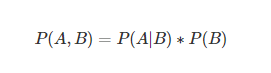
**贝叶斯网络笔记**

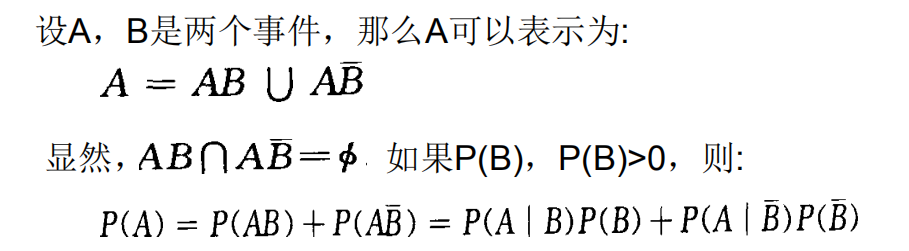
构建贝叶斯网络之前先列出故障表现，然后用可能导致这些故障的原因指向故障表现，注意Pgmpy

1. **几个必要的公式**

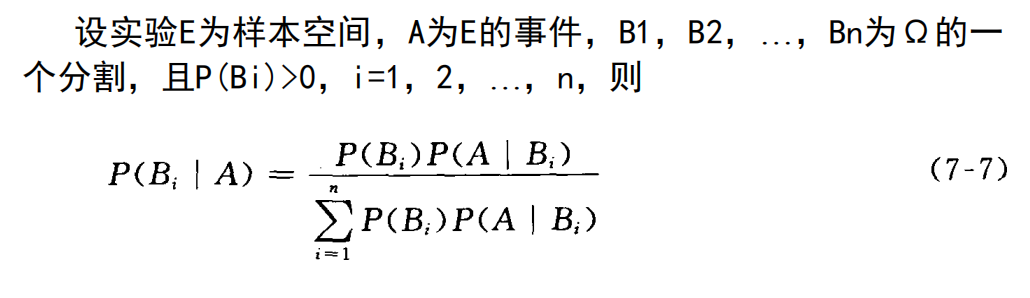
**1.条件概率公式**



**2.全概率公式**



**3.贝叶斯公式**



**二、贝叶斯网络介绍**

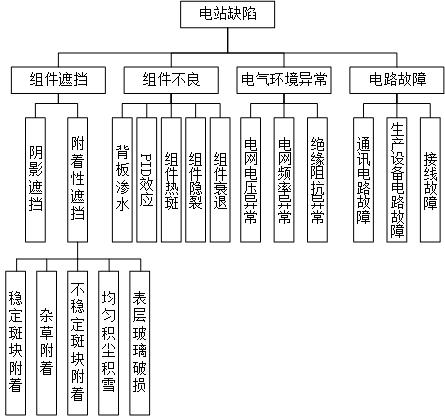
 贝叶斯网络是一个典型的图模型，它对感兴趣变量（variables of interest）及变量之间的关系（relationships）进行建模。当将贝叶斯模型与统计技术一起使用时，这种图模型分析数据具有如下几个优势：

(1)    贝叶斯学习能够方便的**处理不完全数据**。例如考虑具有相关关系的多个输入变量的分类或回归问题，对标准的监督学习算法而言，变量间的相关性并不是它们处理的关键因素，当这些变量中有某个缺值时，它们的预测结果就会出现很大的偏差。而贝叶斯学习则提供了较为直观的概率关联关系模型。  
(2)    贝叶斯学习能够**学习变量间的因果关系**。因果关系是数据挖掘中极为重要的模式。原因有二：在数据分析中，因果关系有利于对领域知识的理解；在干扰较多时，便于作出精确的预测。  
(3)    贝叶斯网络与贝叶斯统计相结合能够**充分利用领域知识和样本数据的信息。**任何从事过实际建模任务的人都会知道先验信息或领域知识在建模方面的重要性，尤其是在样本数据稀疏或数据较难获得的时候，一些商业方面的专家系统完全根据领域专家知识来构建就是一个很好的例证。贝叶斯网络用弧表示变量间的依赖关系，用概率分布表来表示依赖关系的强弱，将先验信息与样本知识有机结合起来。

