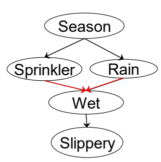
Bayesian Networks
三种常用的推理方式

- Bayesian networks通过父子传递可以在任何一个方向上推理

  1) If the sprinkler is on, the sidewalk is probably wet (prediction预测)

  2) If the sidewalk is wet, it is more likely that the sprinkler is on or it is raining (diagnosis诊断)

  3) If the sidewalk is wet and the sprinkler is on, the likelihood that it is raining is reduced (explaining away解释远离，消解影响)



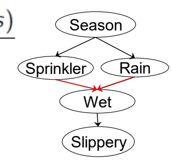
- Explaining away is a special type of reasoning that is especially difficult to model in other network models

解释远离

$$P(r|w,s) = \frac{P(w|r,s)P(r|s)}{P(w|s)}$$

如果 $P(w|s) = P(w|r,s)$

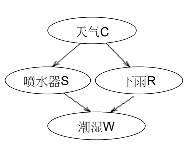
$$P(r|w,s) = P(r|s) = P(r)$$



这说明*S*决定了*W*的发生，*R*则发生了对*W*的解释远离

Bayesian Network的分布表

P(C=F)	P(C=T)
0.5	0.5

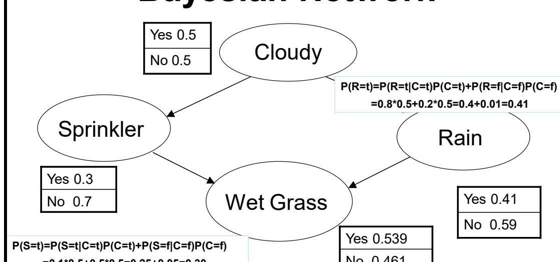


C	P(S=F)	P(S=T)
F	0.5	0.5
T	0.9	0.1

C	P(R=F)	P(R=T)
F	0.8	0.2
T	0.2	0.8

S	R	P(W=F)	P(W=T)
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

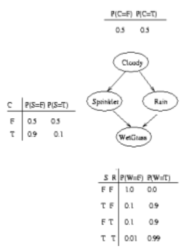
Bayesian Network



P(W=t)=P(W=t|S=t,R=t)P(S=t)P(R=t)+P(W=t|S=t,R=f)P(S=t)P(R=f)
+P(W=t|S=f,R=t)P(S=f)P(R=t)+P(W=t|S=f,R=f)P(S=f)P(R=f)
=0.99*0.41*0.30+0.9*(1-0.41)*0.30+0.9*0.41*(1-0.30)+0.0*(1-0.30)*(1-0.41)
= 0.53937

Diagnosis

We observe the grass is wet- 2 causes sprinkler or rain .. Which is more likely ???



Pr(S=1|W=1) = Σ Pr(S=1, W=1) / Pr(W=1) = 0.2781/0.539

Pr(R=1|W=1) = Σ Pr(R=1, W=1) / Pr(W=1) = 0.4581/0.539

Normalizing Pr(W=1) = 0.539

贝叶斯网络建模的方法

贝叶斯网络建模一般有三种方法：

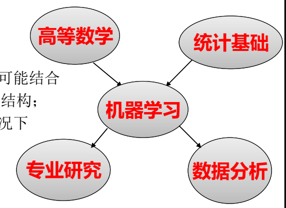
1）依靠专家和背景建模；

2）从数据中学习；

3）从知识库中创建。

贝叶斯网络学习主要研究**结构学习**与**参数学习**两个方面。

1. 结构学习是指利用训练本集，尽可能结合先验知识，确定合适的贝叶斯网络拓扑结构；
2. 参数学习是指在网络结构确定的情况下从数据中学习每一个节点的条件概率表

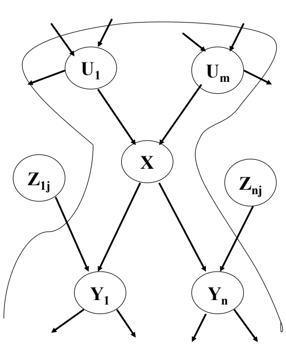


贝叶斯网络的特性：

- 作为对域的一种完备而无冗余的表示，贝叶斯网络比全联合概率分布紧凑得多
- BN的紧凑性是**局部结构化**(Locally structured, 也称**稀疏**, Sparse)系统一个非常普遍特性的实例
- BN中每个节点只与数量有限的其它节点发生**直接的**相互作用
- 假设节点数n=30, 每节点有5个父节点，则BN需$30\text{x}2^5=960$个数据，而全联合概率分布需要2^{30}= 10亿个！

