法律声明

- 本课件包括演示文稿、示例、代码、题库、视频和声音等内容，小象学院和主讲老师拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意及内容，我们保留一切通过法律手段追究违反者的权利。

- 课程详情请咨询
  - 微信公众号：小象
  - 新浪微博：ChinaHadoop

贝叶斯网络

邹博

主要内容

- 复习本次将用到的知识
  - 相对熵、互信息(信息增益)
- 朴素贝叶斯
- 贝叶斯网络的表达
  - 条件概率表参数个数分析
  - 马尔科夫模型
- D-separation
  - 条件独立的三种类型
  - Markov Blanket
- 网络的构建流程
  - 混合(离散+连续)网络：线性高斯模型
- Chow-Liu算法：最大权生成树MSWT

复习：相对熵

- 相对熵，又称互熵，交叉熵，鉴别信息，Kullback熵，Kullback-Leible散度等
- 设p(x)、q(x)是X中取值的两个概率分布，则p对q的相对熵是

$$D(p\|q)=\sum_{x}p(x)\log\frac{p(x)}{q(x)}=E_{p(x)}\log\frac{p(x)}{q(x)}$$

- 说明：
  - 相对熵可以度量两个随机变量的“距离”
  - 一般的，D(p||q) ≠D(q||p)
  - D(p||q)≥0、 D(q||p) ≥0

复习：互信息

- 两个随机变量X，Y的互信息，定义为X，Y的联合分布和独立分布乘积的相对熵。
- I(X,Y)=D(P(X,Y) || P(X)P(Y))

$$I(X,Y)=\sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$$

复习：信息增益

- 信息增益表示得知特征A的信息而使得类X的信息的不确定性减少的程度。
- 定义：特征A对训练数据集D的信息增益g(D,A)，定义为集合D的经验熵H(D)与特征A给定条件下D的经验条件熵H(D|A)之差，即：
  - $g(D,A)=H(D)-H(D|A)$
  - 显然，这即为训练数据集D和特征A的互信息。

概率

- 条件概率：

$$P(A|B)=\frac{P(AB)}{P(B)}$$

- 全概率公式：

$$P(A)=\sum_i P(A|B_i)P(B_i)$$

- 贝叶斯(Bayes)公式：

$$P(B_i|A)=\frac{P(A|B_i)P(B_i)}{\sum_j P(A|B_j)P(B_j)}$$

贝叶斯公式带来的思考 $P(A|D)=\frac{P(D|A)P(A)}{P(D)}$

- 给定某些样本D，在这些样本中计算某结论A_1、A_2.....A_n出现的概率，即$P(A_i|D)$

$$\max P(A_i|D)=\max\frac{P(D|A_i)P(A_i)}{P(D)}=\max(P(D|A_i)P(A_i))\rightarrow\max P(D|A_i)$$

$$\Rightarrow \max P(A_i|D)\rightarrow\max P(D|A_i)$$

  - 第一个等式：贝叶斯公式；
  - 第二个等式：样本给定，则对于任何Ai,P(D)是常数，仅为归一化因子；
  - 第三个箭头：若这些结论A_1、A_2.....A_n的先验概率相等(或近似)，则得到最后一个等式：即第二行的公式。

贝叶斯公式的应用

- 8支步枪中有5支已校准过，3支未校准。一名射手用校准过的枪射击，中靶概率为0.8；用未校准的枪射击，中靶概率为0.3；现从8支枪中随机取一支射击，结果中靶。求该枪是已校准过的概率。

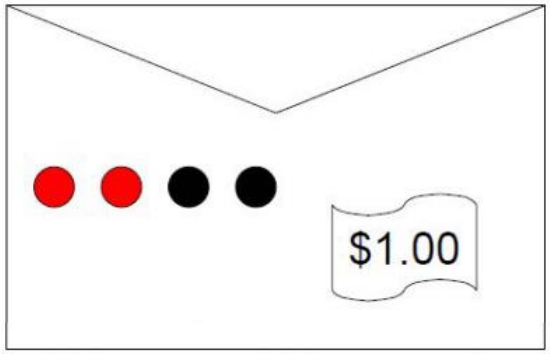
$$P(G=1)=\frac{5}{8} \qquad P(G=0)=\frac{3}{8}$$

- 解：

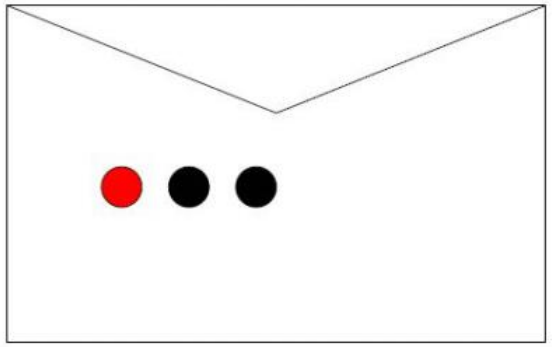
$$P(A=1|G=1)=0.8 \quad P(A=0|G=1)=0.2$$
$$P(A=1|G=0)=0.3 \quad P(A=0|G=0)=0.7$$
$$P(G=1|A=1)=?$$

$$P(G=1|A=1)=\frac{P(A=1|G=1)P(G=1)}{\sum_{i\in G}P(A=1|G=i)P(G=i)}=\frac{0.8\times\frac{5}{8}}{0.8\times\frac{5}{8}+0.3\times\frac{3}{8}}=0.8163$$

另一个实例



The "Win" envelope has a dollar and four beads in it



The "Lose" envelope has three beads and no money

Interesting question: before deciding, you are allowed to see one bead drawn from the envelope.

Suppose it's black: How much should you pay?

Suppose it's red: How much should you pay?

后验概率

- c1、c2表示左右两个信封。
- P(R)，P(B)表示摸到红球、黑球的概率。
- P(R)=P(R|c1)*P(c1) + P(R|c2)*P(c2)：全概率公式
- P(c1|R)=P(R|c1)*P(c1)/P(R)
  - P(R|c1)=2/4
  - P(R|c2)=1/3
  - P(c1)=P(c2)=1/2
- 如果摸到红球，则该信封有1美元的概率是0.6
- 如果摸到黑球，则该信封有1美元的概率是3/7

朴素贝叶斯的假设

- 一个特征出现的概率，与其他特征(条件)独立(特征独立性)
  - 其实是：对于给定分类的条件下，特征独立
- 每个特征同等重要(特征均衡性)

以文本分类为例

- 样本：10000封邮件，每个邮件被标记为垃圾邮件或者非垃圾邮件
- 分类目标：给定第10001封邮件，确定它是垃圾邮件还是非垃圾邮件
- 方法：朴素贝叶斯

分析

- 类别c：垃圾邮件c_1，非垃圾邮件c_2
- 词汇表，两种建立方法：
  - 使用现成的单词词典；
  - 将所有邮件中出现的单词都统计出来，得到词典。
  - 记单词数目为N
- 将每个邮件m映射成维度为N的向量**x**
  - 若单词w_i在邮件m中出现过，则$x_i=1$，否则，$x_i=0$。即邮件的向量化：$m \rightarrow (x_1,x_2.....x_N)$
- 贝叶斯公式：$P(c|\mathbf{x})=P(\mathbf{x}|c)*P(c) / P(\mathbf{x})$
  - $P(c_1|\mathbf{x})=P(\mathbf{x}|c_1)*P(c_1) / P(\mathbf{x})$
  - $P(c_2|\mathbf{x})=P(\mathbf{x}|c_2)*P(c_2) / P(\mathbf{x})$
    - 注意这里**x**是向量

分解

- $P(c|\mathbf{x})=P(\mathbf{x}|c)*P(c) / P(\mathbf{x})$
- $P(\mathbf{x}|c)=P(x_1,x_2\ldots x_N|c)=P(x_1|c)*P(x_2|c)\ldots P(x_N|c)$
  - 特征条件独立假设
- $P(\mathbf{x})=P(x_1,x_2\ldots x_N)=P(x_1)*P(x_2)\ldots P(x_N)$
  - 特征独立假设
- 带入公式： $P(c|\mathbf{x})=P(\mathbf{x}|c)*P(c) / P(\mathbf{x})$

- 等式右侧各项的含义：
  - $P(x_i|c_j)$：在c_j(此题目，cj要么为垃圾邮件1，要么为非垃圾邮件2)的前提下，第i个单词x_i出现的概率
  - $P(x_i)$：在所有样本中，单词x_i出现的概率
  - $P(c_j)$：在所有样本中，邮件类别c_j出现的概率

拉普拉斯平滑

- $p(x_1|c_1)$是指的:在垃圾邮件c_1这个类别中，单词x1出现的概率。
  - x_1是待考察的邮件中的某个单词
- 定义符号
  - n_1：在所有垃圾邮件中单词x_1出现的次数。如果x_1没有出现过，则$n_1=0$。
  - n：属于c_1类的所有文档的出现过的单词总数目。
- 得到公式：$$p(x_1|c_1)=\frac{n_1}{n}$$
- 拉普拉斯平滑：$$p(x_1|c_1)=\frac{n_1+1}{n+N}$$
  - 其中，N是所有单词的数目。修正分母是为了保证概率和为1
- 同理，以同样的平滑方案处理$p(x_1)$

对朴素贝叶斯的思考

- 拉普拉斯平滑能够避免0/0带来的算法异常
- 要比较的是P(c1|x)和P(c2|x) 的相对大小，而根据公式P(c|x) =P(x|c)*P(c) / P(x)，二者的分母都是除以P(x)，实践时可以不计算该系数。
- 编程的限制：小数乘积下溢出怎么办？
- 问题：一个词在样本中出现多次，和一个词在样本中出现一次，形成的词向量相同
  - 由0/1改成记数
- 如何判断两个文档的距离
  - 夹角余弦
- 如何判定该分类器的正确率
  - 交叉验证

贝叶斯网络

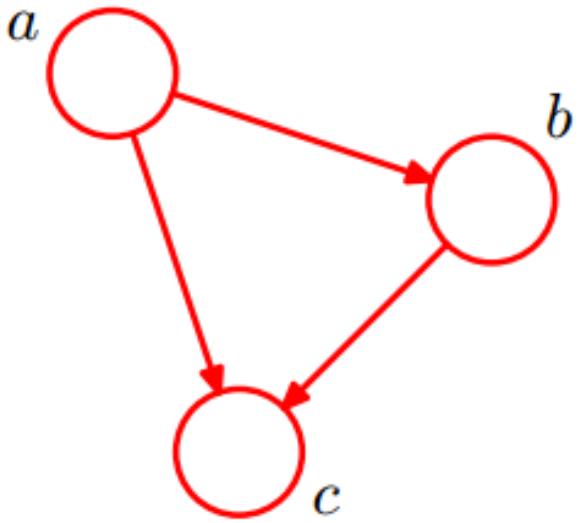
- 把某个研究系统中涉及的随机变量，根据是否条件独立绘制在一个有向图中，就形成了贝叶斯网络。
- 贝叶斯网络(Bayesian Network)，又称有向无环图模型(directed acyclic graphical model ,DAG)，是一种概率图模型，根据概率图的拓扑结构，考察一组随机变量$\{X_1,X_2...X_n\}$及其n组条件概率分布(Conditional Probability Distributions, CPD)的性质。

贝叶斯网络

- 一般而言，贝叶斯网络的有向无环图中的节点表示随机变量，它们可以是可观察到的变量，或隐变量、未知参数等。连接两个节点的箭头代表此两个随机变量是具有因果关系(或非条件独立)。若两个节点间以一个单箭头连接在一起，表示其中一个节点是“因(parents)”，另一个是“果(children)”，两节点就会产生一个条件概率值。
- 每个结点在给定其直接前驱时，条件独立于其非后继。
  - 稍后详细解释此结论

一个简单的贝叶斯网络

$$p(a,b,c) = p(c|a,b)p(b|a)p(a)$$



全连接贝叶斯网络

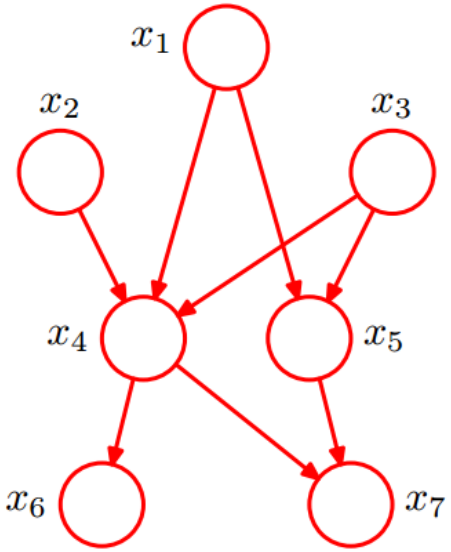
- 每一对结点之间都有边连接

$$p(x_1,\ldots,x_K) = p(x_K|x_1,\ldots,x_{K-1})\ldots p(x_2|x_1)p(x_1)$$

$$\mathrm{P}(X_1 = x_1,\ldots,X_n = x_n) = \prod_{i=1}^{n} \mathrm{P}(X_i = x_i \mid X_{i+1} = x_{i+1},\ldots,X_n = x_n)$$