1. 贝叶斯统计方法可以和贝叶斯网络一起使用，**避免了数据过度拟合**（the overfiting of data）。

贝叶斯学习理论在数据挖掘中获得了成功的应用。对贝叶斯学习理论研究最大的动力就是它在实际应用中的巨大作用和潜力。目前，贝叶斯学习理论已成功地应用到智能用户接口、信息滤波、车辆自动导航、武器制导、医疗诊断、经济预测和文本分类等诸多领域。

三、**贝叶斯模型构建**

****

构建贝叶斯网络包括以下三部分内容：

(1)变量的定义；

(2)结构学习；

(3)参数学习。

网络结构学习的目标是找到和样本数据D匹配度最好的贝叶斯网络结构。贝叶斯网络的参数学习实质上是在已知网络结构的条件下，来学习每个节点的概率分布表。

这三个任务之间一般是顺序进行的，然而在构造过程中一般需要在以下两个方面作折中:一方面为了达到足够的精度，需要构建一个足够大的、丰富的网络模型；另一方面，要考虑构建、维护模型的费用和考虑概率推理的复杂性。实际上建立一个贝叶斯网络往往是上述三个过程迭代地、反复地交互过程。其中第二、三个任务是构建贝叶斯网络的关键点也是难点所在，主要是构建出一个有向无环图并给出图中每个结点的分布参数，即每个节点都对应一个条件概率分布表。

第一个任务主要是在领域专家的指导下选取适当研究问题领域的变量，同时在有些情况下也需要一定的策略从专家提供的变量中选择重要的因子。

第二个任务是构建贝叶斯网络的关键点也是难点所在，主要是构建 出一个有向无环图并给出图中每个结点的分布参数，即每个节点都对应一个 条件概率分布表(CPT)。

1. 人工定义：根据先前对研究事物的统计研究，完全由人工定义网络结构和参数。没有网络学习的过程。  
2. 部分学习：根据先前的研究定义网络的结构，包括节点（变量）和有向边（因果关系），通过数据训练网络参数，相应的学习算法有**最大似然法和贝叶斯估计**。  
3.完全学习：根据具体研究目的定义节点（变量），通过数据学习网络结构，相应的算法有**爬山法，K2**等。然后再通过数据训练网络的参数。[e01]