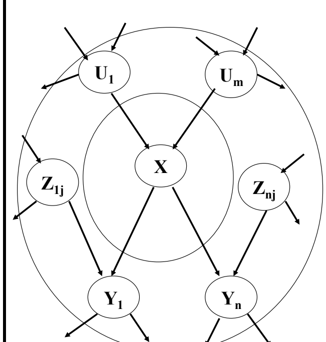
贝叶斯网络中的条件独立关系：

- 给定父节点，一个节点与它的**非后代节点**是条件独立的
- 给定一个节点的父节点、子节点以及子节点的父节点——**马尔可夫覆盖**(Markov blanket)，这个节点和网络中的所有其它节点是条件独立的



【说明】：

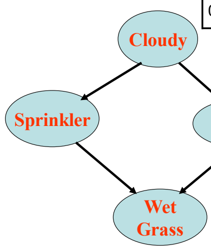
给定节点X的父节点U_1…U_m，节点X与它的非后代节点（即Z_{ij}）是条件独立的。



【说明】：

给定**马尔可夫覆盖**（两圆圈之间的区域），节点X和网络中所有其它节点都是条件独立的。

多连通网络及其CPT（Conditional Probability Table）：



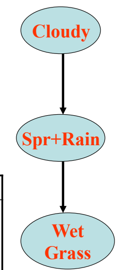
P(C)
0.50

C	P(S)
t	0.10
f	0.90

C	P(R)
t	0.80
f	0.20

S	R	P(W)
t	t	0.99
t	f	0.90
f	t	0.90
f	f	0.00

等价的联合树及其CPT：



P(C)
0.50

C	P(S+R = x) tt	tf	ft	ff
t	.08	.02	.72	.18
f	.10	.40	.10	.40

S + R		P(W)
t	t	0.99
t	f	0.90
f	t	0.90
f	f	0.00

网络模型结构G和参数θ

小规模贝叶斯网络构造原则：

- 首先，添加“**根本原因**”节点
- 然后，加入受它们**直接影响的变量**
- 依次类推，直到**叶节点**，即对其它变量没有直接因果影响的节点
- 两节点间的有向边的取舍原则：更高精度概率的重要性与指定额外信息的代价的折衷
- “因果模型”比“诊断模型”需要更少的数据，且这些数据也更容易得到

贝叶斯网络的结构学习的要点

贝叶斯网络的结构学习经常被视为是一种最优化问题，计算的任务就是找到最佳的结构使得统计学意义上的得分最高。在构建贝叶斯网络的过程中，并不能对所有的结构分别进行计算评估，只能采用启发式搜索算法，在有限的搜索空间中寻优。

贝叶斯网络算法的两种类型

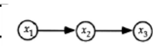
- **变量相关性分析：**
  最大加权生成树算法(m a x i m u m w e i g h t e d s p a n n i n g t r e e，M W S T）
  - 三阶段相关性分析算法（t h r e e p h a s e d e p e n d e n c y a n a l y s i s，T P D A）等；
- **基于搜索评分的学习：比如网络得分**，如B算法、K2算法、基于蚁群优化的B算法（a n t c o l o n y o p t i m i z a t i o n－B a l g o r i t h m，A C O B；
- 例：K2算法,1992年,目标是估计后验概率
  - 该算法要求先确定网络中节点变量的次序，对先验知识的依赖性很大。
  - 在该算法中提出了节点模块化思路，即各节点的父节点集相互独立

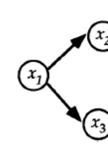
$$\max_{B_s}[P(G,D)] = C\prod_{i=1}^{n}\max_{\Pi_i}[\prod_{j=1}^{q_j}\frac{(r_i-1)!}{(N_{ij}+r_i-1)!}\prod_{k=1}^{r_i}N_{ijk}!]$$

  - 困难：不了解相关的领域知识或没有专家指导的情况下，确定变量的次序就变得相当困难。

Table 1. A database example. The term *case* in the first column denotes a single training instance (record) in the database—as for example, a patient case. For brevity, in the text we sometimes use 0 to denote *absent* and 1 to denote *present*.