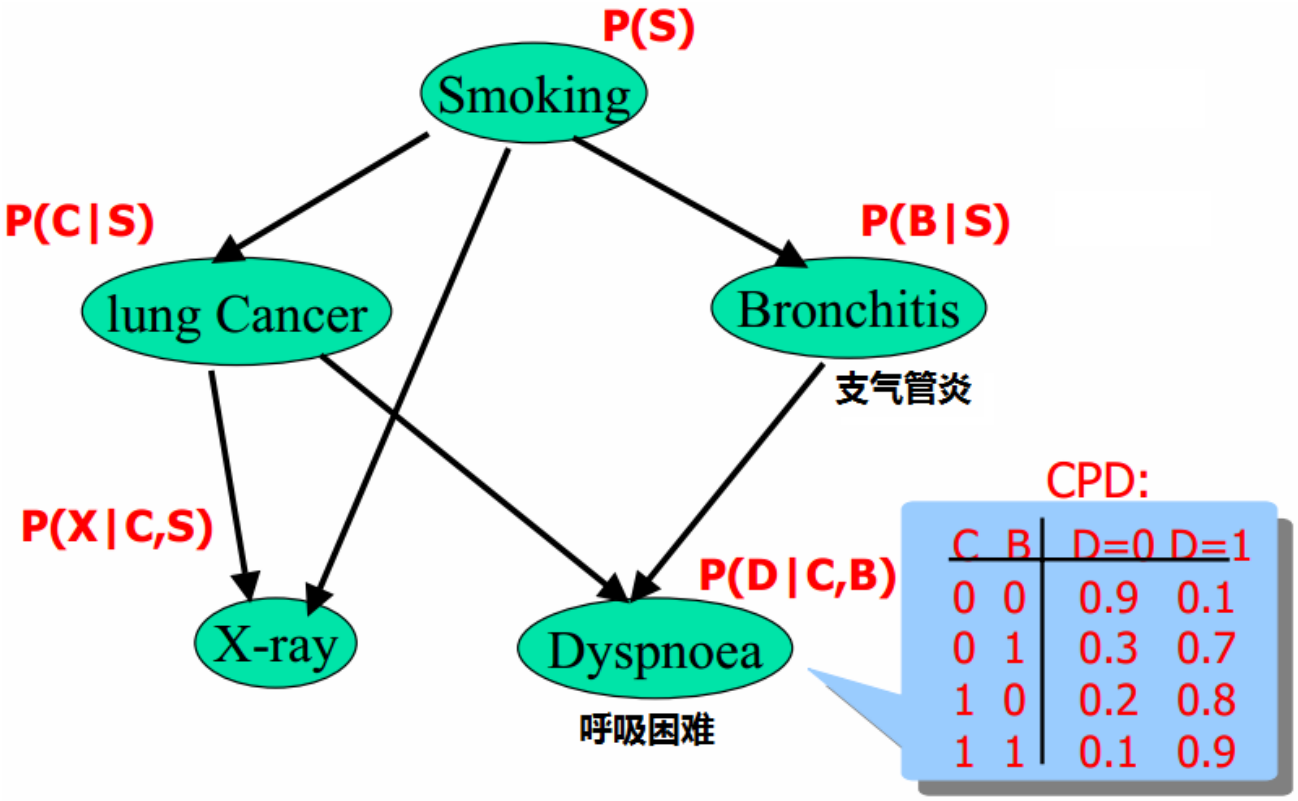
一个“正常”的贝叶斯网络

- 有些边缺失
- 直观上：
  - x1和x2独立
  - x6和x7在x4给定的条件下独立
- x1,x2,...x7的联合分布：



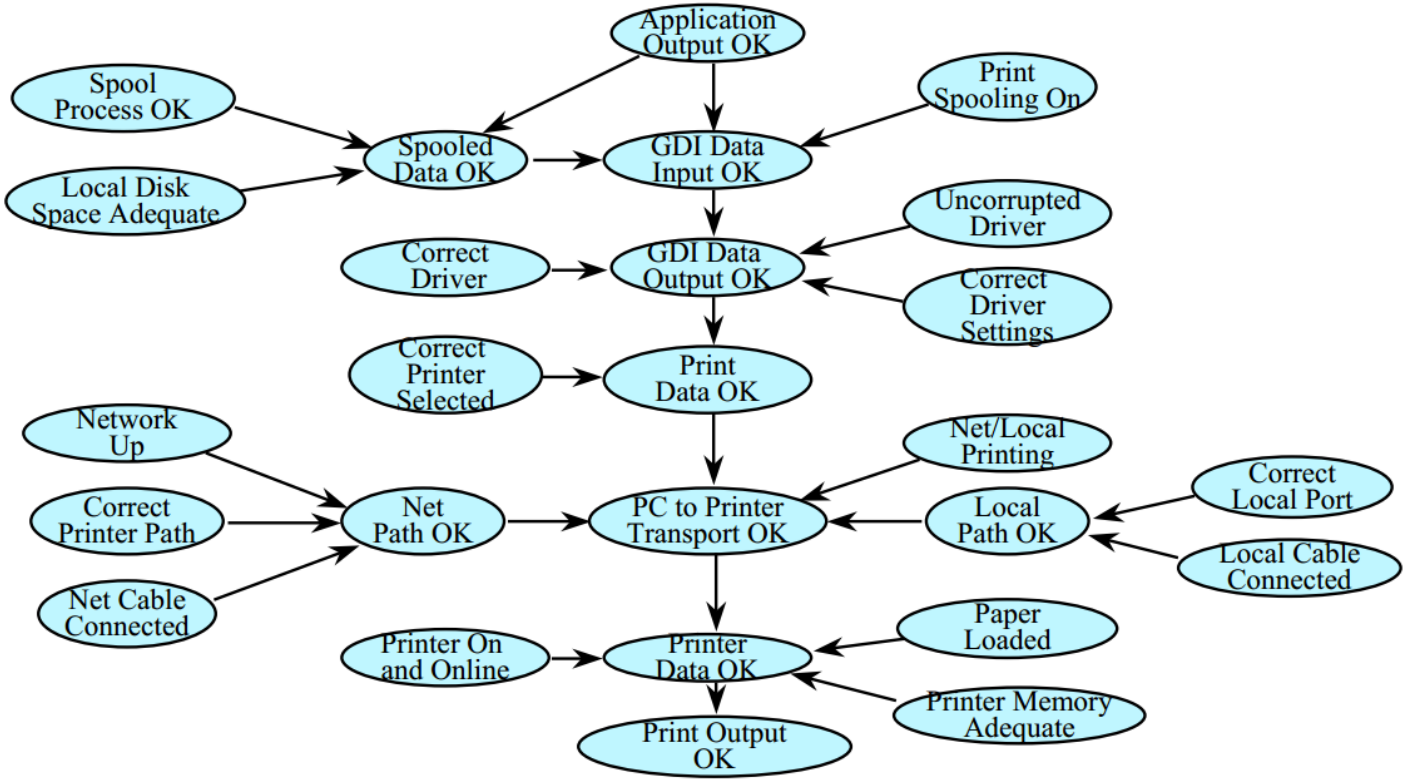
$$p(x_1)p(x_2)p(x_3)p(x_4|x_1,x_2,x_3)p(x_5|x_1,x_3)p(x_6|x_4)p(x_7|x_4,x_5)$$

对一个实际贝叶斯网络的分析



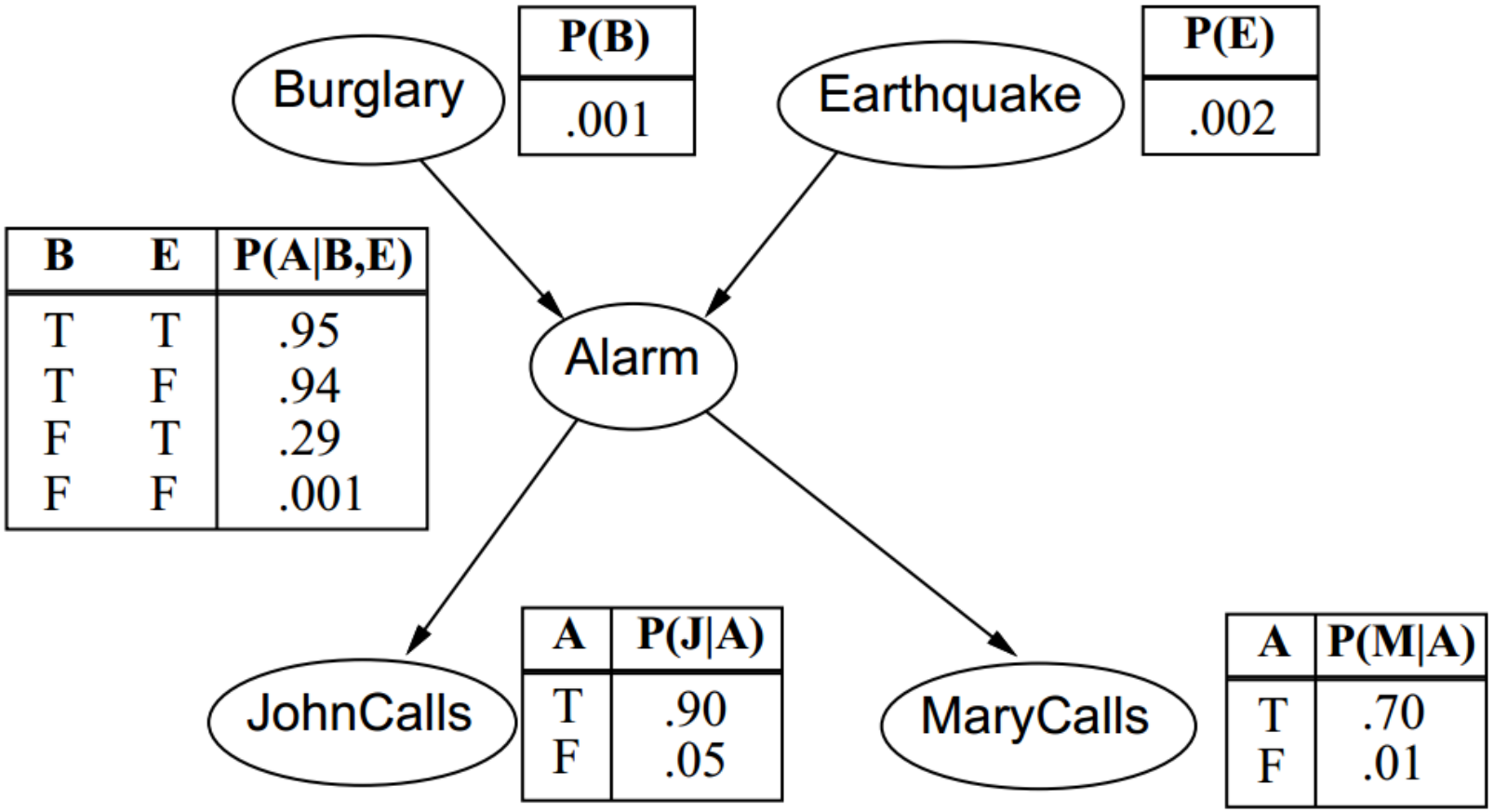
$1+2+2+4+4=13$ vs 2^5

贝叶斯网络：打印机故障诊断

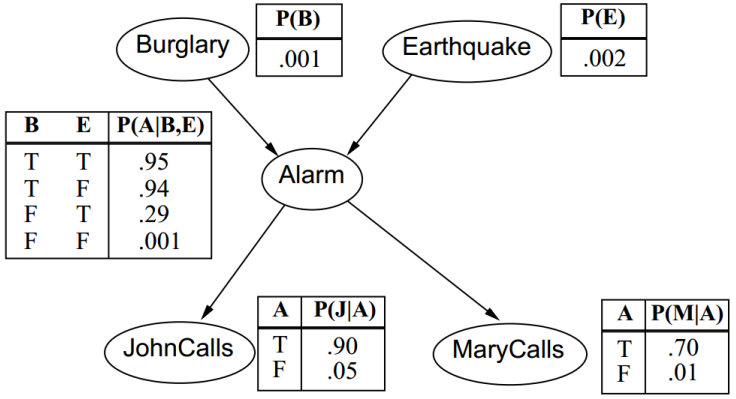


- $17*1 + 1*2 + 2*2^2 + 3*2^3 + 3*2^4 = 99$
- $2^{26} = 67108864$

贝叶斯网络：警报


Burglary
P(B)
.001
Earthquake
P(E)
.002
B E P(A|B,E)
T T .95
T F .94
F T .29
F F .001
Alarm
JohnCalls
A P(J|A)
T .90
F .05
MaryCalls
A P(M|A)
T .70
F .01