四、概率分布（Distributions）

如图1所示，这是最简单的联合分布案例，姑且称之为学生模型。

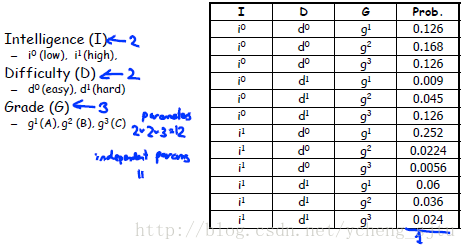


图1

其中包含3个变量，分别是：I（学生智力，有0和1两个状态）、D（试卷难度，有0和1两个状态）、G（成绩等级，有1、2、3三个状态）。

表中就是概率的联合分布了，表中随便去掉所有包含某个值的行，就能对分布表进行缩减。

例如可以去掉所有G不为1的行，这样就只剩下了1、4、7、10行，这样他们的概率之和就不为1了，所以可以重新标准化（Renormalization）。如图2所示。

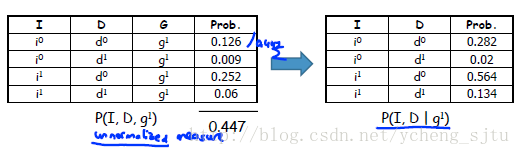


图2

反之也可以把所有含有某个值得行相加，就是边缘化（Marginalization），如图3所示。

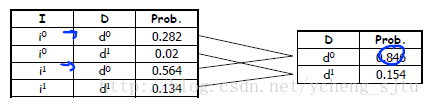


图3

条件概率分布（Conditional ProbabilityDistribution, CPD）

已知学生的智力和试卷难度，学生得分的分布就是条件概率。如图4所示。

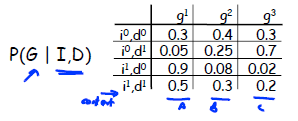


图4

因子（Factors）

因子是随机变量的函数。

因子是处理概率分布的的基本手段。

因子是高维空间中用以定义概率分布的基本单元。

IMG_260

因子可以相乘（图5）、边缘化（图6）以及缩减（图7）。

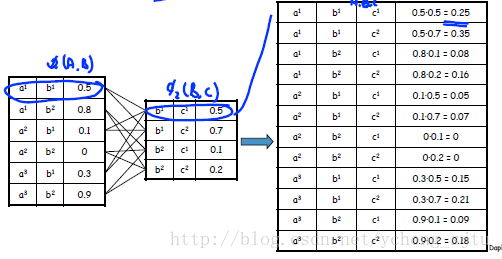


图5

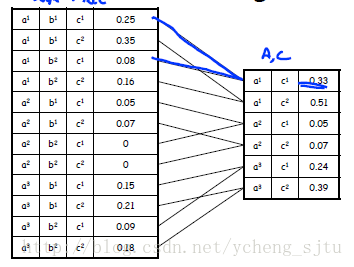


图6

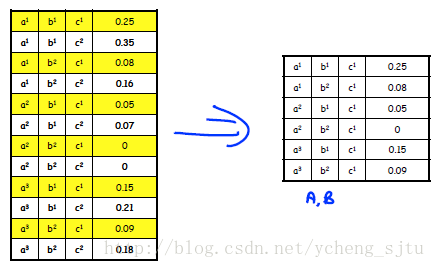


图7

前面提到的学生模型，其条件概率分布可以画在一张图里面，如图8.

每个节点代表一个因子，其中有些CPD已经蜕化成非条件概率了。

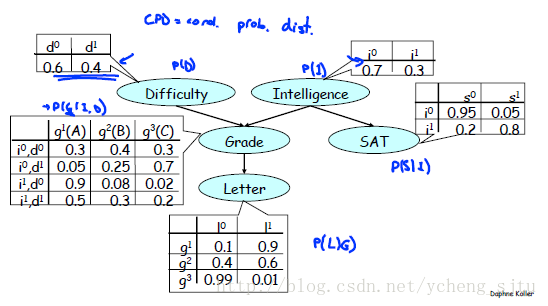


图8

贝叶斯网络的链式法则（Chain Rule）

如图9所示。概率分布由因子的积来定义。

IMG_265

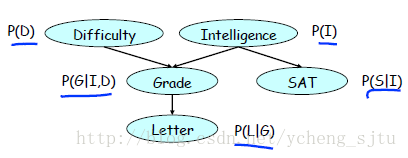


图9

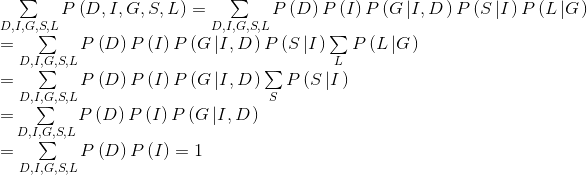
例如

IMG_267

因此，通过链式法则，贝叶斯网络能够表示联合概率分布：

IMG_268

贝叶斯网络的重要性质是概率和为1



一个简单的概率图是血型模型

其中G指基因型，B指血型。可以看到血型只由自己的基因型决定，而基因型则由父母两人的基因型决定。如图10.

