北京郵電大學
Beijing University of Posts and Telecommunications

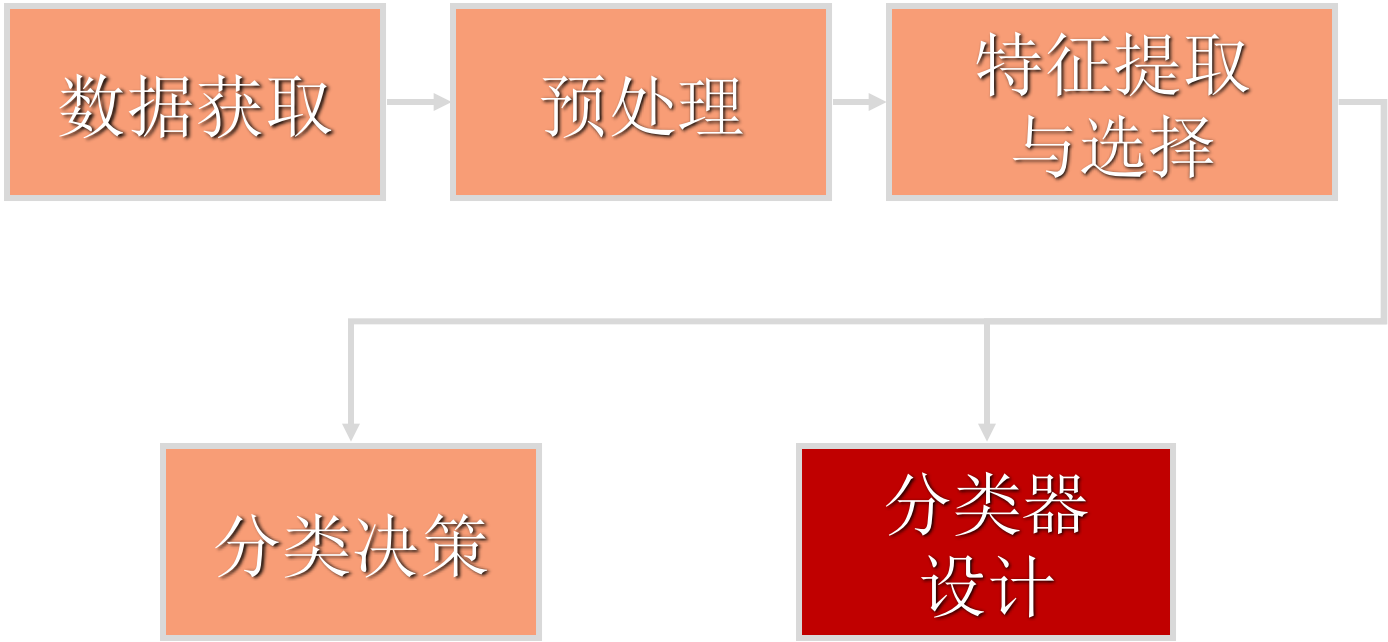
贝叶斯模型

戚 琦

网络与交换技术国家重点实验室　网络智能研究中心　科研楼511

qiqi8266@bupt.edu.cn　　13466759972

回顾：监督学习



大纲

- 贝叶斯决策论
- 朴素贝叶斯分类器
- 半朴素贝叶斯分类器
- 贝叶斯网络

基础知识

条件概率

描述两个有因果关系的随机事件之间的概率关系，$p(b|a)$定义为在事件a发生的前提下事件b发生的概率。

贝叶斯公式

两个随机事件之间的关系：

$$p(b|a) = \frac{p(a|b)p(b)}{p(a)}$$

贝叶斯，英国数学家。1702年出生于伦敦，做过牧师。1742年成为英国皇家学会会员。1763年4月7日逝世。贝叶斯在数学方面主要研究概率论。他首先将归纳推理法用于概率论基础理论，并创立了贝叶斯统计理论，对于统计决策函数、统计推断、统计的估算等做出了贡献。

