**第四章 贝叶斯分类器**

**1. 简述朴素贝叶斯的优缺点.**

  优点：

    （1）朴素贝叶斯模型发源于古典数学理论，有稳定的分类效率。

    （2）对小规模的数据表现很好，能个处理多分类任务，适合增量式训练，

尤其是数据量超出内存时，我们可以一批批的去增量训练。

    （3）对缺失数据不太敏感，算法也比较简单，常用于文本分类。

缺点：

    （1）理论上，朴素贝叶斯模型与其他分类方法相比具有最小的误差率。但、是实际上并非总是如此，这是因为朴素贝叶斯模型给定输出类别的情况下,假设属性之间相互独立，这个假设在实际应用中往往是不成立的，在属性个数比较多或者属性之间相关性较大时，分类效果不好。而在属性相关性较小时，朴素贝叶斯性能最为良好。对于这一点，有半朴素贝叶斯之类的算法通过考虑部分关联性适度改进。

    （2）需要知道先验概率，且先验概率很多时候取决于假设，假设的模型可

以有很多种，因此在某些时候会由于假设的先验模型的原因导致预测

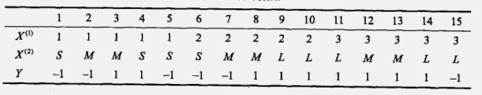
效果不佳。

    （3）由于我们是通过先验和数据来决定后验的概率从而决定分类，所以分

类决策存在一定的错误率。

    （4）对输入数据的表达形式很敏感。

**2. 试由下表的训练数据学习一个朴素贝叶斯分类器并确定 的类标记y. 表中 ， 为特征，取值的集合分别为， , Y为类标记，**













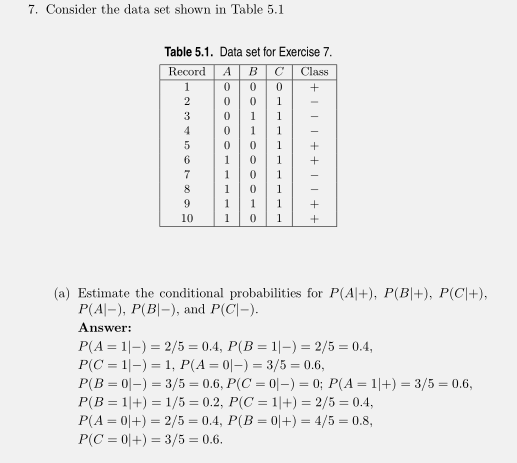
对于给定的计算：





因为最大，所以

**3. 考虑下表中的数据集**



1. **估计条件概率P(A|+)，P(B|+)，P(C|+)，P(A|-)，P(B|-)和P(C|-).**









1. **根据(a)中的条件概率，使用朴素贝叶斯方法预测测试样本(A=0，B=1，C=0)的类标号**

假设，则K属于两个类的概率为：







因此可以得到，此样本的类标号是：+

1. **使用m估计方法(p=1/2且m=4)估计条件概率**

M-估计公式：









1. **同(b)，使用(c)中的条件概率**

假设，则K属于两个类的概率为：





因此可以得到，此样本的类标号是：-

1. **比较估计概率的两种方法。哪一种更好？为什么？**

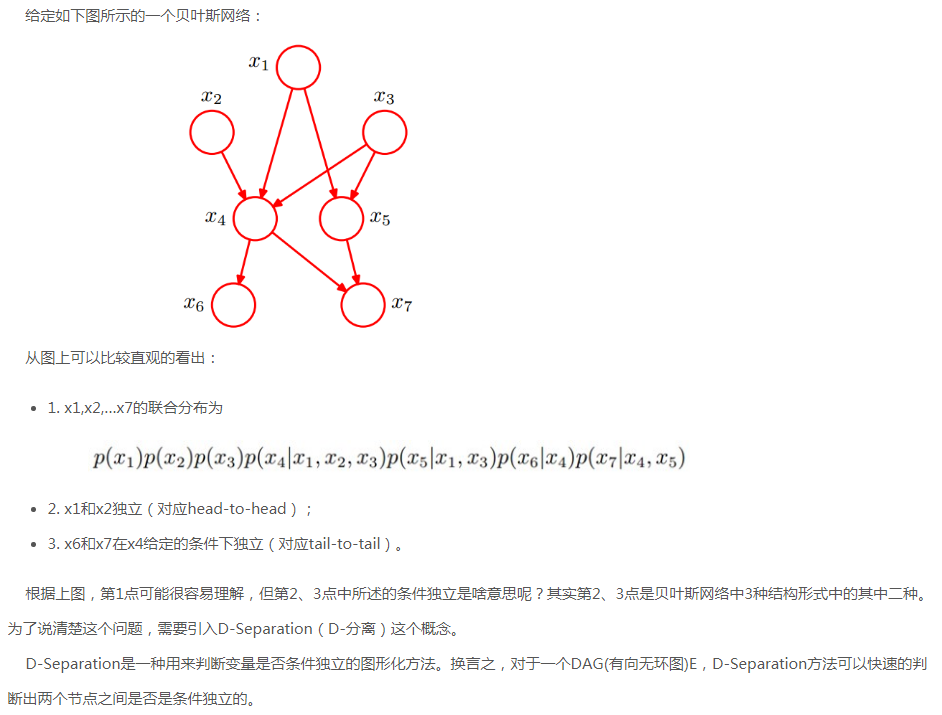