贝叶斯网络：警报



- 全部随机变量的联合分布

$$P(x_1, x_2, \cdots x_n) = \prod_{i=1}^{n} P(x_i \mid parents(x_i))$$

$$P(j, m, a, \bar{b}, \bar{e})$$

$$= P(j \mid a) P(m \mid a) P(a \mid \bar{b}, \bar{e}) P(\bar{b}) P(\bar{e})$$

$$= 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998$$

$$\approx 0.00063$$

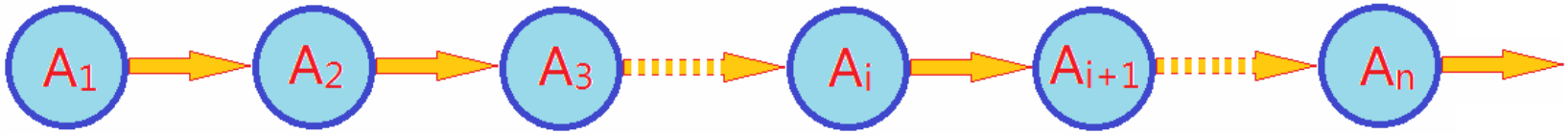
贝叶斯网络的形式化定义

- BN(G, Θ)
  - G:有向无环图
  - G的结点：随机变量
  - G的边：结点间的有向依赖
  - Θ：所有条件概率分布的参数集合
  - 结点X的条件概率：P(X|parent(X))

$$P(S, C, B, X, D) = P(S)\,P(C|S)\,P(B|S)\,P(X|C,S)\,P(D|C,B)$$

- 思考：需要多少参数才能确定上述网络呢？
  - 每个结点所需参数的个数：结点的parent数目是M，结点和parent的可取值数目都是K：$K^M*(K-1)$
    - 考察结点的parent对该结点形成了多少种情况(条件分布)

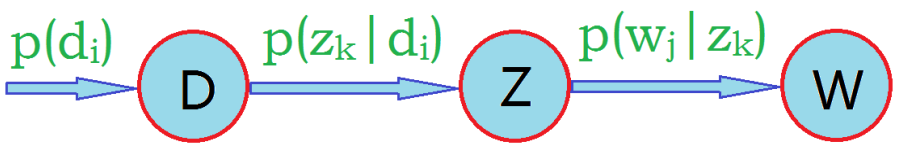
特殊的贝叶斯网络



- 结点形成一条链式网络，称作马尔科夫模型
  - A_{i+1}只与A_i有关，与$A_1,\ldots,A_{i-1}$无关
- 思考：
  - 伪随机数发生器

```
return(((holdrand = holdrand * 214013L + 2531011L) >> 16) & 0x7fff);
```

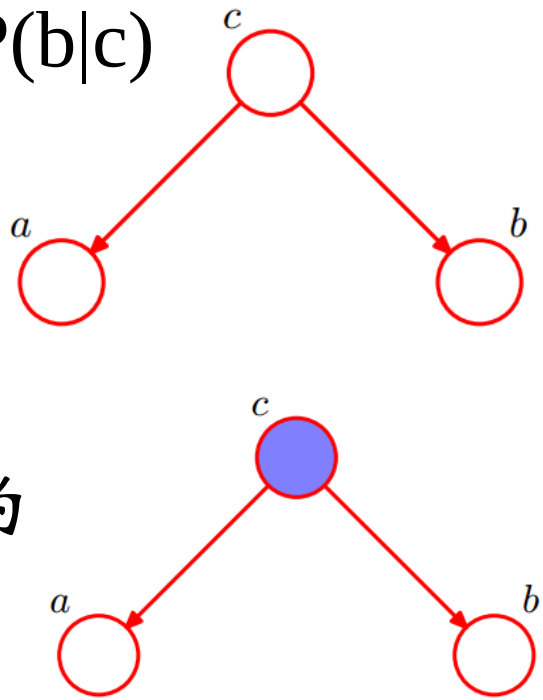
  - pLSA主题模型



通过贝叶斯网络判定条件独立—1

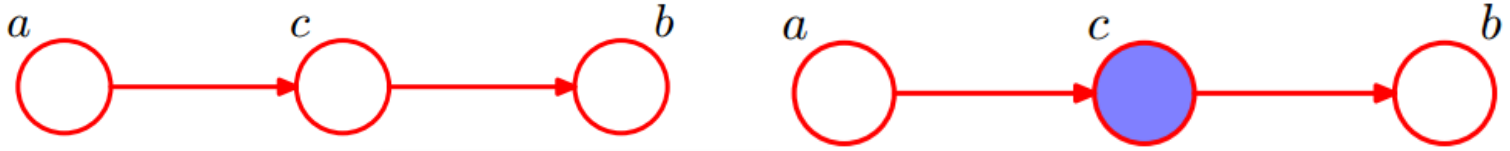
- 根据图模型，得：P(a,b,c)=P(c)*P(a|c)*P(b|c)
- 从而： P(a,b,c)/P(c)= P(a|c)*P(b|c)
- 因为 P(a,b|c)=P(a,b,c)/P(c)
- 得：P(a,b|c)=P(a|c)*P(b|c)
- 即：在c给定的条件下，

  a，b被阻断(blocked)是独立的
  - 条件独立：tail-to-tail



通过贝叶斯网络判定条件独立—2

- P(a,b,c)=P(a)*P(c|a)*P(b|c)



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

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

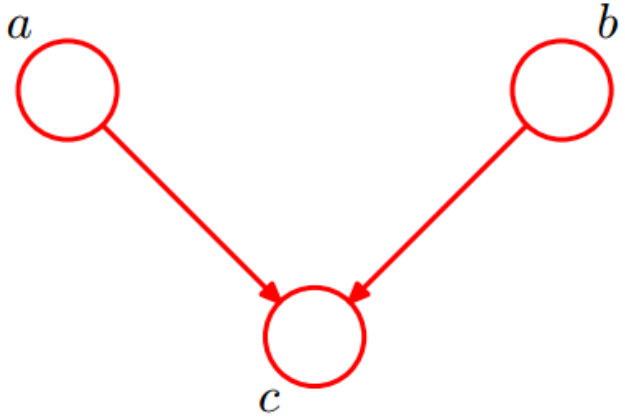
通过贝叶斯网络判定条件独立—3

- $$P(a,b,c) = P(a)*P(b)*P(c|a,b)$$

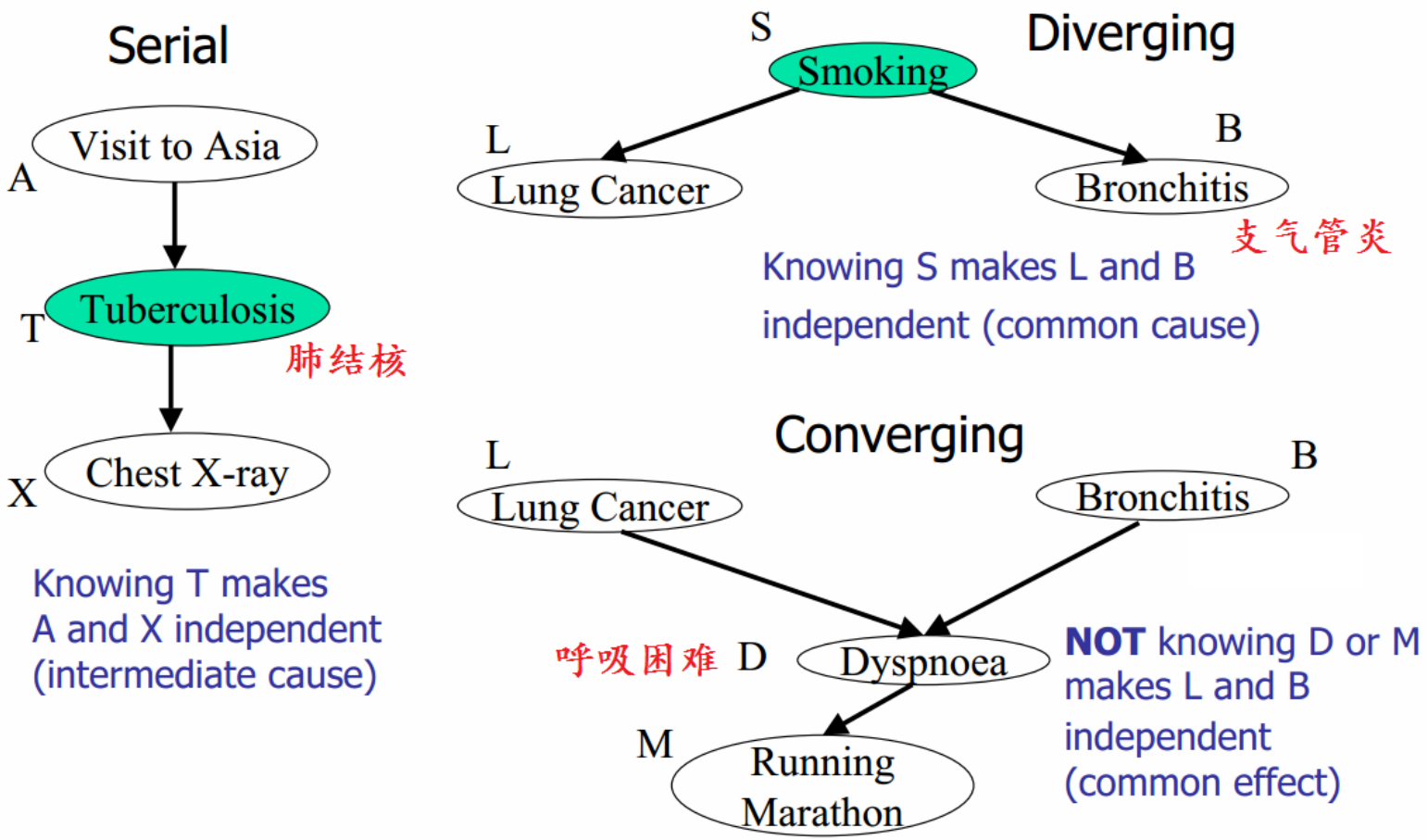
  $$\sum_c P(a,b,c) = \sum_c P(a)*P(b)*P(c|a,b)$$

  $$\Rightarrow P(a,b) = P(a)*P(b)$$

- 在c未知的条件下，a，b被阻断(blocked)，是独立的： head-to-head



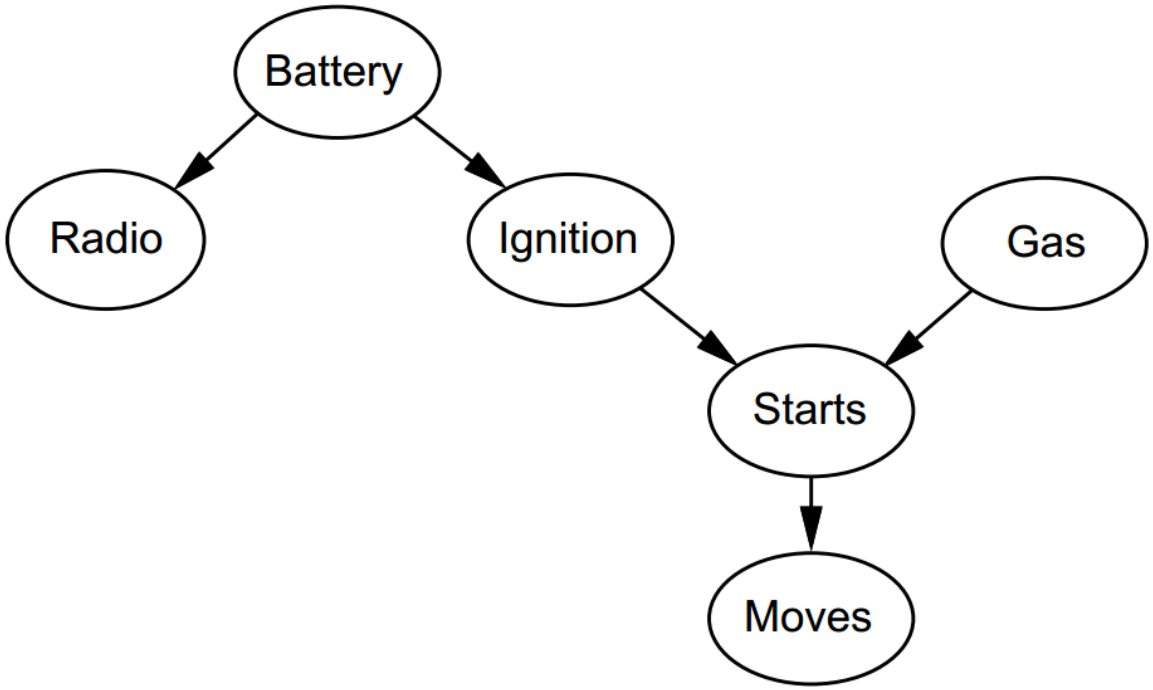
举例说明这三种情况


Serial
A
Visit to Asia
T
Tuberculosis
肺结核
X
Chest X-ray
Knowing T makes
A and X independent
(intermediate cause)
S
Smoking
Diverging
L
Lung Cancer
B
Bronchitis
支气管炎
Knowing S makes L and B
independent (common cause)
Converging
L
Lung Cancer
B
Bronchitis
呼吸困难 D
Dyspnoea
M
Running
Marathon
NOT knowing D or M
makes L and B
independent
(common effect)

将上述结点推广到结点集

- D-separation：有向分离
- 对于任意的结点集A，B，C，考察所有通过A中任意结点到B中任意结点的路径，若要求A，B条件独立，则需要所有的路径都被阻断(blocked)，即满足下列两个前提之一：
  - A和B的“head-to-tail型”和“tail-to-tail型”路径都通过C；
  - A和B的“head-to-head型”路径不通过C以及C的子孙；
- 如果A,B不满足D-separation，A,B有时被称为D-connected.