分类问题

- 分类问题：特征向量取值x与样本所属类型y有因果关系。**因为样本属于类型y，所以具有特征值x**。
- 分类器：已知样本的特征向量x的条件下**反推样本所属的类别**。根据贝叶斯公式有

$$p(y|\boldsymbol{x}) = \frac{p(x|y)p(y)}{p(x)}$$

$p(x)$	**特征**向量的分布
$p(y)$	**类别**出现的概率，即先验分布
$p(x\|y)$	每个类别下，样本的某个特征的条件概率
$p(y\|x)$	样本属于每一类的概率，即后验概率

贝叶斯最优分类器

$$h^*(x) = arg\max_{c \in y} P(c|x)$$

对于每个样本x，选择使后验概率$P(c|x)$最大的类别标记

现实中难以直接获得

最小化分类错误率的贝叶斯最优分类器

判别模型与生成模型

- 机器学习所要实现的是基于有限的训练样本尽可能准确地估计出后验概率。

- 两种基本策略:
  - 判别式模型
    直接建模 P(c | x)：决策树、BP 神经网络、SVM
  - 生成式模型
    先建模联合概率分布 P(x, c)，再计算 P(c | x)：

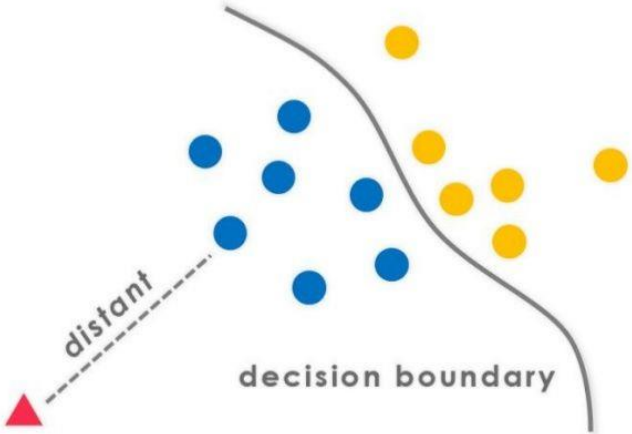
$$P(c|\boldsymbol{x}) = \frac{P(\boldsymbol{x}, c)}{P(\boldsymbol{x})}$$

    贝叶斯分类器、GAN、VAE等

判别模型与生成模型

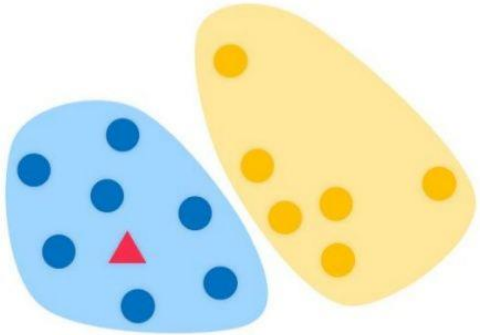
Discriminative vs. Generative

Discriminative



- Only care about estimating the conditional probabilities
- Very good when underlying distribution of data is really complicated (e.g. texts, images, movies)