当条件概率有为0的时候，用M-估计的方法比条件概率的贝叶斯更好，因为我们不想整个概率计算结果为0，这样一般是不符合实际的。M-估计可以避免概率计算结果为0的情况

**4. 给定如下图所示的一个贝叶斯网络**

**(a) 请写出的联合概率分布**

**(b)和是否相互独立？**

**(c) 和在给定的条件下是否相互独立？**



1. 由上图可以看出x1,x2,x3,…,x7的联合概率分布为：

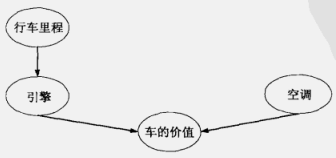


1. X1和x2独立（对应head-to-head），因为x1和x2都是两个head，head-head之间是相互独立的
2. x6和x7在x4给定的条件下独立（对应tail-to-tail）；因为x6的概率只与x4的概率有关，x7的概率与x4，x5有关，若给定x4的条件下，显然此时x7的概率取决于x5，与x6无关，x6-x7等价于tail-tail，二者相互独立

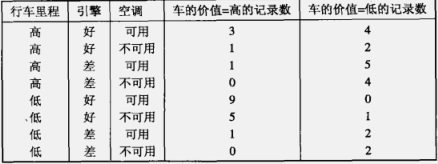
**5. 下图给出了表中的数据集对应的贝叶斯信念网络（假设所有的属性都是二元的）。**

**(a) 画出网络中每个结点对应的概率表。**

**(b) 使用贝叶斯网络计算P(引擎=差，空调=不可用)。**



贝叶斯信念网络



数据集

①行车里程

| **行车里程=高** | **行车里程=低** |
| --- | --- |
| 0.5 | 0.5 |

②空调

| **空调=可用** | **空调=不可用** |
| --- | --- |
| 5/8 | 3/8 |

③引擎

|  | **引擎=好** | **引擎=差** |
| --- | --- | --- |
| **行车里程=高** | 0.5 | 0.5 |
| **行车里程=低** | 0.75 | 0.25 |

④车的价值

| **引擎** | | **空调** | **车的价值=高** | **车的价值=低** | |
| --- | --- | --- | --- | --- | --- |
| **好** | **可用** | | 12/16=0.75 | | 4/16=0.25 |
| **好** | **不可用** | | 6/9=2/3 | | 3/9=1/3 |
| **差** | **可用** | | 2/9 | | 7/9 |
| **差** | **不可用** | | 0 | | 1 |

(b) 计算过程如下：