有向分离的举例



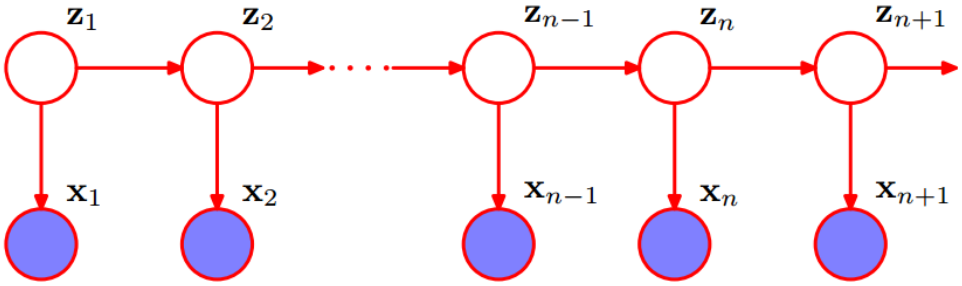
- Gas和Radio是独立的吗？给定Battery呢？Ignition呢？Starts呢？Moves呢？(答：IIDD)

再次分析链式网络

- 有D-separation可知，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模型，叫做马尔科夫模型。

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

HMM



- 隐马尔科夫模型(HMM, Hidden Markov Model)可用标注问题，在语音识别、NLP、生物信息、模式识别等领域被实践证明是有效的算法。
- HMM是关于时序的概率模型，描述由一个隐藏的马尔科夫链随机生成不可观测的状态随机序列，再由各个状态生成一个观测而产生观测随机序列的过程。
- 隐马尔科夫模型随机生成的状态的序列，称为状态序列；每个状态生成一个观测，由此产生的观测随机序列，称为观测序列
  - 序列的每个位置可看做是一个时刻
  - 空间序列也可以使用该模型：如分析DNA。

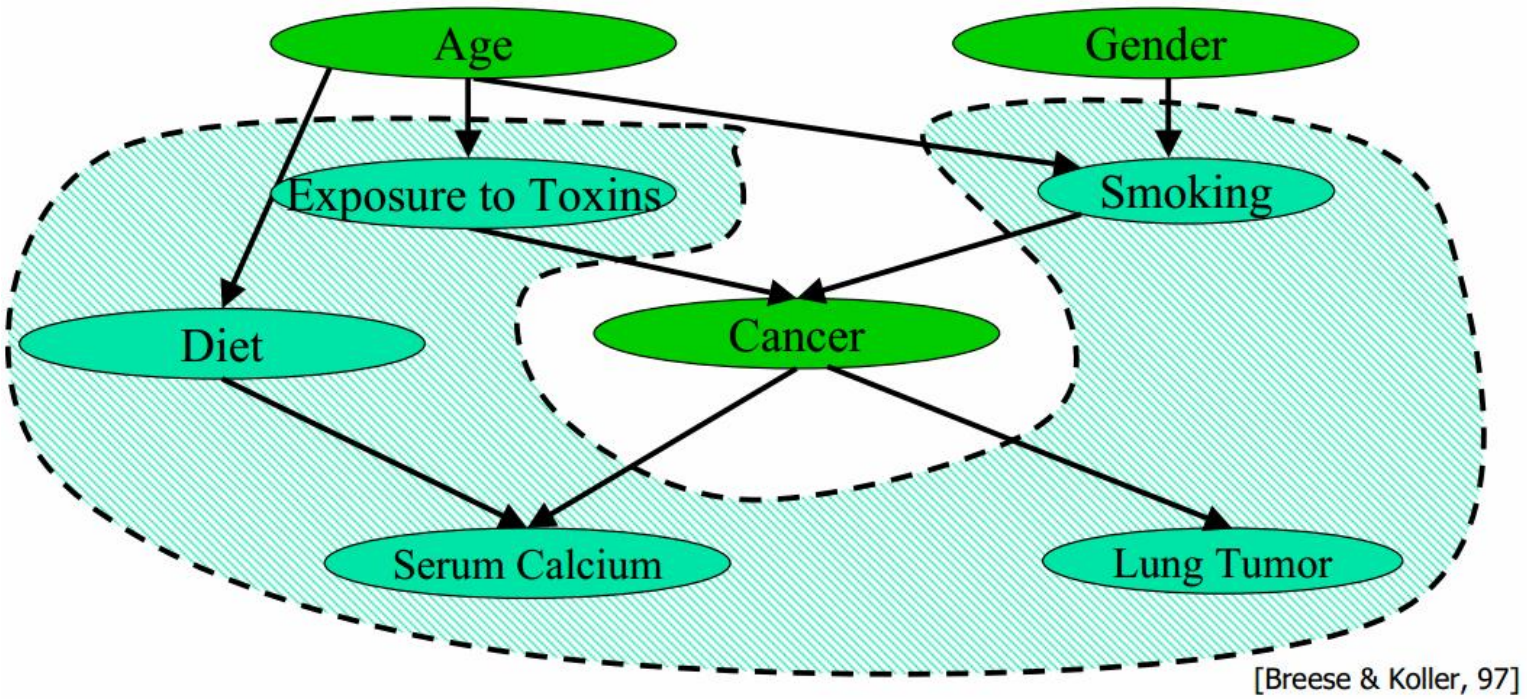
Markov Blanket

- 一个结点的Markov Blanket是一个集合，在这个集合中的结点都给定的条件下，该结点条件独立于其他所有结点。
- 即：一个结点的Markov Blanket是它的parents,children以及spouses(孩子的其他parent)

贝叶斯网络

背景知识：Serum Calcium(血清钙浓度)高于2.75mmol/L即为高钙血症。许多恶性肿瘤可并发高钙血症。恶性肿瘤病人离子钙增高的百分比大于总钙，也许可用于肿瘤的过筛试验。当高钙血症的原因难于确定时，必须考虑到恶性肿瘤的存在。

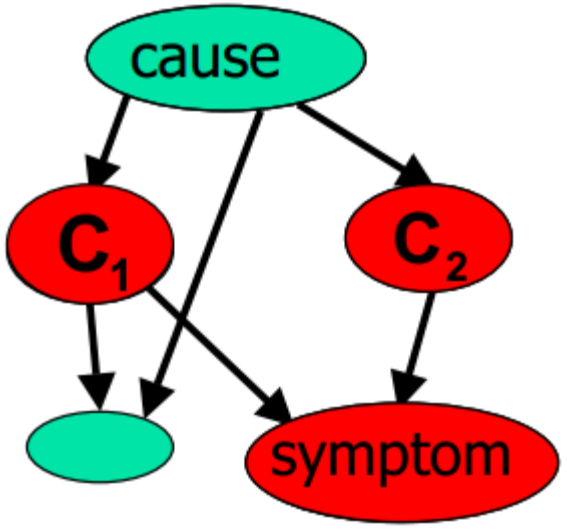
http://www.wiki8.com/xueqinggai_131584/



阴影部分的结点集合，是Cancer的“马尔科夫毯”(Markov Blanket)

条件独立：$P(S,L|C) = P(S|C) * P(L|C)$

贝叶斯网络的用途



- 诊断：P(病因|症状)
- 预测：P(症状|病因)
- 分类：$\max_{class} P(类别|数据)$

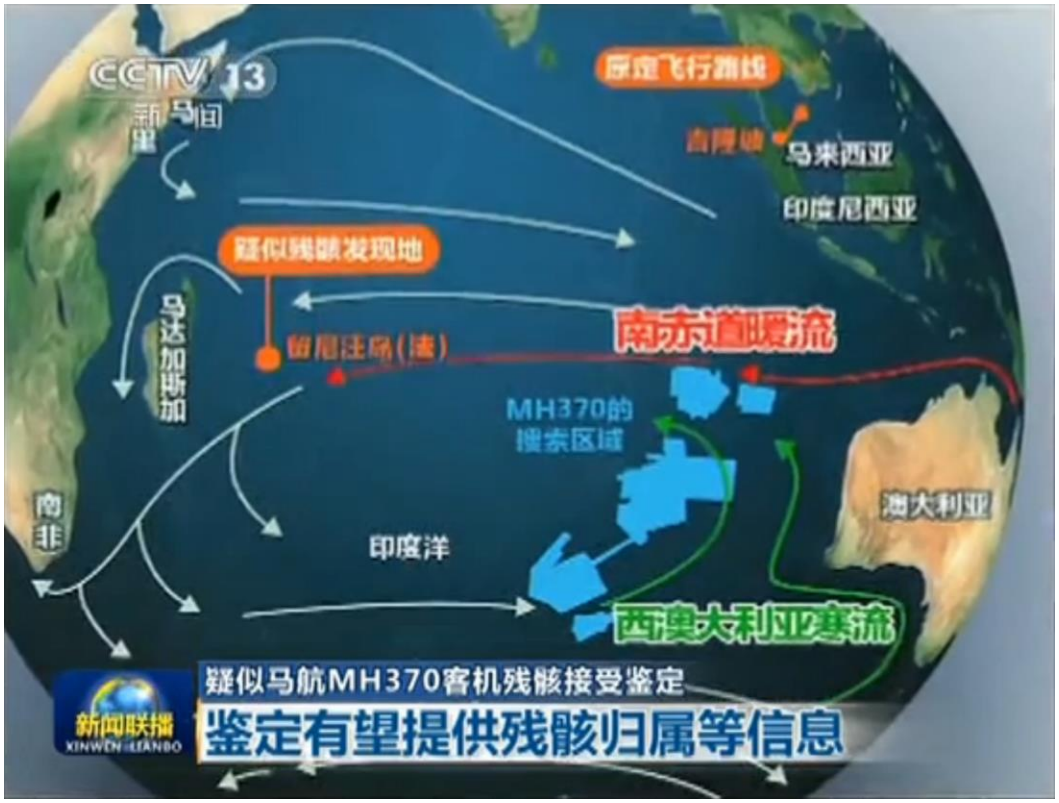
- 通过给定的样本数据，建立贝叶斯网络的拓扑结构和结点的条件概率分布参数。这往往需要借助先验知识和极大似然估计来完成。
- 在贝叶斯网络确定的结点拓扑结构和条件概率分布的前提下，可以使用该网络，对未知数据计算条件概率或后验概率，从而达到诊断、预测或者分类的目的。

寻找马航MH370

- 2014年3月8日，马来西亚航空公司MH370航班(波音777-200ER)客机凌晨0:41分从吉隆坡飞往北京；凌晨1:19分，马航MH370与空管失去联系。凌晨2:14分飞机最后一次出现在军事雷达上之后人间消失。
- 2015年7月29日在法属留尼汪岛(l'île de la Reunion)发现襟副翼残骸；2015年8月6日，马来西亚宣布，该残骸确属马航MH370。随后法国谨慎宣布，“有很强的理由推测认为，...残骸属于马航MH370航班的波音777客机...但最终的比对结果还需要进一步的技术验证加以确认。”

MH370最后消失区域

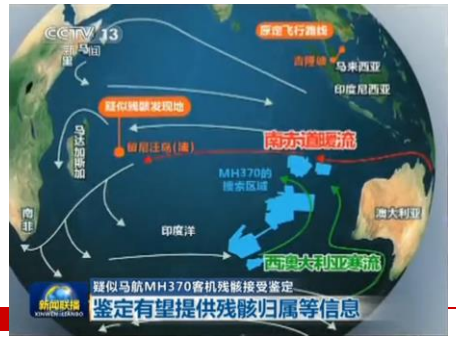
- 可否根据雷达最后消失区域和洋流、大气等因素；
- 判断留尼汪岛是否位于可能区域？
- 残骸漂流到该岛屿的概率有多大？