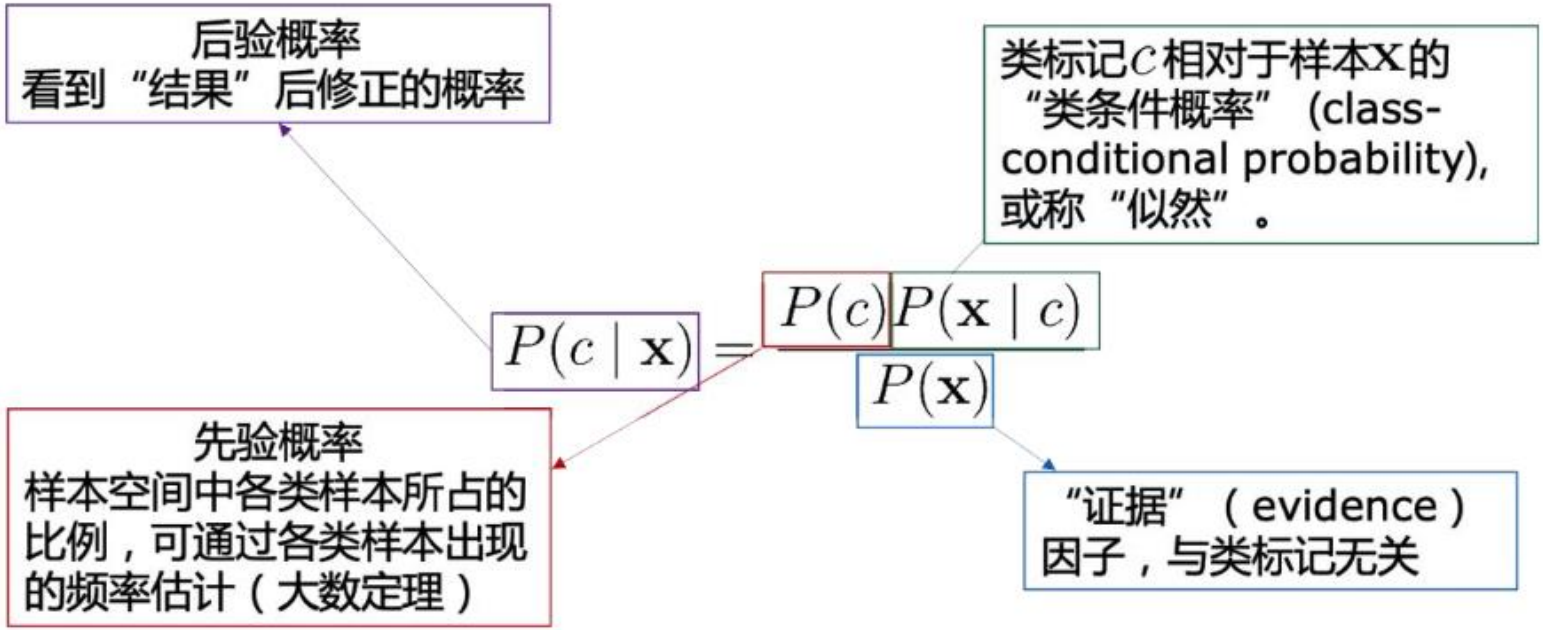
Generative



- Model observations (x,y) first, then infer p(y|x)
- Good for missing variables, better diagnostics
- Easy to add prior knowledge about data

生成式模型

- 基于贝叶斯定理，生成式模型可写成



大纲

- 贝叶斯决策论
- 朴素贝叶斯分类器
- 半朴素贝叶斯分类器
- 贝叶斯网络

条件独立性

给定类标号y，朴素贝叶斯分类器在估计类条件概率时**假设属性之间条件独立**。
条件独立假设可以形式化的表达如下：

$$P(X \mid Y = y) = \prod_{1}^{n} P(x_i \mid Y = y)$$

其中每个训练样本可用一个属性向量$X=(x_1,x_2,x_3\ldots,x_n)$表示，各个属性之间条件独立。

对于一篇文章“Good good study,Day day up.”

用一个文本特征向量来表示：x=(Good, good, study, Day, day , up)

一般各个词语之间肯定不是相互独立的，有一定的上下文联系。但在朴素贝叶斯文本分类时，我们假设个单词之间没有联系，可以用一个文本特征向量来表示这篇文章，这就是“朴素”的来历。

