**第七章 贝叶斯分类器**

7.1 为什么朴素贝叶斯如此“朴素”？

  参考答案：因为它假定所有的特征在数据集中的作用是同样重要且独立的。这个假设在现实世界中是不真实的，因此说朴素贝叶斯很“朴素”。

7.2 简述朴素贝叶斯的优缺点.

  参考答案：

优点：对小规模的数据表现很好，适合多分类任务，适合增量式训练。

缺点：对输入数据的表达形式很敏感。

7.3 什么是线性分类器？朴素贝叶斯分类器是线性分类器吗？

  参考答案：如果模型是参数的线性函数，并且存在线性分类面，那么就是线性分类器，否则不是。朴素贝叶斯分类器是线性分类器。

7.4 为什么说朴素贝叶斯是高偏差低方差？

参考答案：它简单的假设了各个特征之间是无关的，是一个被严重简化了的模型。所以，对于这样一个简单模型，大部分场合都会bias部分大于variance部分，也就是高偏差，低方差

7.5 什么是置信区间？

参考答案：置信区间不能用贝叶斯学派的概率来描述，它属于频率学派的范畴。真值要么在，要么不在。由于在频率学派当中，真值是一个常数，而非随机变量（后者是贝叶斯学派），所以我们不对真值做概率描述。比如，95%置信区间，并不是真值在这个区间内的概率是95%，而应该为100次随机抽样中构造的100个区间如果95次包含了参数真值，那么置信度为95%.

7.6 Google的拼写检查基于贝叶斯方法，怎么利用贝叶斯方法实现"拼写检查"的功能？

参考答案：

用户输入一个单词时，可能拼写正确，也可能拼写错误。如果把拼写正确的情况记做c（代表correct），拼写错误的情况记做w（代表wrong），那么"拼写检查"要做的事情就是：在发生w的情况下，试图推断出c。换言之：已知w，然后在若干个备选方案中，找出可能性最大的那个c，也就是求http://img.blog.csdn.net/20141112230745208的最大值。

根据贝叶斯定理，有：

http://img.blog.csdn.net/20141112230845578

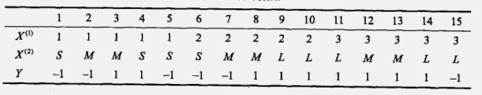
由于对于所有备选的c来说，对应的都是同一个w，所以它们的P(w)是相同的，因此我们只要最大化http://img.blog.csdn.net/20141112231004421 即可。其中：

P(c)表示某个正确的词的出现"概率"，它可以用"频率"代替。如果我们有一个足够大的文本库，那么这个文本库中每个单词的出现频率，就相当于它的发生概率。某个词的出现频率越高，P(c)就越大。比如在你输入一个错误的词“Julw”时，系统更倾向于去猜测你可能想输入的词是“July”，而不是“Jult”，因为“July”更常见。

P(w|c)表示在试图拼写c的情况下，出现拼写错误w的概率。为了简化问题，假定两个单词在字形上越接近，就有越可能拼错，P(w|c)就越大。

所以，我们比较所有拼写相近的词在文本库中的出现频率，再从中挑出出现频率最高的一个，即是用户最想输入的那个词。

7.7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，



参考答案：

根据贝叶斯算法，容易计算下列概率：

，

， ，

， ，

， ，

， ，

对于给定的计算：

所以y = -1

7.8 问题同上题，按照拉普拉斯平滑估计概率，即取λ=1.

参考答案：

，

， ，

， ，

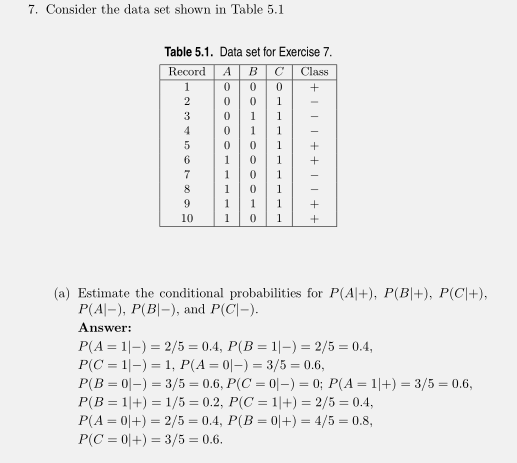
， ，

， ，

对于给定的计算：

所以y = -1

7.9考虑下表中的数据集



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

(b) 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

(c) 使用m估计方法(p=1/2且m=4)估计条件概率

(d) 同(b)，使用(c)中的条件概率

(e) 比较估计概率的两种方法。哪一种更好？为什么？

参考答案：

(a)