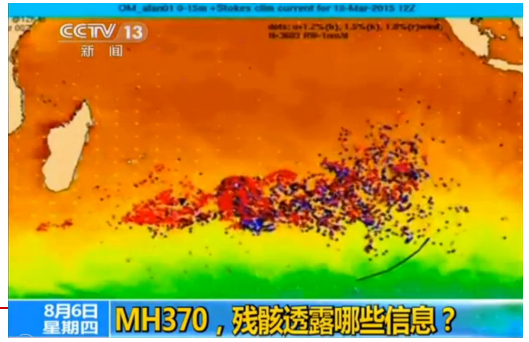


建模分析

- 记MH370最后消失区域坐标为(x_0,y_0)，经过短暂时间后，如何推断它的可能区域？
- 以(x_0,y_0)为圆心、以x轴正向为0°，逆时针为旋转角正向建立坐标系，将旋转角按照0.1°为步长，分成3600份。经过时间Δt后，残骸落在哪个区域？
- 根据洋流大气等综合因素，判断在(x_0,y_0)移动方向d的概率为p_d。$d\in[0,360]$，步长为0.1；
- 得到3600维的初始行向量π。
  - 移动速度V_0*微小时间段Δt，得到位于新位置(x_1,y_1)的概率。这些新位置的外包围盒构成区域R_1。

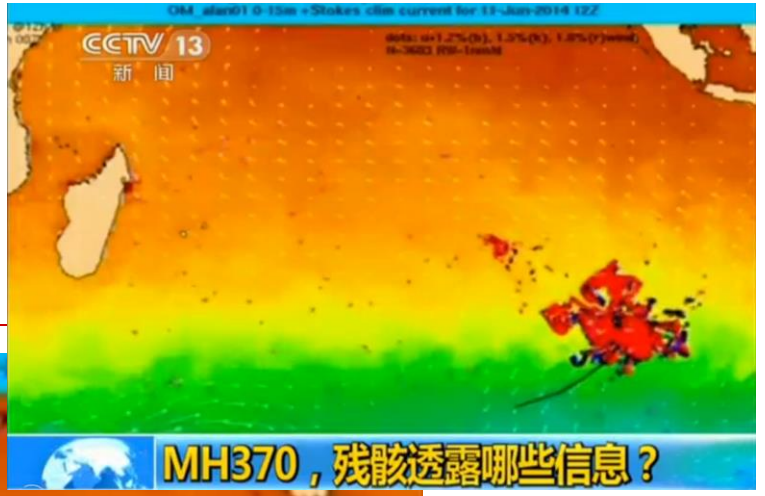
转移概率矩阵

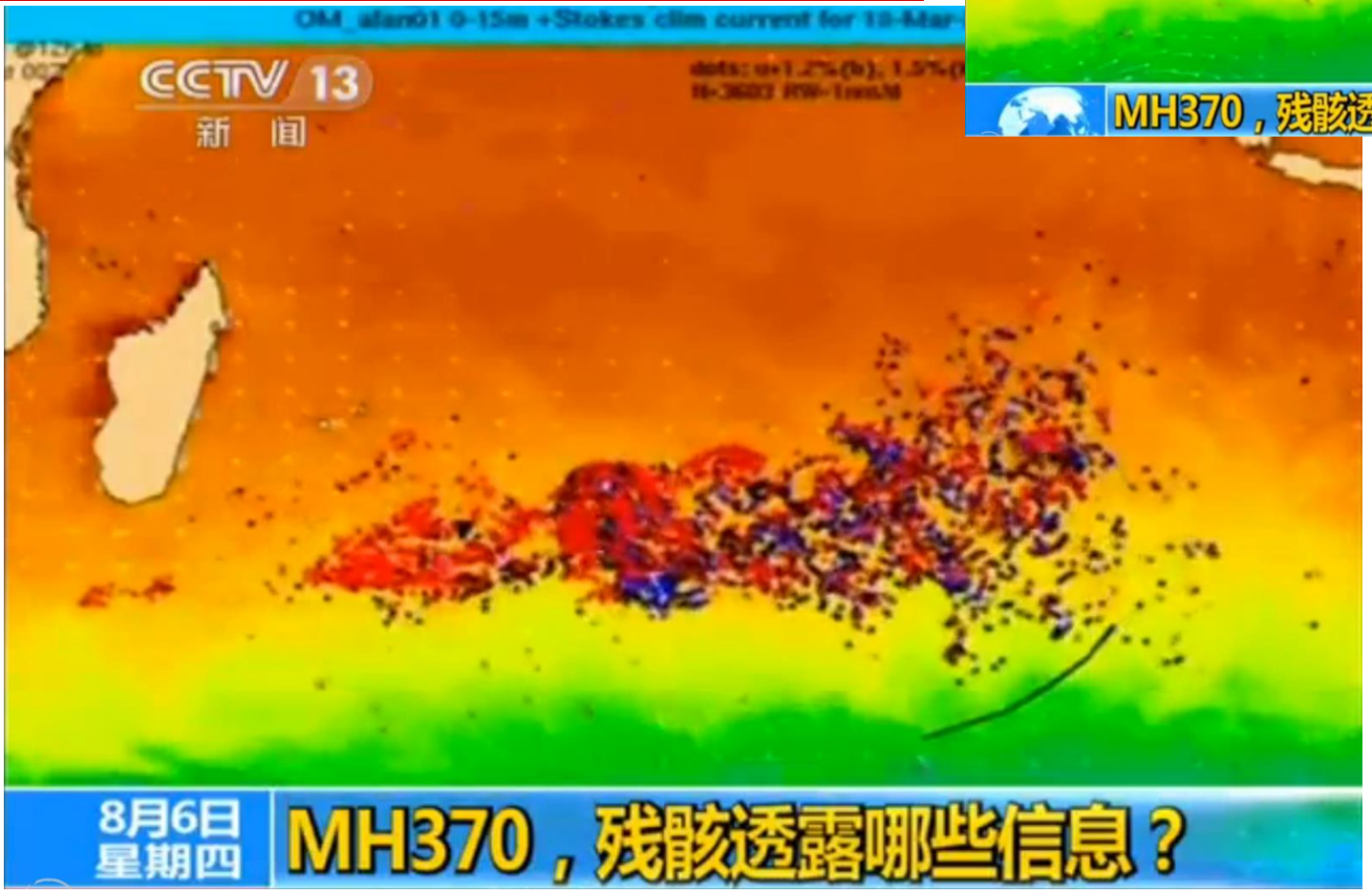




- 建模任务：在某个时刻t_i，物体位于区域R_i，计算下一时刻位于R_{i+1}的概率。
- 遍历R_i上所有概率非零的坐标点，用$Vi*\Delta t$确定R_{i+1}的先验范围。将R_i增广到R_{i+1}上。
  - 将二维区域R_i和R_{i+1}拉直成行向量，假定维度为N
- 根据领域知识建立R_i上所有概率非零点的概率P_i。
  - P_i表示区域R_i到R_{i+1}的转移概率，因此，是N×N的二维矩阵(方阵)，
- 区域R_{i+1}的后验概率为：$R_{i+1}=R_i \times P_i$
  - 以上述方法，计算每个时刻的可能区域$R_0,R_1,\ldots R_n$

MH370可能区域





总结

- 在每个时刻，物体的当前可能区域是上一时刻所有可能区域和相应转移概率的乘积和，这恰好是矩阵乘法(矩阵和向量乘法)的定义。
- 当前可能区域只和上一个时刻的区域有关，而与更上一个时刻无关，因此，是马尔科夫模型。
- 思考：可以使用“漂流位置”建立马尔科夫模型，该可能位置是不可观察的，而将“转移位置”认为是“漂流位置”的转换结果，“转移位置”是残骸的最终真实位置，使用增强的隐马尔科夫模型。
  - 不要过得累加模型的复杂度，适时使用奥卡姆剃刀(Occam‘s Razor)。
  - 该模型仅个人观点。

贝叶斯网络的构建

- 依次计算每个变量的D-separation的局部测试结果，综合每个结点得到贝叶斯网络。
- 算法过程：
  - 选择变量的一个合理顺序：$X_1,X_2,...X_n$
  - 对于i=1到n
    - 在网络中添加X_i结点
    - 在$X_1,X_2,...X_{i-1}$中选择X_i的父母，使得：

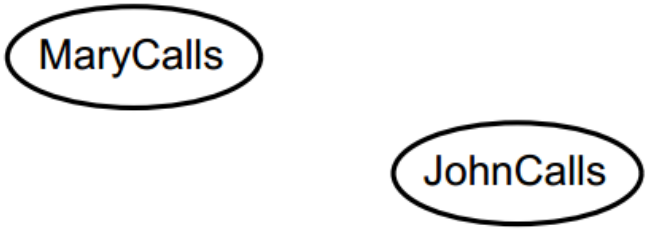
$$P\left(X_i \middle| Parent\left(X_i\right)\right) = P\left(X_i \middle| X_1, X_2 \cdots X_{i-1}\right)$$

- 这种构造方法，显然保证了全局的语义要求：

$$P\left(X_1, X_2 \cdots X_n\right) = \prod_{i=1}^{n} P\left(X_i \middle| X_1, X_2 \cdots X_{i-1}\right) = \prod_{i=1}^{n} P\left(X_i \middle| Parent\left(X_i\right)\right)$$

贝叶斯网络的构建举例

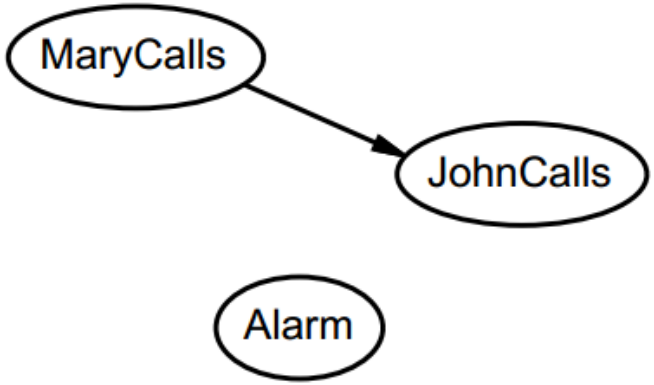
Suppose we choose the ordering M, J, A, B, E



$P(J|M) = P(J)$?

构建举例

Suppose we choose the ordering M, J, A, B, E

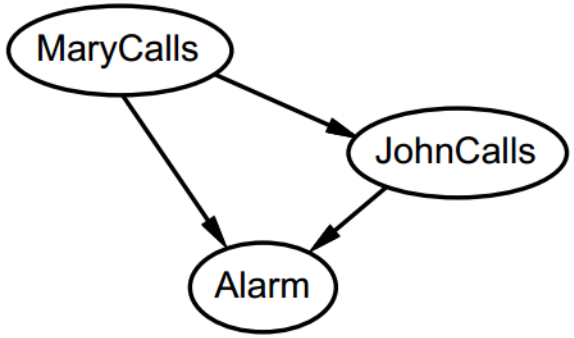


$P(J|M) = P(J)$? No

$P(A|J, M) = P(A|J)$? $P(A|J, M) = P(A)$?

构建举例

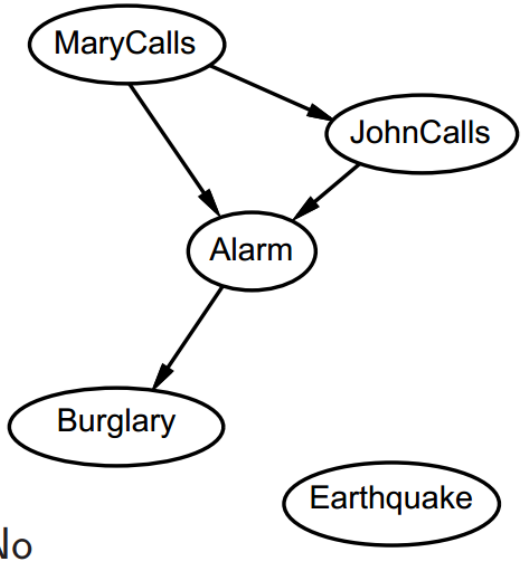
Suppose we choose the ordering M, J, A, B, E



$P(J|M) = P(J)$? No
$P(A|J, M) = P(A|J)$? $P(A|J, M) = P(A)$? No
$P(B|A, J, M) = P(B|A)$?
$P(B|A, J, M) = P(B)$?