朴素贝叶斯如何工作

- 有了条件独立假设，就不必计算X和Y的每一种组合的类条件概率，只需对给定的Y，计算每个x_i的条件概率。不需要很大的训练集就能获得较好的概率估计，更实用。

- 估计分类属性的条件概率：$P(x_i \mid Y=y)$怎么计算呢？

一般根据类别y下包含属性x_i的实例的比例来估计。

以文本分类（比如积极或消极）为例，x_i表示一个单词，

$$P(x_i \mid Y=y)=\frac{\text{该类别下包含单词的xi的文章总数}}{\text{该类别下的文章总数}}$$

贝叶斯分类器举例

假设给出如下训练样本数据（14天打球情况、4种环境变化），学习的目标是根据给定的天气状况判断对PlayTennis这个请求的回答是Yes还是No。

先验概率

$$p(y = yes) = 9/14$$

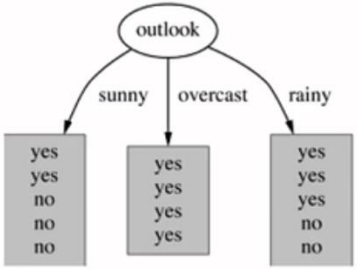
$$p(y = no) = 5/14$$

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

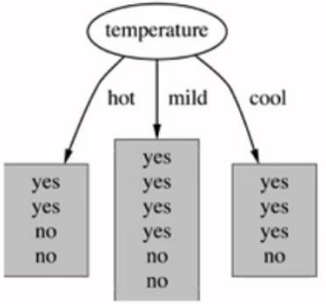
*回顾决策树

- 决策树构造实例
- 划分方式：4种
- 问题：谁当根节点呢？
- 依据：信息增益

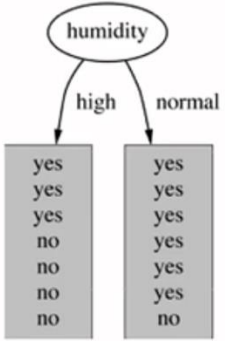
1.基于天气的划分



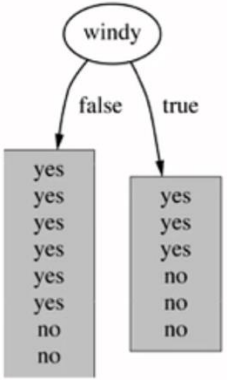
2.基于温度的划分



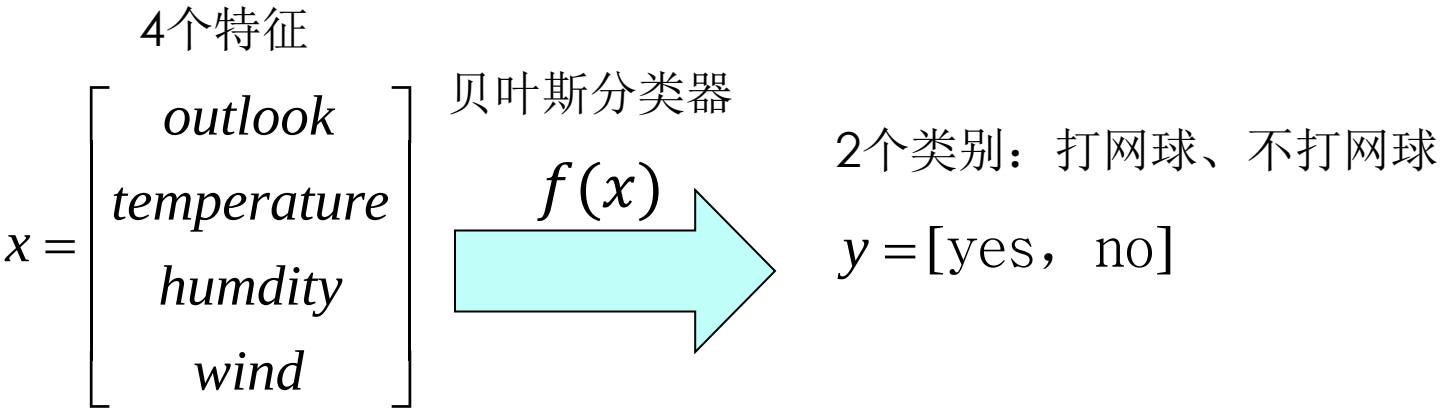
3.基于湿度的划分



4.基于有风的划分



贝叶斯分类器举例



利用训练数据**计算后验概率P(Yes|x)和P(No|x)**，

测试样本，如果P(Yes|x)>P(No|x)，那么分类为Yes，否则为No。

贝叶斯分类器举例

朴素贝叶斯分类器来分类左侧测试样本：

$$x = \begin{bmatrix} outlook = sunny \\ temperature = cool \\ humdity = high \\ wind = strong \end{bmatrix}$$

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D8	Sunny	Mild	High	Weak	No
D14	Rain	Mild	High	Strong	No
D6	Rain	Cool	Normal	Strong	No