(b) 令 P(A=0，B=1，C=0)=K

故测试样本的类标记为 ”+”

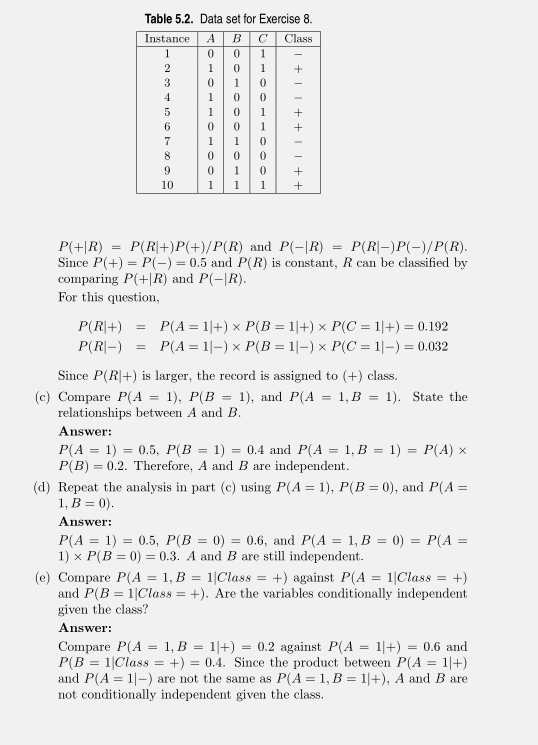
(c)

(d) 令 P(A=0，B=1，C=0)=K

故该测试集的类标签为 ”+”.

(e) 当其中有为0的条件概率时，用m估计方法更好.可以避免连乘后计算出的概率值为0.

7.10 考虑下表中的数据集



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

(b)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=1,B=1,C=1)的类标号。

(c)比较P(A=1),P(B=1)和P(A=1,B=1)，陈述A、B之间的关系。

(d)对P(A=1)，P(B=0)和P(A=1,B=0)重复(c)分析。

(e)比较P(A=1,B=1|类=+)与P(A=1|类=+)和P(B=1|类=+)。给定类+，变量A，B条件独立吗？

参考答案：

(a)

(b) 令测试集样本为 R：(A=1,B=1,C=1)，我们需要比较

由于P(+)=P(-)=0.5，且P(R)为常数，可以通过比较P(R|+)和P(R|-)的大小来判断R的类标号。

由于所以测试样本的类标号为“+”

(c)

因此，A和B相互独立

(d)

因此，A和B仍相互独立

(e)

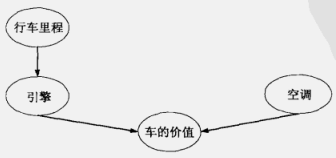
由于

所以，给定类“+”，变量A，B条件不独立

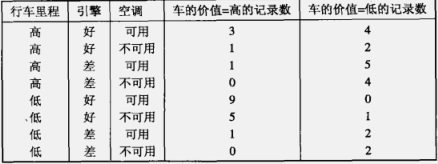
7.11 下图给出了表中的数据集对应的贝叶斯信念网络（假设所有的属性都是二元的）。

(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

参考答案：

(a) P(行车里程=高)=0.5

P(空调=可用)=0.625

P(引擎=好|行车里程=高)=0.5

P(引擎=好|行车里程=低)=0.75

P(车的价值=高|引擎=好，空调=可用)=0.75

P(车的价值=高|引擎=好，空调=不可用)=0.667

P(车的价值=高|引擎=差，空调=可用)=0.222

P(车的价值=高|引擎=差，空调=不可用)=0

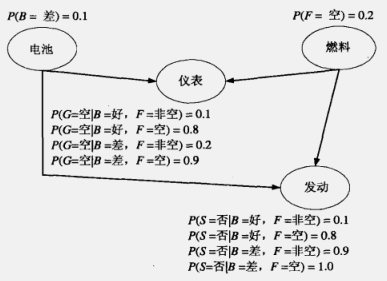
(b)

7.12 给定如下图所示的贝叶斯网络，计算下面概率.