构建举例

Suppose we choose the ordering M, J, A, B, E



$P(J|M) = P(J)$? No

$P(A|J, M) = P(A|J)$? $P(A|J, M) = P(A)$? No

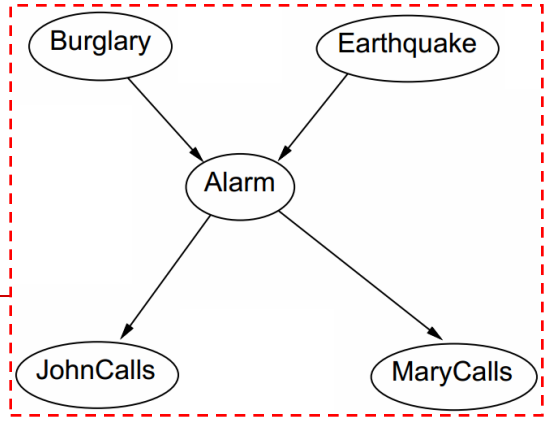
$P(B|A, J, M) = P(B|A)$? Yes

$P(B|A, J, M) = P(B)$? No

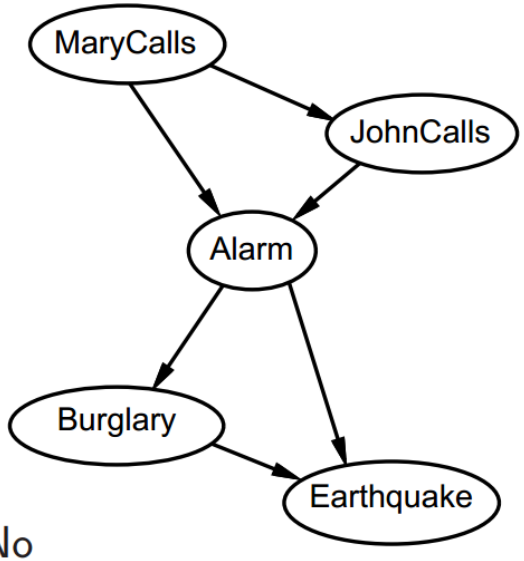
$P(E|B, A, J, M) = P(E|A)$?

$P(E|B, A, J, M) = P(E|A, B)$?

构建举例



Suppose we choose the ordering M, J, A, B, E

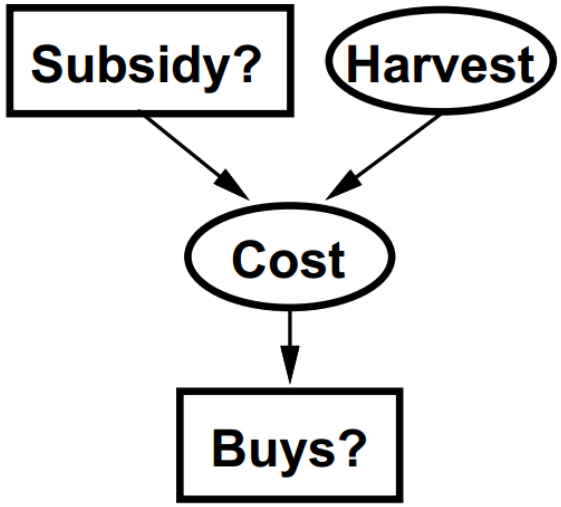


$P(J|M) = P(J)$? No
$P(A|J,M) = P(A|J)$? $P(A|J,M) = P(A)$? No
$P(B|A,J,M) = P(B|A)$? Yes
$P(B|A,J,M) = P(B)$? No
$P(E|B,A,J,M) = P(E|A)$? No
$P(E|B,A,J,M) = P(E|A,B)$? Yes

1+2+4+2+4=13

混合(离散+连续)网络

Discrete ($Subsidy?$ and $Buys?$); continuous ($Harvest$ and $Cost$)



Option 1: discretization—possibly large errors, large CPTs
Option 2: finitely parameterized canonical families

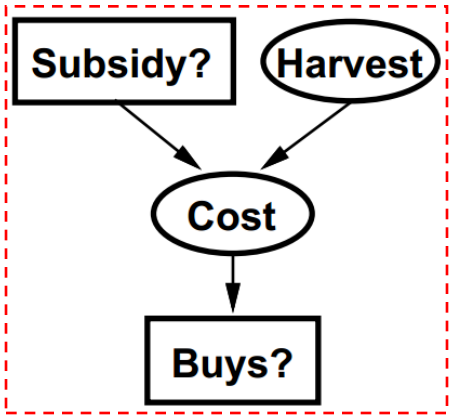
1) Continuous variable, discrete+continuous parents (e.g., $Cost$)
2) Discrete variable, continuous parents (e.g., $Buys?$)

孩子结点是连续的

Need one conditional density function for child variable given continuous parents, for each possible assignment to discrete parents

Most common is the linear Gaussian model, e.g.,:

$$
\begin{aligned}
&P(Cost = c | Harvest = h, Subsidy? = true) \\
&= N(a_t h + b_t, \sigma_t)(c) \\
&= \frac{1}{\sigma_t \sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{c - (a_t h + b_t)}{\sigma_t}\right)^2\right)
\end{aligned}
$$

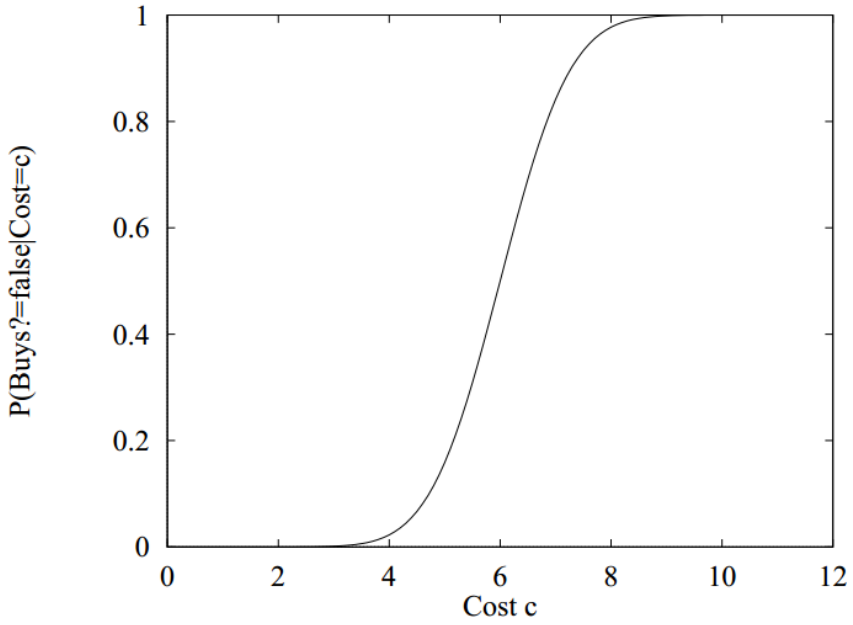


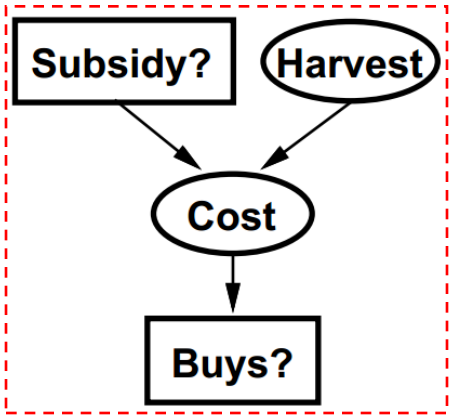
Mean $Cost$ varies linearly with $Harvest$, variance is fixed

Linear variation is unreasonable over the full range
but works OK if the **likely** range of $Harvest$ is narrow

孩子结点是离散的，父节点是连续的

Probability of $Buys?$ given $Cost$ should be a "soft" threshold:





Probit distribution uses integral of Gaussian:

$\Phi(x) = \int_{-\infty}^{x} N(0,1)(x)dx$

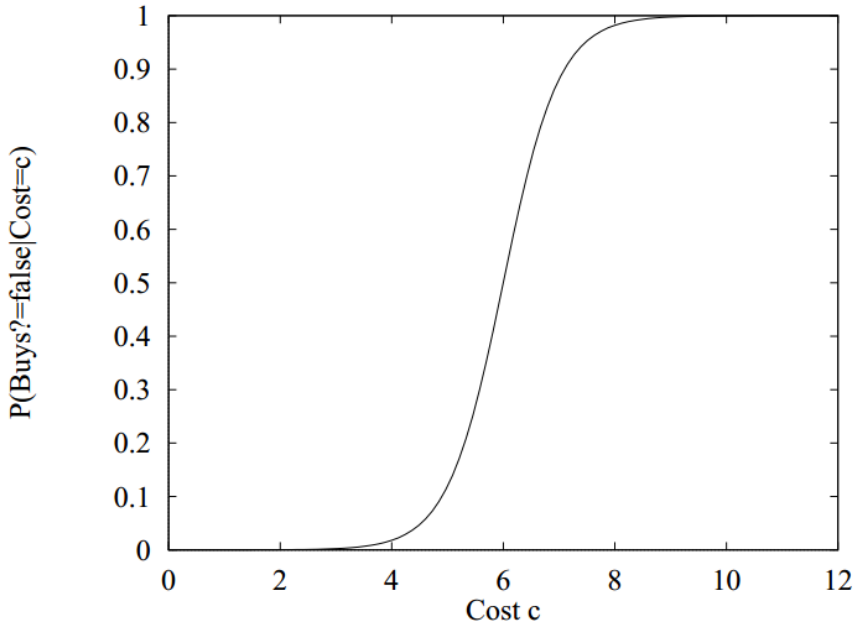
$P(Buys? = true \mid Cost = c) = \Phi((-c + \mu)/\sigma)$

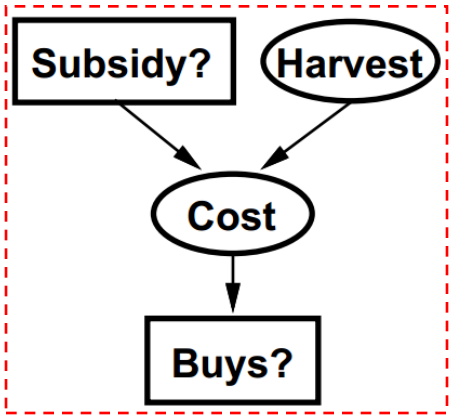
孩子结点是离散的，父节点是连续的

Sigmoid (or logit) distribution also used in neural networks:

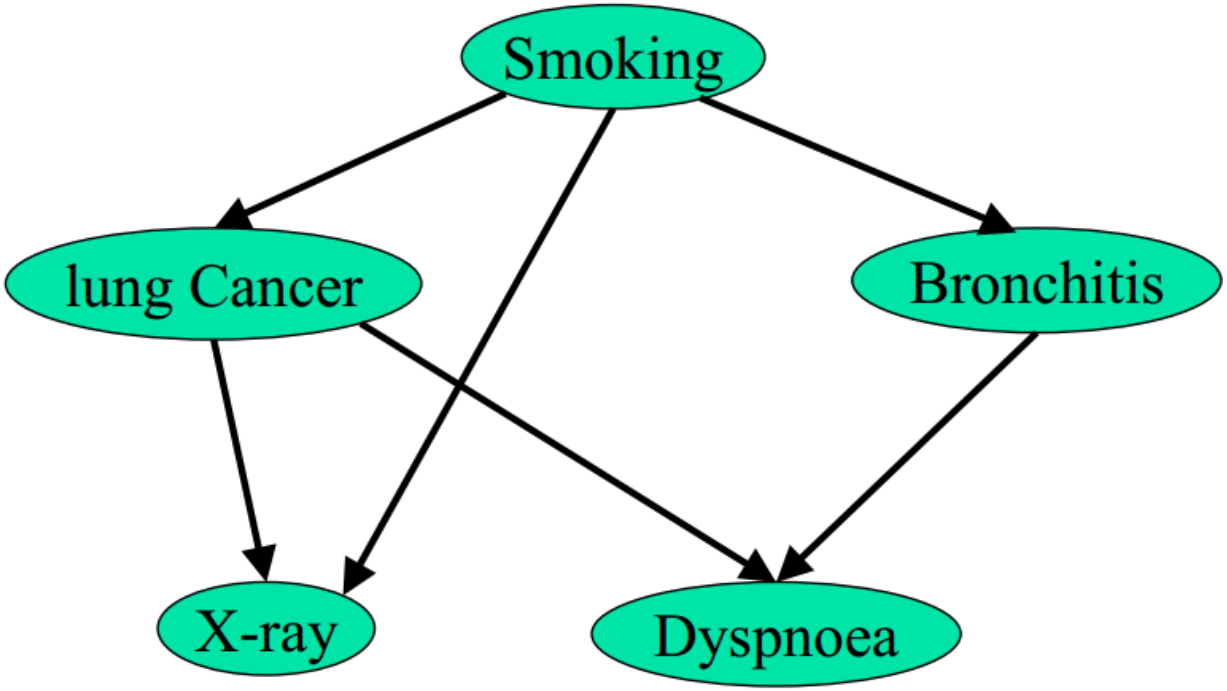
$$P(Buys? = true \mid Cost = c) = \frac{1}{1 + exp(-2\frac{-c+\mu}{\sigma})}$$

Sigmoid has similar shape to probit but much longer tails:



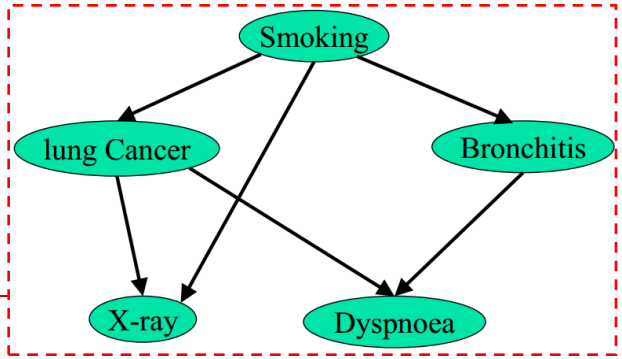


贝叶斯网络的推导



P (smoking| dyspnoea=yes) = ?