P(Outlook = Sunny | No)=3/5

P(Temperature = Cool | No) =1/5

P(Humidity = High | No) =4/5

P(Wind = Strong |No) =3/5

$$\Rightarrow \mathbf{P(X \mid Y = NO)} = \frac{3}{5} * \frac{1}{5} * \frac{4}{5} * \frac{3}{5} = \frac{36}{625} \qquad \mathbf{p(Y = no)} = 5/14$$

$$\Rightarrow \mathbf{P(X \mid Y = NO) * P(Y = NO)} = \frac{36}{625} * \frac{5}{14} = \frac{18}{875}$$

贝叶斯分类器举例

P(Outlook = Sunny|Yes)=2/9

P(Temprature = Cool |Yes) =3/9

P(Humidity = High |Yes) =3/9

P(Wind = Strong |Yes) =3/9

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D7	Overcast	Cool	Normal	Strong	Yes
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes

$$P(P = yes) = 9/14$$

$$P(X \mid Y = YES) = \frac{2}{9} * \frac{3}{9} * \frac{3}{9} * \frac{3}{9} = \frac{2}{283}$$

$$P(X \mid Y = YES)P(P = YES) = \frac{2}{9} * \frac{3}{9} * \frac{3}{9} * \frac{3}{9} * \frac{9}{14} = \frac{1}{189}$$

所以该样本分类为No

$$P(X \mid Y = NO)P(Y = NO) = \frac{18}{875} > P(X \mid Y = YES)P(P = YES) = \frac{1}{189}$$

朴素贝叶斯分类器表达式

周志华，《机器学习》

由于对所有类别来说 $P(\boldsymbol{x})$ 相同, 因此基于式(7.6)的贝叶斯判定准则有

$$h_{nb}(\boldsymbol{x}) = \arg\max_{c \in \mathcal{Y}} P(c) \prod_{i=1}^{d} P(x_i \mid c) , \tag{7.15}$$

这就是朴素贝叶斯分类器的表达式.

显然, 朴素贝叶斯分类器的训练过程就是基于训练集 D 来估计类先验概率 $P(c)$, 并为每个属性估计条件概率 $P(x_i \mid c)$.

令 D_c 表示训练集 D 中第 c 类样本组成的集合, 若有充足的独立同分布样本, 则可容易地估计出类先验概率

$$P(c) = \frac{|D_c|}{|D|} . \tag{7.16}$$

离散

$$P(x_i \mid c) = \frac{|D_{c,x_i}|}{|D_c|} .$$

连续

例如，高斯分布

$$p(x_i \mid c) = \frac{1}{\sqrt{2\pi}\sigma_{c,i}} \exp\left(-\frac{(x_i - \mu_{c,i})^2}{2\sigma_{c,i}^2}\right) .$$

拉普拉斯修正

- 若某个属性值在训练集中没有与某个类同时出现过，则直接计算连乘式的概率值为 0，导致分类结果显然不合理
- 为了避免其他属性携带的信息被训练集中未出现的属性值“抹去”，在估计概率值时通常要进行“**拉普拉斯修正**”
- 令 N 表示数据集 D 中可能的类别数，N_i 表示第 i 个属性可能的取值数，则

先验概率 $\hat{P}(c) = \frac{|D_c| + 1}{|D| + N}$

类条件概率 $\hat{P}(x_i|c) = \frac{|D_{c,x_i}| + 1}{|D_c| + N_i}$

拉布拉斯修正，实际上是假设了属性与类别均匀分布

多种使用方式

- 若任务对预测速度要求较高:
  计算所有概率估值存储，使用时“查表”