(a) P(B=好，F=空，G=空，S=是)

(b) P(B=差，F=空，G=非空，S=否)

(c) 如果电池是差的，计算车发动起来的概率。



贝叶斯信念网络

解：

(a)

(b)

=空

(c)

7.13 (a) 假设本科生中抽烟的比例是15%，研究生中抽烟的比例是23%.如果大学生中研究生占1/5，其余是本科生，那么抽烟的学生是研究生的概率是多少？

(b) 根据(a)中的信息，随机选择一个大学里的学生，那么，该生是研究生或本科生的可能性哪个大？

(c) 同(b)，假设学生是个抽烟者.

(d) 假设30%的研究生住学生宿舍，只有10%的本科生住学生宿舍。如果一个学生抽烟又住宿舍，那么他（她）是研究生或本科生的可能性哪个大？可以假设住宿舍的学生和抽烟的学生相互独立。

参考答案：

(a) P(S|UG) = 0.15, P(S|G) = 0.23, P(G) = 0.2, P(UG) = 0.8

(b)由于P(UG)>P(G)，所以是本科生的可能性大.

(c)由于P(UG|S)>P(G|S)，所以是本科生的可能性大.

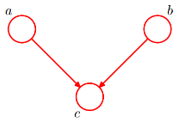
(d)

由于P(G|DS)>P(UG|DS)，故他（她）是研究生的可能性大.

7.14 贝叶斯网络的三种形式，并简述它们的特点

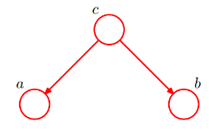
参考答案：

形式1：V型结构



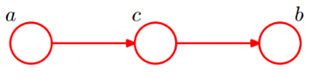
特点：当c未知时，a和b是相互独立的

形式2：同父结构



特点：给定c时，a和b是相互独立的

形式3：顺序结构



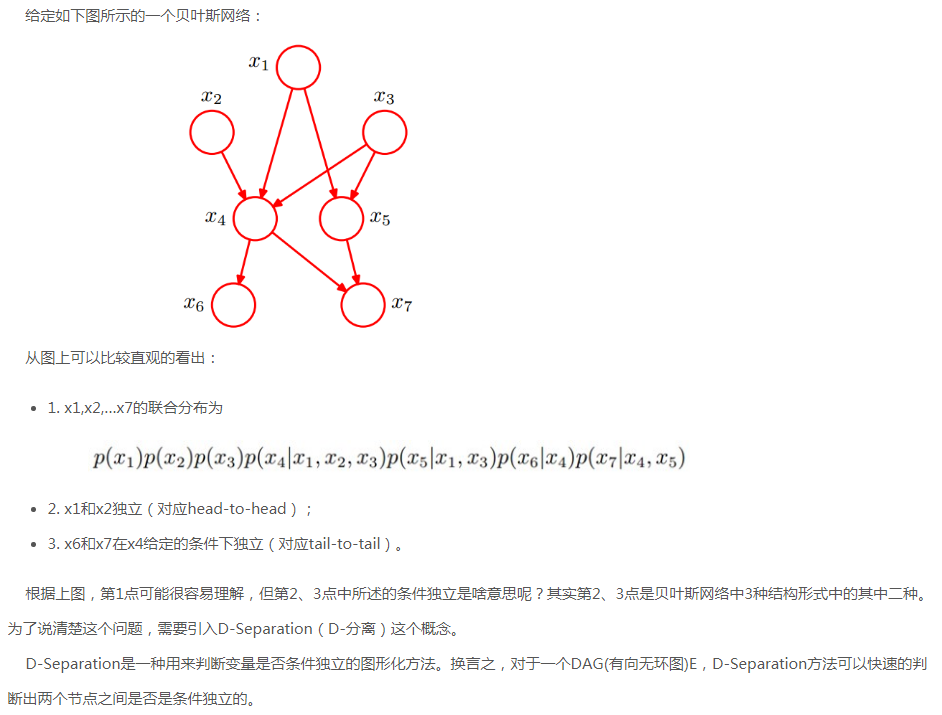
特点：给定c，a和b是相互独立的

7.15 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？



参考答案:

(a)

(b) 是

(c) 是