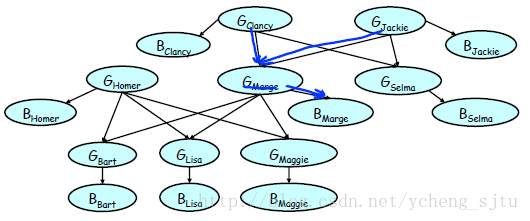
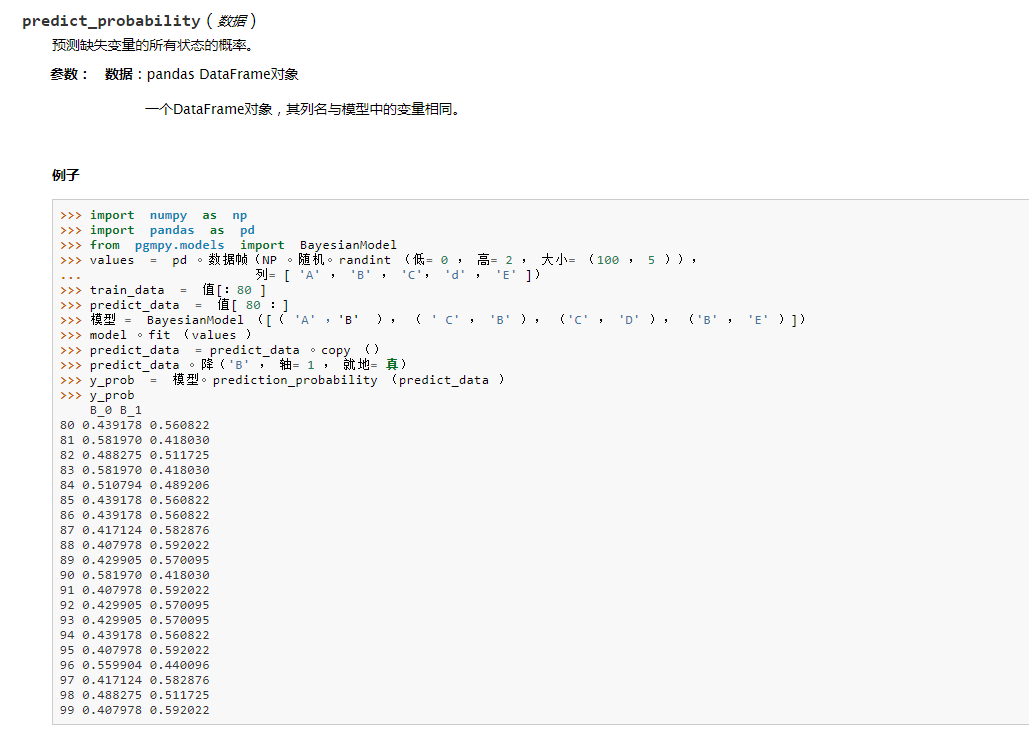