- 若任务数据更替频繁:
  使用“懒惰学习”，即先不进行任务训练，收到预测请求时再估值

- 若任务数据不断增加:
  使用“增量学习”，基于现有估值，对新样本涉及的概率估值进行修正

大纲

- 贝叶斯理论
- 朴素贝叶斯分类器
- 半朴素贝叶斯分类器
- 贝叶斯网络

半朴素贝叶斯分类器

- 为了降低贝叶斯公式中估计后验概率的困难，朴素贝叶斯分类器采用了**属性条件独立性假设**；
- 对属性条件独立假设进行一定程度的放松，由此产生了一类称为“**半朴素贝叶斯分类器**”。
- 半朴素贝叶斯分类器**最常用**的一种策略：“独依赖估计” (**One-Dependent Estimator**, **ODE**)，假设每个属性在类别之外最多仅依赖一个其他属性，即

$$P(c|\boldsymbol{x}) \propto P(c) \prod_{i=1}^{d} P(x_i|c, pa_i)$$

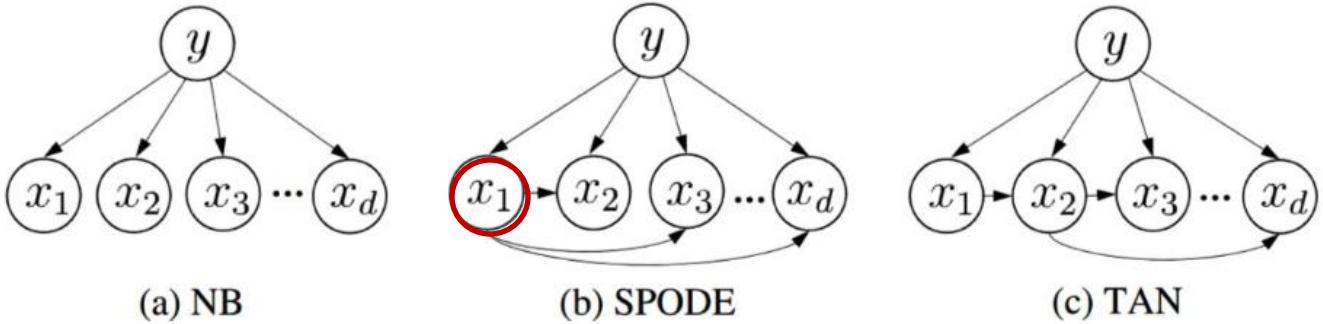
pa_i为属性 x_i 所依赖的属性，称为 x_i 的父属性

- 对每个属性 x_i，若其父属性 pa_i 已知，则可估计概率值 $\mathbf{P(x_i|c, pa_i)}$，于是问题的关键转化为如何确定每个属性的父属性。

常见方法

- **SPODE**(Super-Parent ODE) 假设所有属性都依赖于同一属性，称为“超父”(Super-Parent)，可通过交叉验证等模型选择方法来确定超父属性。

- **TAN**(Tree Augmented naïve Bayes) 以属性间的条件“互信息”(mutual information) 为边的权重，构建完全图，再利用最大带权生成树算法，仅保留强相关属性间的依赖性。



(a) NB (b) SPODE (c) TAN

朴素贝叶斯 x_1为超父

大纲

- 贝叶斯理论
- 朴素贝叶斯分类器
- 半朴素贝叶斯分类器
- 贝叶斯网络

贝叶斯网络

1985年由Judea Pearl首先提出

- 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），是有向无环图模型，也是一种概率图模型，借由有向无环图DAG中得知一组随机变量$\{X_1, X_2 ... X_n\}$及其n组条件概率分布的性质。
- 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。