	Variable values for each case		
Case	x_1	x_2	x_3
1	*present*	*absent*	*absent*
2	*present*	*present*	*present*
3	*absent*	*absent*	*present*
4	*present*	*present*	*present*
5	*absent*	*absent*	*absent*
6	*absent*	*present*	*present*
7	*present*	*present*	*present*
8	*absent*	*absent*	*absent*
9	*present*	*present*	*present*
10	*absent*	*absent*	*absent*





$$P(B_S, D) = \int_{B_p} P(D \mid B_S, B_P) f(B_P \mid B_S) P(B_S) dB_P,$$

- BP是一个向量，其值表示在bayes网络结构Bs下的条件概率分布。

$$P(B_S, D) = \int_{B_P} \left[\prod_{h=1}^{m} P(C_h \mid B_S, B_P) \right] f(B_P \mid B_S) P(B_S) dB_P,$$

where m is the number of cases in D and C_h is the hth case in D.

Theorem 1. Let Z be a set of n discrete variables, where a variable x_i in Z has r_i possible value assignments: $(v_{i1}, \ldots, v_{ir_i})$. Let D be a database of m cases, where each case contains a value assignment for each variable in Z. Let B_S denote a belief-network structure containing just the variables in Z. Each variable x_i in B_S has a set of parents, which we represent with a list of variables π_i. Let w_{ij} denote the jth unique instantiation of π_i relative to D. Suppose there are q_i such unique instantiations of π_i. Define N_{ijk} to be the number of cases in D in which variable x_i has the value v_{ik} and π_i is instantiated as w_{ij}. Let

$$N_{ij} = \sum_{k=1}^{r_i} N_{ijk}.$$

1. The variables in Z are discrete
2. Cases occur independently, given a belief-network model
3. There are no cases that have variables with missing values
4. Before observing D, we are indifferent regarding which numerical probabilities to assign to the belief network with structure B_S.

$$P(B_S, D) = P(B_S) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}!$$

大规模贝叶斯网络的近似推理

- **大规模多连通BN的精确推理是不可操作的，只能通过近似推理来解决。**
- **后验概率计算的主要采样方法**

  直接采样方法,直接采样算法,拒绝采样(Rejection sampling)算法,似然加权(Likelihood weighting)算法

  - **马尔可夫链蒙特卡罗（MCMC）方法**
  - 变分法(Variational method)
  - 环传播(Loopy propagation)方法

马尔可夫链蒙特卡罗结构学习算法（Metropolis Hasting）

算法的主要思想是：由于网络结构的后验分布 $P(G \mid D)$是无法直接计算得到的，可以首先构造一个 markov 链，使其极限分布收敛于网络结构的后验分布 $P(G \mid D)$；然后使用 Monte Carlo 方法对此 markov 链进行抽样，得到网络结构的样本序列，即 $G^0, G^1, \ldots, G^i, \ldots$)；最后从此序列中挑出具有最大后验概率的网络结构，来近似网络的最优结构。算法中从第 i 个网络结构 G 转移到新网络结构 G' 的接受概率为：

$$\alpha(G, G') = \min\{1, R_\alpha\} \tag{2}$$

$$R_\alpha = \frac{\#(nbd(G)) P(G' \mid D)}{\#(nbd(G')) P(G \mid D)} \tag{3}$$

式中， # nbd G表示由G 和那些对G 实行一次边的简单操作（删除边、增加边、改变边方向）得到的图构成的集合，称为G 的邻近域。#(( )) nbd G 为G 的邻近域中元素的个数。