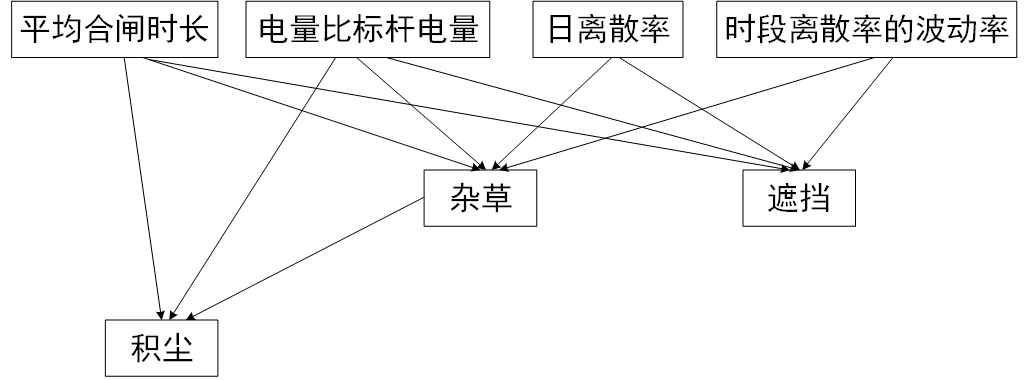
图10

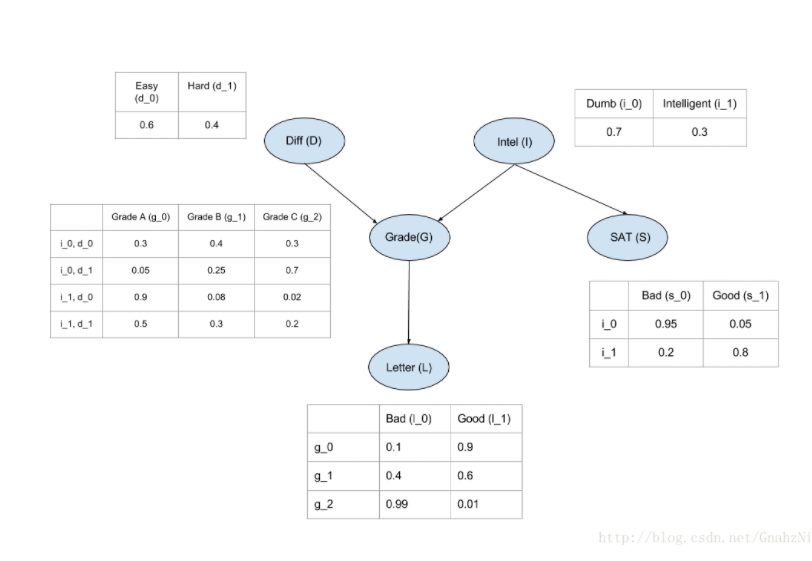
1.变量定义

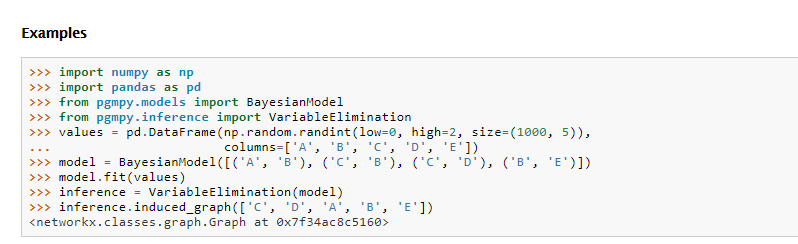
发电量 离散率 告警信息

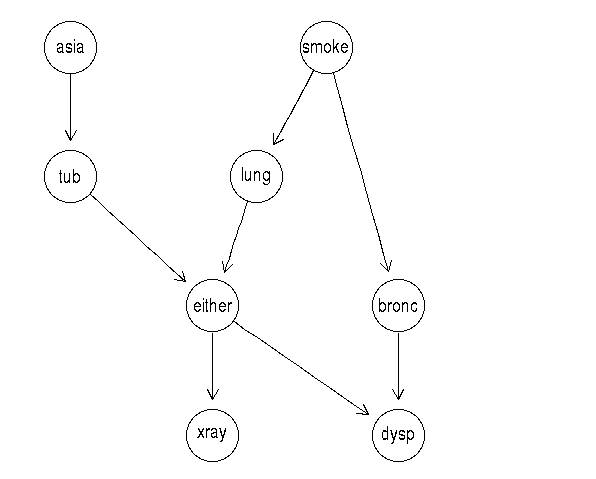
固定物遮挡 杂草遮挡 积尘 光伏板破损 热斑











Pgmpy官方文档

<http://pgmpy.org/models.html#module-pgmpy.models.NoisyOrModel>

贝叶斯网络各种数据

<http://www.bnlearn.com/bnrepository/#asia>

贝叶斯网络亚洲数据实例

<https://github.com/pgmpy/pgmpy/blob/dev/examples/Inference%20in%20Bayesian%20Networks.ipynb>

斯坦福概率图讲课

第一讲 贝叶斯网络基础

<https://blog.csdn.net/yangliuy/article/details/8067261>

第二讲 Template Models and Structured CPDs

<https://blog.csdn.net/yangliuy/article/details/8091630>

连续属性离散化

<https://blog.csdn.net/jbb0523/article/details/78806593>

补充1：

<https://blog.csdn.net/jbb0523/article/details/79843496>

补充2：

<https://blog.csdn.net/jbb0523/article/details/80597506>

贝叶斯网络参数学习实例

<https://max.book118.com/html/2017/1203/142597585.shtm>

贝叶斯网络参数学习（建模全过程）

<https://blog.csdn.net/snoopy_yuan/article/details/66477014>

Matlab贝叶斯工具箱

<http://www.cs.ubc.ca/~murphyk/Software/BNT/usage.html#examples>