*概率图模型

- 概率图模型：用图来表示变量概率依赖关系的理论
- 结合概率论与图论，利用图表示与模型有关的变量的联合概率分布

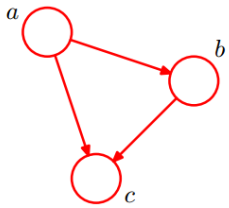
贝叶斯网络（Bayesian Network）：有向图结构表示

马尔可夫网络（Markov Network）：无向图结构表示

概率图模型包括朴素贝叶斯模型、最大熵模型、隐马尔可夫模型、条件随机场、主题模型等，在机器学习的诸多场景中都有着广泛的应用

常见的贝叶斯网络

- 一个简单的贝叶斯网络



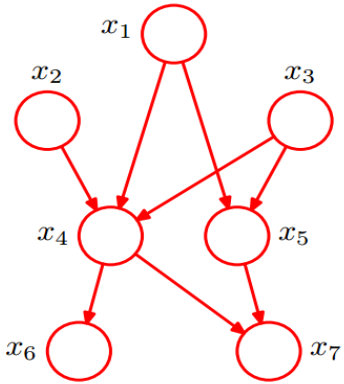
$$p(a,b,c) = p(c|a,b)p(b|a)p(a)$$

- 全连接贝叶斯网络：每一对结点之间都有边连接

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

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

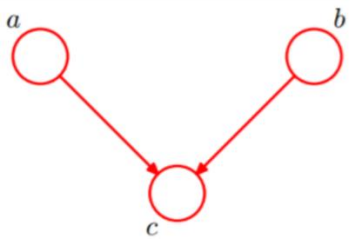
- 正常的贝叶斯网络：有些边缺失，直观上
  - x_1和x_2独立
  - x_6和x_7在x_4给定的条件下独立
  - $x_1,x_2,\ldots x_7$的联合分布：



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

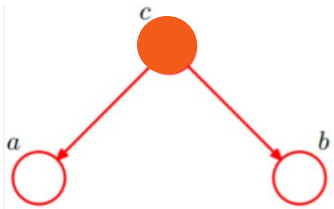
贝叶斯网络的结构形式

1. head-to-head



$p(a,b,c) = p(a)p(b)p(c|a,b)$ 成立，即在c未知的条件下，a、b被阻断(blocked)，是独立的，称之为head-to-head条件独立。

2. tail-to-tail

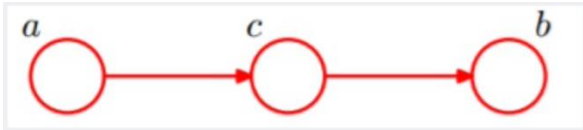


在c未知的条件下：$p(a,b,c) = p(c)p(a|c)p(b|c)$，此时，无法得出 $p(a,b) = p(a)p(b)$，即c未知时，a、b不独立；

在c已知的条件下：$p(a,b|c) = \frac{p(a,b,c)}{p(c)}$，将 $p(a,b,c) = p(c)p(a|c)p(b|c)$，带入式子得到：$p(a,b|c) = p(a|c)p(b|c)$，即c已知时，a、b独立。

贝叶斯网络的结构形式

3. head-to-tail



- **c未知时：**

$p(a,b,c) = p(c)p(a|c)p(b|c)$，此时，无法得出$p(a,b) = p(a)p(b)$，即c未知时，a、b不独立；

- **c已知时：**

$p(a,b|c) = \frac{p(a,b,c)}{p(c)}$，且根据$p(a,c) = p(a)p(c|a) = p(c)p(a|c)$，可得到：

$$p(a,b|c) = \frac{p(a,b,c)}{p(c)} = \frac{p(a)p(c|a)*p(b|c)}{p(c)} = \frac{p(a,c)p(b|c)}{p(c)} = p(a|c)p(b|c)$$