贝叶斯信息准则（Bayesian information criterion）是实际中最常用的模型评分函数，简称 BIC 评分。BIC 评分是在大样本前提下对边缘似然函数的近似。通常利用拉普拉斯近似方法对 $P(D|G)$ 进行大样本近似从而得出 BIC 评分函数。基本的想法就是在最大似然估计附近将对数似然函数做 Taylor 展开，将计算转化为一个多元正态分布函数在极值点的邻域的积分，具体如下：

用 θ^*表示 θ 的极大似然估计，假设 $P(D|G)$ 在 θ^*的附近光滑的且不为零，记

$l(\theta) = P(D|G,\Theta)$，将 $l(\theta)$ 在 θ^*的一个邻域展开得：

$$l(\theta) \approx l(\theta^*) + \tfrac{1}{2}(\theta - \theta^*)^T l''(\theta^*)(\theta - \theta^*)$$

用A表示 $-l''(\theta^*)$， $\log P(D|G)$ 的拉普拉斯近似为

$$\log P(D|G) \approx \log P(D|G,\theta^*) - \tfrac{d}{2}\log m$$

上式称为模型结构G的BIC评分，记为BIC($G|D$)。

对MCMC算法的评价

- 该方法是解决结构未知和数据完备情况的网络结构学习问题。
- 优点：不需要领域的先验知识，对于大多数知识发现领域具有极大优势。
- 缺点：该算法受初始网络结构影响较大，且学习结果易陷入局部最大化。

为什么需要MCMC

产生独立样本
- 基本方法：概率积分变换（第一部分已讲）
- 接受—拒绝（舍选）采样
- 重要性采样

产生相关样本：
Markov Chain Monte Carlo
- Metropolis-Hastings算法
- Gibbs Sampler

需要MCMC的原因

重要性采样和接受—拒绝采样都只$q(x)$与$\pi(x)$很相近似时才表现很好

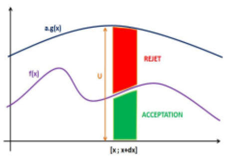
在高维空间问题中，标准的采样方法会失败：
- 接受—拒绝采样：维数增高时，拒绝率→100%
- 重要性采样：大多数的样本权重→ 0

对高维复杂问题，用马尔科夫链（Markov Chain） 产生一些列相关样本，实现对分布的采样

拒绝抽样的原理（Wiki）

- 当我们需要随机生成一个随机变量 X 满足概率分布 f 时，我们可以选择一个更容易生成的概率分布 g 。当然，不是任意的 g 都可以胜任，g 必须满足:

$$\exists a \geq 1 / \forall x \in \mathbb{R}^d, f(x) \leq a \cdot g(x)$$



拒绝性算法

1. 模拟
   - 从概率分布 g 中抽取 Y
   - 从一个定义域为 [0,1] 并与 Y 相互独立的均匀概率分布中抽取 U
2. 接受或者拒绝
   - 如果 $U < \frac{f(Y)}{a \cdot g(Y)}$，那么就令 $X=Y$
   - 否则重新开始

Metropolis-Hastings算法

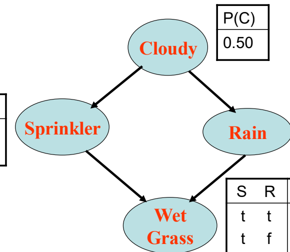
等价于

- 给定在当前状态 $X_t = x$，
- 1、 产生 $Y_t \sim q(y \mid x)$
- 2、$U \sim Uniform[0,1]$
- 3、 $X_{t+1} = \begin{cases} Y_t & U \leq \alpha(x, Y_t) \\ X_t & 其它 \end{cases}$
- 其中 $\alpha(x,y) = \min\left(1, \frac{q(x \mid y)}{q(y \mid x)} \frac{\pi(y)}{\pi(x)}\right)$

马尔可夫链蒙特卡罗（MCMC）参数估计算法思想:

- 对前一个事件进行随机改变而生成事件样本
- BN为每个变量指定了一个特定的当前状态
- 下一个状态是通过对某个非证据变量X_i进行采样来产生，取决于X_i的马尔可夫覆盖中的变量当前值
- MCMC方法可视为：在状态空间中—所有可能的完整赋值空间—的随机游走—每次改变一个变量，但是证据变量的值固定不变。

MCMC算法进行参数学习的执行过程示例:



P(C)
0.50

C	P(S)
t	0.10
f	0.50

C	P(R)
t	0.80
f	0.20

S	R	P(W)
t	t	0.99
t	f	0.90
f	t	0.90
f	f	0.00

【要求】：查询P(Rain | Sprinkler = true, WetGrass = true)的概率，理论值：0.523

MCMC算法执行步骤:

- **证据变量**Sprinkler, WetGrass固定为true
- **隐变量**Cloudy和**查询变量**Rain随机初始化，例如， Cloudy = true, Rain = false，初始状态为：**[C=true, S=true, R=false, W=true]**
- 反复执行如下步骤：

（1） 根据Cloudy的马尔可夫覆盖(MB)变量的当前值，**对Cloudy采样**，即根据P(Cloudy|Sprinkler= true, Rain=false)（即转移概率）来采样。即：

P(C|S, ~R) = P(C,S,~R) / P(S, ~R) = P(C)P(S|C)P(~R|C) / [P(C)P(S|C)P(~R|C)+P(~C)P(S|~C)P(~R|~C)]
=(0.5×0.1×0.2) / [0.5×0.1×0.2+0.5×0.5×0.8]=0.04762

再由计算机生成一个随机数$q \in [0,1]$（可参照《概率统计》中的随机数生成方法）。比较转移概率值与随机数q的大小，以决定是继续停留在原状态，还是转移到下一个新的状态。判别方法如下:

if q < P(C|S, ~R) then 转移到下一个新状态;
otherwise 停留在原状态.

对于本例子，假设生成的随机数$q=0.0389$，可知，转移概率$P(Cloudy|Sprinkler= true, Rain=false)= 0.04762 > q=0.0389$，所以，Cloudy由true状态转移到新状态false，即采样结果为：Cloudy = false。故新的当前状态为:

[C=false, S=true, R=false, W=true]

（2） 根据Rain节点的马尔可夫覆盖(MB)变量的当前值，对Rain采样，即根据P(Rain | Cloudy = false, Sprinkler = true, WetGrass = true)来采样。假设采样结果为：Rain = true。故新的当前状态为：

[C=false, S=true, R=true, W=true]

【注】： 上述过程中所访问的每一个状态都是一个样本，能对查询变量Rain的估计有贡献。

（3）重复上述步骤，直到所要求的访问次数N。

若为true, false的次数分别为n_1, n_2，则查询解为：

Normalize(<n_1, n_2>) = < n_1 /N, n_1 /N>

若上述过程访问了20个Rain=true的状态和60个Rain = false的状态，则所求查询的解为<0.25, 0.75>。

马尔可夫链蒙特卡罗（MCMC）算法描述：

function MCMC-Ask(*X, e, bn, N*) **return** P(X | e)

local variables: N[X]， //关于查询变量X的向量计数，初值0

Z，//*bn*中的非证据变量集

x， // bn的当前状态

利用Z中变量的随机值来初始化x；

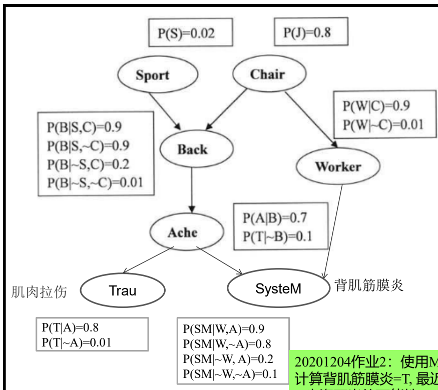
for *j* = *1* to *N* **do**

N(*x*) ← N(*x*) + 1； //*x*是当前状态**x**中的查询变量X的值

for each Z_i in Z **do**

给出Z_i的马尔可夫覆盖MB(Z_i)，并根据P(Z_i |mb(Z_i))来采样的Z_i值；

return Normalize(N[X]) //对N[X]进行归一化



20201204作业2：使用MCMC算法
计算背肌筋膜炎=T, 最近做运动=F的条件下,
1.坐姿不当的可能性：1.P(C=T| SM=T,S=F)
2.伏案过劳的可能性：2.P(W=T|SM=T,S=F)