贝叶斯网络的推导



$$P(s|d=1) = \frac{P(s,d=1)}{P(d=1)} \propto P(s,d=1) =$$

$$\sum_{d=1,b,x,c} P(s)P(c|s)P(b|s)P(x|c,s)P(d|c,b) =$$

$$P(s)\sum_{d=1}\sum_{b}P(b|s)\sum_{x}\sum_{c}P(c|s)P(x|c,s)P(d|c,b)$$

Variable Elimination

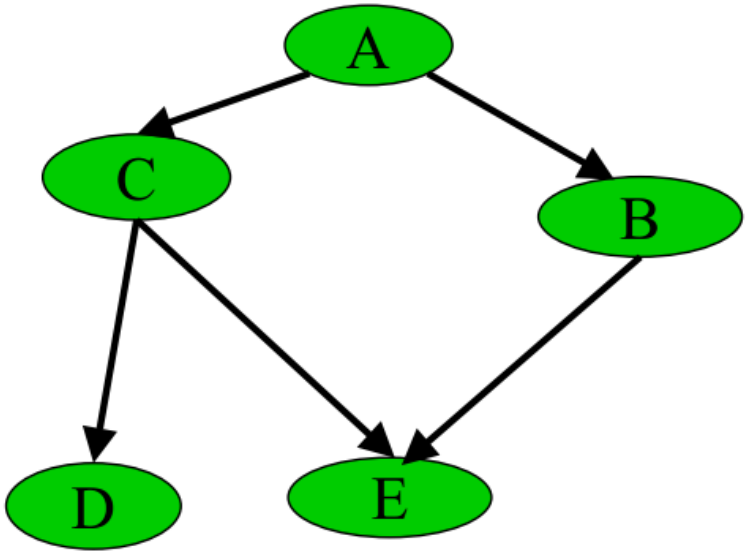
$$f(s,d,b,x)$$

无向环

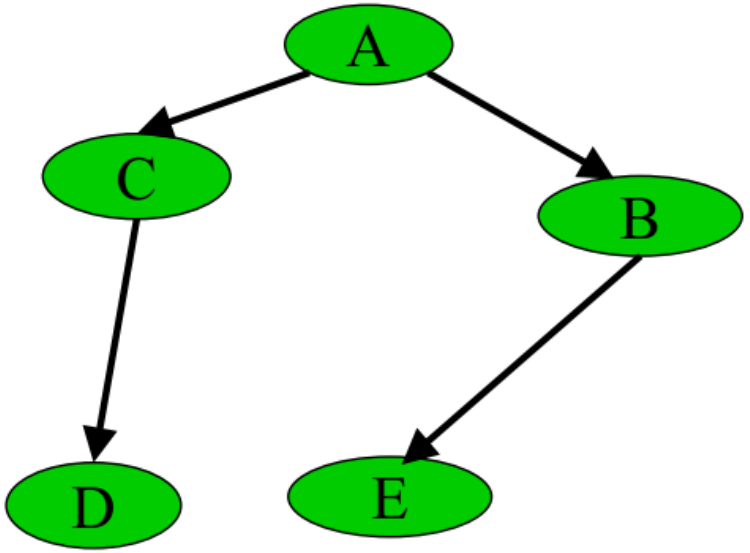
- 可以发现，若贝叶斯网络中存在“环”(无向)，则因此构造的因子图会得到环。而使用消息传递的思想，这个消息将无限传输下去，不利于概率计算。
- 解决方法：
  - 删除贝叶斯网络中的若干条边，使得它不含有无向环
  - 重新构造没有环的贝叶斯网络

原贝叶斯网络的近似树结构

True distribution P(X)



Tree-approximation P′(X)



$$D(p \| q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = E_{p(x)} \log \frac{p(x)}{q(x)}$$

将两图的相对熵转换成变量的互信息

Theorem [Chow and Liu, 1968]

Given a joint PDF P(x), the KL-divergence D(P,P') is minimized by projecting P(x) on a *maximum-weight spanning tree (MSWT)* over nodes in X, where the weight on the edge (X_i, X_j) is defined by the mutual information measure

$$I(X_i; X_j) = \sum_{x_i, x_j} P(x_i, x_j) \log \frac{P(x_i, x_j)}{P(x_i) P(x_j)}$$

最大权生成树MSWT的建立过程

- 对于分布P(x)，对于所有的i≠j，计算联合分布P(Xi|Yj)；
- 使用第1步得到的概率分布，计算任意两个结点的互信息I(Xi,Yj)，并把I(Xi,Yj)作为这两个结点连接边的权值；
- 计算最大权生成树(Maximum-weight spanning tree)
  - a. 初始状态：n个变量(结点)，0条边
  - b. 插入最大权重的边
  - c. 找到下一个最大的边，并且加入到树中；要求加入后，没有环生成。否则，查找次大的边；
  - d. 重复上述过程c过程直到插入了n-1条边(树建立完成)
- 选择任意结点作为根，从根到叶子标识边的方向；

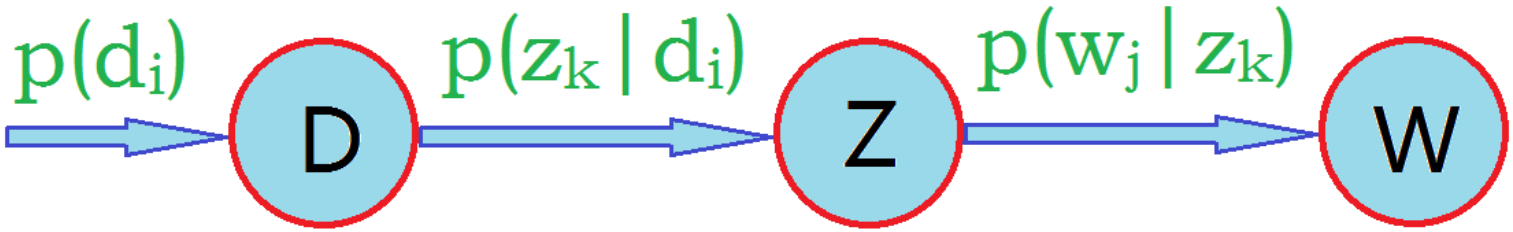
- 该生成树的近似联合概率P'(x)和原贝叶斯网络的联合概率P(x)的相对熵最小。

附：Chow-Liu算法

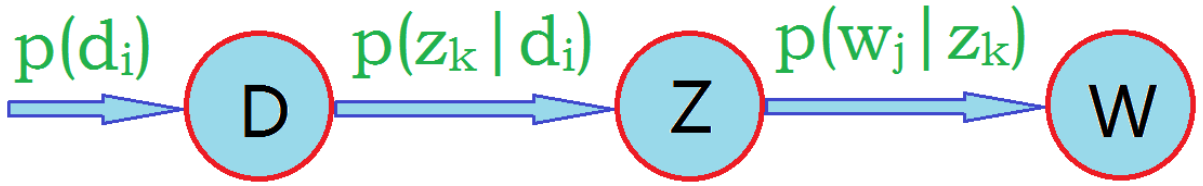
1. From the given distribution P(x) (or from data generated by P(x)), compute the joint distribution $P(x_i \mid x_j)$ for all $i \neq j$
2. Using the pairwise distributions from step 1, compute the mutual information $I(X_i; X_j)$ for each pair of nodes and assign it as the weight to the corresponding edge (X_i, X_j).
3. Compute the maximum-weight spanning tree (MSWT):
   a. Start from the empty tree over n variables
   b. Insert the two largest-weight edges
   c. Find the next largest-weight edge and add it to the tree if no cycle is formed; otherwise, discard the edge and repeat this step.
   d. Repeat step (c) until n-1 edges have been selected (a tree is constructed).
4. Select an arbitrary root node, and direct the edges outwards from the root.
5. Tree approximation P'(x) can be computed as a projection of P(x) on the resulting directed tree (using the product-form of P'(x)).

pLSA模型

- 基于概率统计的pLSA模型(probabilistic Latent Semantic Analysis，概率隐语义分析)，增加了主题模型，形成简单的贝叶斯网络，可以使用EM算法学习模型参数。

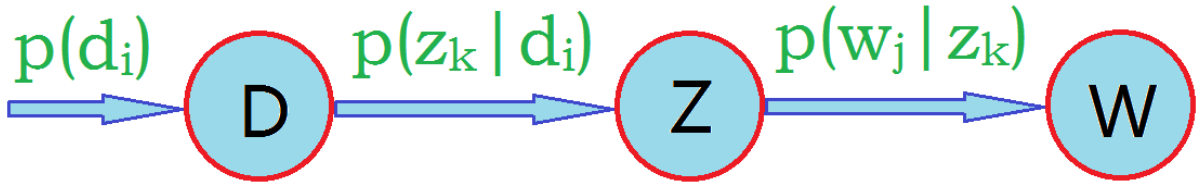


pLSA模型



- D代表文档，Z代表主题(隐含类别)，W代表单词；
  - P(d_i)表示文档d_i的出现概率，
  - P(z_k|d_i)表示文档d_i中主题z_k的出现概率，
  - P(w_j|z_k)表示给定主题z_k出现单词w_j的概率。
- 每个主题在所有词项上服从多项分布，每个文档在所有主题上服从多项分布。
- 整个文档的生成过程是这样的：
  - 以P(d_i)的概率选中文档d_i；
  - 以P(z_k|d_i)的概率选中主题z_k；
  - 以P(w_j|z_k)的概率产生一个单词w_j。

pLSA模型



- 观察数据为(d_i,w_j)对，主题z_k是隐含变量。
- (d_i,w_j)的联合分布为

$$P(d_i,w_j)=P(w_j\mid d_i)P(d_i)$$

$$P(w_j\mid d_i)=\sum_{k=1}^{K}P(w_j\mid z_k)P(z_k\mid d_i)$$

- 而$P(w_j\mid z_k),P(z_k\mid d_i)$对应了两组多项分布，而计算每个文档的主题分布，就是该模型的任务目标。

最大似然估计：w_j在d_i中出现的次数 $n(d_i, w_j)$

$$L = \prod_{i=1}^{N}\prod_{j=1}^{M} P(d_i, w_j) = \prod_i \prod_j P(d_i, w_j)^{n(d_i, w_j)}$$

$$l = \sum_i \sum_j n(d_i, w_j) \log P(d_i, w_j)$$

$$= \sum_i \sum_j n(d_i, w_j) \log P(w_j \mid d_i) P(d_i)$$

$$= \sum_i \sum_j n(d_i, w_j) \log \left(\sum_{k=1}^{K} P(w_j \mid z_k) P(z_k \mid d_i) \right) P(d_i)$$

$$= \sum_i \sum_j n(d_i, w_j) \log \left(\sum_{k=1}^{K} P(w_j \mid z_k) P(z_k \mid d_i) P(d_i) \right)$$

$$P(d_i, w_j) = P(w_j \mid d_i) P(d_i)$$

$$P(w_j \mid d_i) = \sum_{k=1}^{K} P(w_j \mid z_k) P(z_k \mid d_i)$$

目标函数分析

- 观察数据为(d_i,w_j)对，主题z_k是隐含变量。
- 目标函数 $$l=\sum_i\sum_j n(d_i,w_j)\log\left(\sum_{k=1}^{K}P(w_j\mid z_k)P(z_k\mid d_i)P(d_i)\right)$$
- 未知变量/自变量 $P(w_j\mid z_k),P(z_k\mid d_i)$
- 使用逐次逼近的办法：
  - 假定P(z_k|d_i)、P(w_j|z_k)已知，求隐含变量z_k的后验概率；
  - 在(d_i,w_j,z_k)已知的前提下，求关于参数P(z_k|d_i)、P(w_j|z_k)的似然函数期望的最大值，得到最优解P(z_k|d_i)、P(w_j|z_k)，带入上一步，从而循环迭代；
  - 即：EM算法。