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

head-to-tail是一个链式网络

在x_i给定的条件下，x_{i+1}的分布和$x_1,x_2...x_{i-1}$条件独立

x_{i+1}的分布状态只和x_i有关，和其他变量条件独立。

当前状态只跟上一状态有关，跟上上或上上之前的状态无关-----**马尔科夫链**（Markov chain）

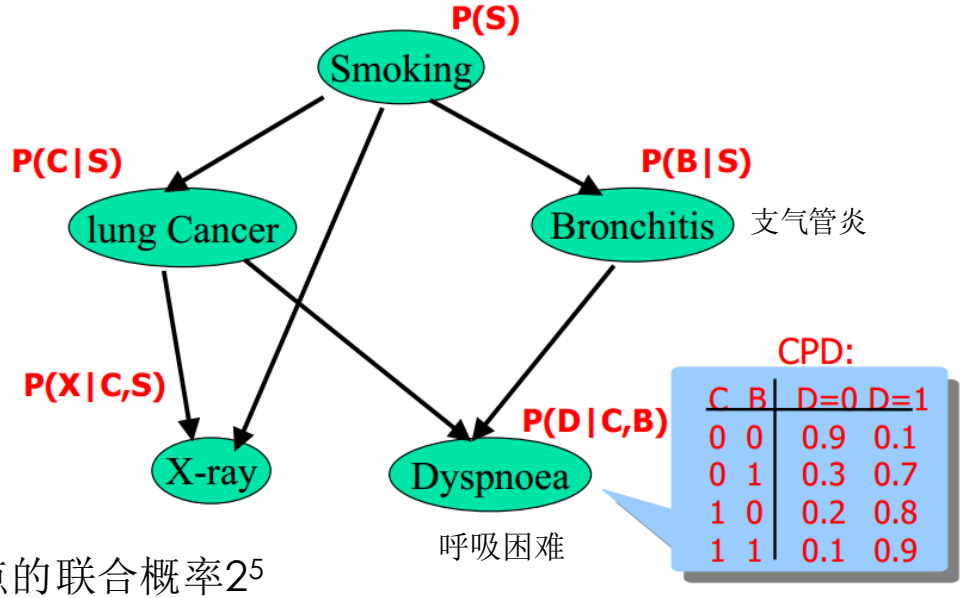
贝叶斯网络举例

通过先验知识，根据节点因果关系构建

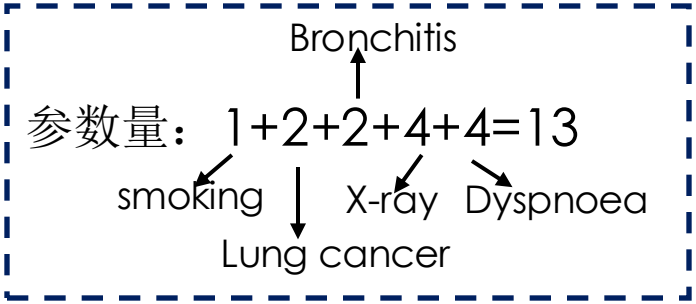
节点X的条件概率：P(X | parent(X))

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

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



条件概率表

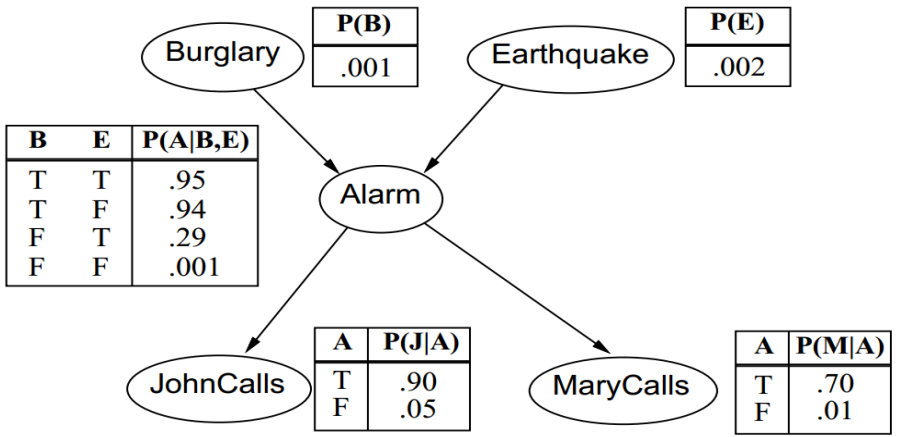


所有节点的联合概率2^5

两个依赖的父节点就是4行（00,01,10,11）

“西瓜视频：机器学习课程”

贝叶斯网络举例



全部随机变量的联合分布

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | parents(X_i))$$

$$P(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$$

$$= P(j|a)P(m|a)P(a|\neg b, \neg e)P(\neg b)P(\neg e)$$

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

$$\approx 0.00063$$

盗窃报警，地震也报警，警报响后，John可以打电话，Mary也可以打电话

求：John打了电话，Mary也打了电话，警报也响了，但是没有盗窃发生，也没有地震的概率？

总结：判别模型与生成模型

假设有四个samples：

	sample 1	sample 2	sample 3	sample 4
x	0	0	1	1
y	0	0	0	1

生成模型：

	y = 0	y = 1
x = 0	1/2	0
x = 1	1/4	1/4

判别模型：

	y = 0	y = 1
x = 0	1	0
x = 1	1/2	1/2

同一套数据，
统计方法不同