求隐含变量主题z_k的后验概率

- 假定$P(z_k|d_i)$、$P(w_j|z_k)$已知，求隐含变量z_k的后验概率；

$$P(z_k \mid d_i, w_j) = \frac{P(w_j \mid z_k)P(z_k \mid d_i)}{\sum_{l=1}^{K} P(w_j \mid z_l)P(z_l \mid d_i)}$$

- 在(d_i, w_j, z_k)已知的前提下，求关于参数$P(z_k|d_i)$、$P(w_j|z_k)$的似然函数期望的最大值，得到最优解$P(z_k|d_i)$、$P(w_j|z_k)$，带入上一步，从而循环迭代；

分析似然函数期望

- 在(d_i,w_j,z_k)已知的前提下，求关于参数$P(z_k|d_i)$、$P(w_j|z_k)$的似然函数期望的最大值，得到最优解$P(z_k|d_i)$、$P(w_j|z_k)$，带入上一步，从而循环迭代；

关于参数$P(z_k|d_i)P(w_j|z_k)$的似然函数期望

$$l = \sum_i \sum_j n(d_i, w_j) \log P(d_i, w_j)$$

$$= \sum_i \sum_j n(d_i, w_j) \log(P(w_j \mid d_i) P(d_i))$$

$$= \sum_i \sum_j n(d_i, w_j) (\log P(w_j \mid d_i) + \log P(d_i))$$

$$= \left(\sum_i \sum_j n(d_i, w_j) \log P(w_j \mid d_i) \right) + \left(\sum_i \sum_j n(d_i, w_j) \log P(d_i) \right)$$

$$\Rightarrow$$

$$l_{new} = \left(\sum_i \sum_j n(d_i, w_j) \log P(w_j \mid d_i) \right)$$

$$E(l_{new}) = \sum_i \sum_j n(d_i, w_j) \sum_{k=1}^{K} P(z_k \mid d_i, w_j) \log P(w_j, z_k \mid d_i)$$

$$= \sum_i \sum_j n(d_i, w_j) \sum_{k=1}^{K} P(z_k \mid d_i, w_j) \log P(w_j \mid z_k) P(z_k \mid d_i)$$

完成目标函数的建立

- 关于参数$P(z_k|d_i)$、$P(w_j|z_k)$的函数E，并且，带有概率加和为1的约束条件：

$$E=\sum_i\sum_j n(d_i,w_j)\sum_{k=1}^{K}P(z_k\mid d_i,w_j)\log P(w_j\mid z_k)P(z_k\mid d_i)$$

$$s.t.\begin{cases}\sum_{j=1}^{M}P(w_j\mid z_k)=1\\ \sum_{k=1}^{K}P(z_k\mid d_i)=1\end{cases}$$

- 显然，这是只有等式约束的求极值问题，使用Lagrange乘子法解决。

目标函数的求解

- Lagrange函数为：

$$Lag = \sum_i \sum_j n(d_i, w_j) \sum_{k=1}^{K} P(z_k \mid d_i, w_j) \log P(w_j \mid z_k) P(z_k \mid d_i)$$

$$+ \sum_{k=1}^{K} \tau_k \left(1 - \sum_{j=1}^{M} P(w_j \mid z_k) \right) + \sum_{i=1}^{N} \rho_i \left(1 - \sum_{k=1}^{K} P(z_k \mid d_i) \right)$$

- 求驻点：

$$\frac{\partial Lag}{\partial P(w_j \mid z_k)} = \frac{\sum_i n(d_i, w_j) P(z_k \mid d_i, w_j)}{P(w_j \mid z_k)} - \tau_k \overset{令}{==} 0$$

$$\frac{\partial Lag}{\partial P(z_k \mid d_i)} = \frac{\sum_i n(d_i, w_j) P(z_k \mid d_i, w_j)}{P(z_k \mid d_i)} - \rho_i \overset{令}{==} 0$$

分析第一个等式

$$\frac{\partial Lag}{\partial P(w_j \mid z_k)} = \frac{\sum_i n(d_i, w_j) P(z_k \mid d_i, w_j)}{P(w_j \mid z_k)} - \tau_k \overset{令}{==} 0$$

$$\Rightarrow \sum_i n(d_i, w_j) P(z_k \mid d_i, w_j) = \tau_k P(w_j \mid z_k)$$

$$\Rightarrow \sum_{m=1}^{M} \sum_i n(d_i, w_j) P(z_k \mid d_i, w_j) = \sum_{m=1}^{M} \tau_k P(w_j \mid z_k)$$

$$\Rightarrow \sum_{m=1}^{M} \sum_i n(d_i, w_j) P(z_k \mid d_i, w_j) = \tau_k \sum_{m=1}^{M} P(w_j \mid z_k)$$

$$\Rightarrow \sum_{m=1}^{M} \sum_i n(d_i, w_j) P(z_k \mid d_i, w_j) = \tau_k$$

$$\xrightarrow{将\tau_k代回第二式} \sum_i n(d_i, w_j) P(z_k \mid d_i, w_j) = \sum_{m=1}^{M} \sum_i n(d_i, w_j) P(z_k \mid d_i, w_j) P(w_j \mid z_k)$$

$$\Rightarrow P(w_j \mid z_k) = \frac{\sum_i n(d_i, w_j) P(z_k \mid d_i, w_j)}{\sum_{m=1}^{M} \sum_i n(d_i, w_j) P(z_k \mid d_i, w_j)}$$

同理分析第二个等式

- 求极值时的解——M-Step：

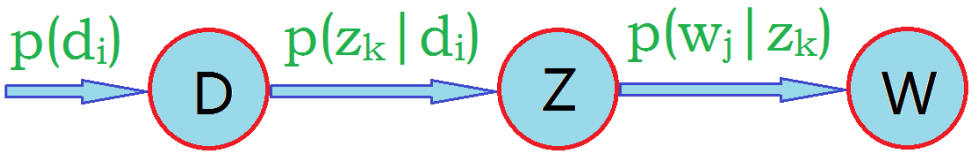
$$\begin{cases} P(w_j \mid z_k) = \dfrac{\sum_i n(d_i, w_j) P(z_k \mid d_i, w_j)}{\sum_{m=1}^{M} \sum_i n(d_i, w_j) P(z_k \mid d_i, w_j)} \\ P(z_k \mid d_i) = \dfrac{\sum_j n(d_i, w_j) P(z_k \mid d_i, w_j)}{\sum_{k=1}^{K} \sum_j n(d_i, w_j) P(z_k \mid d_i, w_j)} \end{cases}$$

- 别忘了E-step：$P(z_k \mid d_i, w_j) = \dfrac{P(w_j \mid z_k) P(z_k \mid d_i)}{\sum_{l=1}^{K} P(w_j \mid z_l) P(z_l \mid d_i)}$

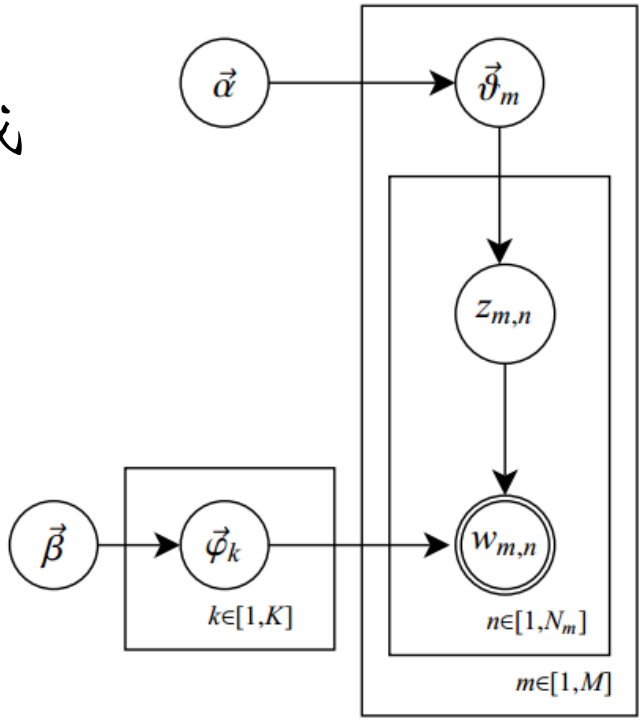
pLSA的总结

- pLSA应用于信息检索、过滤、自然语言处理等领域，pLSA考虑到词分布和主题分布，使用EM算法来学习参数。
- 虽然推导略显复杂，但最终公式简洁清晰，很符合直观理解，需用心琢磨；此外，推导过程使用了EM算法，也是学习EM算法的重要素材。

pLSA进一步思考



- 相对于“简单”的链状贝叶斯网络，可否给出“词”“主题”“文档”更细致的网络拓扑，形成更具一般性的模型？
- pLSA不需要先验信息即可完成自学习——这是它的优势。如果在特定的要求下，需要有先验知识的影响呢？
- 答：LDA模型；
  - 三层结构的贝叶斯模型
  - 需要超参数



参考文献

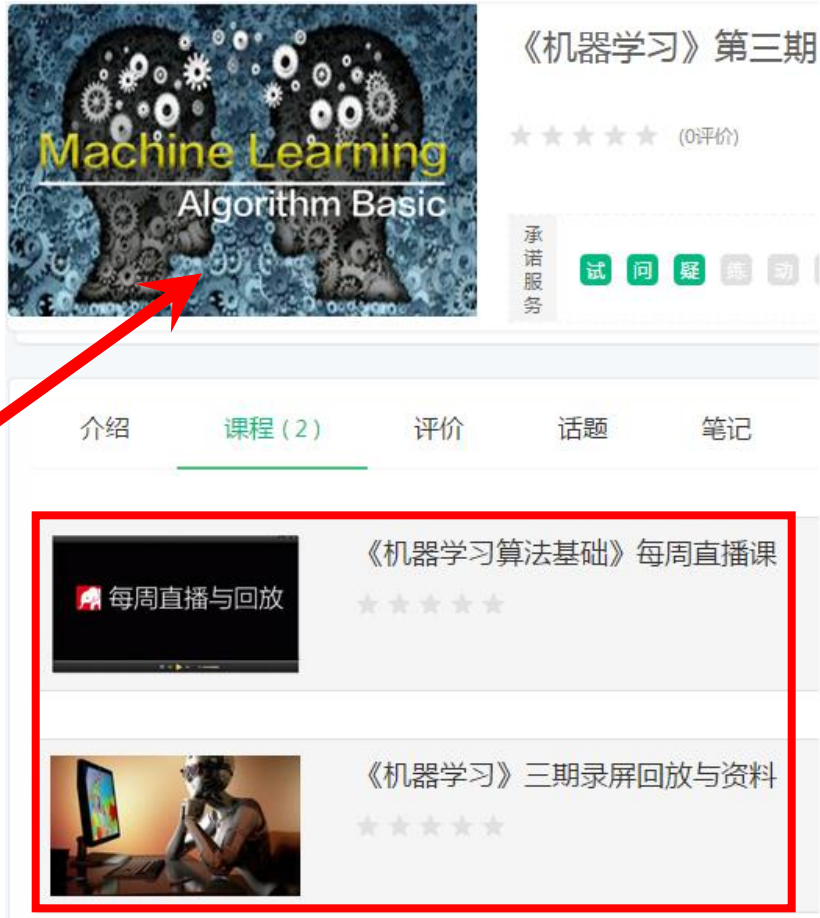
- M. Jordan, J. Kleinberg, ect. *Pattern Recognition and Machine Learning, Chapter 8.* 2006
- Irina Rish. *A Tutorial on Inference and Learning in Bayesian Networks*
- David Heckerman. *A Tutorial on Learning With Bayesian Networks* 1996
- Hans-Andrea Loeliger. *An Introduction to Factor Graphs*, MLSB 2008
- Frank R. Kschischang, Brendan J.Frey, ect. *Factor graph and sum-product algorithm.* 1998
- http://www.eng.yale.edu/pjk/eesrproj_02/luckenbill_html/node4.html(sum-product)

我们在这里

- http://wenda.ChinaHadoop.cn
  - 视频/课程/社区
- 微博
  - @ChinaHadoop
  - @邹博_机器学习
- 微信公众号
  - 小象
  - 大数据分析挖掘

课程资源

- 直播课的入口
- 录播视频和讲义资料



感谢大家！

恳请大家批评指正！