机器学习（西瓜书）

注　解

（第 7 章　贝叶斯分类器）

https://blog.csdn.net/jbb0523

前言

经常听人说南大周老师所著的《机器学习》（以下统称为西瓜书）是一本入门教材，是一本科普性质的教科书。在该书第十次印刷之际，周老师在“如何使用本书”中也提到“这是一本入门级教科书”。然而，本人读起来却感觉该书远不止“科普”“入门”那么简单，书中的很多公式需要思考良久方能推导，很多概念需要反复咀嚼才能消化。边读边想着是不是应该将自己学习时遇到的一些知识难点的解析分享出来，以帮助更多的人入门。自己的确也随手做过一些笔记，但由于怀疑这仅是自己的个别现象，毕竟读书期间，思考更多的是如何使用单片机、DSP、ARM、FGPA 等，而这些基本是不需要推导任何公式的，因此作罢。偶然间在周老师的新浪微博看到如下对话：

此时方知，可能“读不懂”并不是个别现象。因此决定写一本“西瓜书注解”或者称为“西瓜书读书笔记”，对自己研读西瓜书时遇到的“台阶”进行解释和推导，以帮助更多的人能够更快地进入到这个领域。另外，近期越来越强地意识到，扎扎实实地推导一些基础算法的公式，无论是对于理解算法本身机理还是进行学术研究，都是非常有必要的。

自己会根据个人学习进度和研究需要按章发布，不知道能不能坚持写完，加油！

毕竟自己也是一名初学者，所以可能一些概念解释并不完整、一些公式推导并不优美，甚至会存在错误，这是不可避免的，不接受漫骂，但欢迎将问题反馈给我，共同学习进步！

（网盘链接：https://pan.baidu.com/s/1QtEiNnk8jMzmbs0KPBN-_w）

第 7 章目录

第 7 章 贝叶斯分类器

提到贝叶斯，肯定就会联想到概率论课程中的贝叶斯公式，教材中常按如下形式给出：

$$P(B_i \mid A) = \frac{P(B_i)P(A \mid B_i)}{\sum_{j=1}^{n} P(B_j)P(A \mid B_j)}$$

其中$B_1, B_2, \ldots, B_n$互斥且构成一个完全事件。实际上，分母即为全概率公式：

$$P(A) = \sum_{j=1}^{n} P(B_j)P(A \mid B_j)$$

分子$P(A, B_i) = P(B_i)P(A \mid B_i)$为联合概率。机器学习中更常见的形式为

$$P(B \mid A) = \frac{P(B)P(A \mid B)}{P(A)}$$

贝叶斯公式的作用在于将$P(B \mid A)$的估计转化为估计$P(B)$和$P(A \mid B)$。

7.1 贝叶斯决策论

本节内容干货满满，虽然有 8 个公式，但整体以概念为主。

1、式(7.1)的解释

等号左侧$R(c_i \mid \boldsymbol{x})$表示将样本$\boldsymbol{x}$分类为$c_i$所产生的期望损失；也就是说，现在已知样本$\boldsymbol{x}$被分类为$c_i$，想知道的是这一事实的期望损失是多少。

那么，根据样本$\boldsymbol{x}$真实类别标记的不同，$\boldsymbol{x}$被分类为c_i所产生的损失肯定也不同。当样本$\boldsymbol{x}$真实类别标记为c_1时误分类为c_i所产生的损失为λ_{i1}，当样本$\boldsymbol{x}$真实类别标记为c_2时误分类为c_i所产生的损失为λ_{i2}，……，当样本$\boldsymbol{x}$真实类别标记为c_j时误分类为c_i所产生的损失为λ_{ij}……

因此，只要知道样本$\boldsymbol{x}$真实类别标记，就可以知道将其分类为c_i所产生的损失。**现在待求的条件风险$R(c_i \mid \boldsymbol{x})$是将样本$\boldsymbol{x}$分类为$c_i$所产生的期望损失，即所有损失的平均值**。根据数据期望的定义，每个损失乘以其对应的概率，再求和即可。

对于样本$\boldsymbol{x}$来说，它的标记为c_j的概率为$P(c_j \mid \boldsymbol{x})$，即已知$\boldsymbol{x}$的情况下，类别标记为$c_j$的后验概率（在给定$\boldsymbol{x}$的条件下，类别标记为$c_j$的(条件)概率），因此得式(7.1)：

$$R(c_i \mid \boldsymbol{x}) = \sum_{j=1}^{N} \lambda_{ij} P(c_j \mid \boldsymbol{x})$$

其中N为可能的类别标记个数。

2、式(7.2)的解释

式(7.1)是针对单个样本$\boldsymbol{x}$的，而式(7.2)则是针对整个数据集D所有样本的期望，即

$$R(h) = \mathbb{E}_{\boldsymbol{x}}\left[R(h(\boldsymbol{x}) \mid \boldsymbol{x})\right] = \sum_{\boldsymbol{x} \in D} R(h(\boldsymbol{x}) \mid \boldsymbol{x})P(\boldsymbol{x})$$

其中$P(\boldsymbol{x})$为样本$\boldsymbol{x}$出现的概率，且满足$\sum_{\boldsymbol{x} \in D} P(\boldsymbol{x}) = 1$；$h(\boldsymbol{x})$为判定准则$h : \mathcal{X} \mapsto \mathcal{Y}$对样本$\boldsymbol{x}$预测的类别标记；则$R(h)$表示判定准则$h$的总体风险。

3、式(7.3)的解释

解释一下符号“arg min”，其中“arg”是 “argument”的前三个字母，“min” 是“minimum”的前三个字母。维基百科中有“arg max”的解释：https://en.wikipedia.org/wiki/Arg_max，是其反义符号。概括起来，式(7.3)表示求出使目标函数$R(c \mid \boldsymbol{x})$最小的类别标记c，并将该c返回给$h^*(\boldsymbol{x})$作为输出，注意此处$\boldsymbol{x}$为已知常量，$R(c \mid \boldsymbol{x})$表达式参见式(7.1)。

此式共涉及到三个概念：**贝叶斯判定准则**(Bayes decision rule)，**贝叶斯最优分类器**(Bayes optimal classifier)，**贝叶斯风险**(Bayes risk)。式(7.3)即为贝叶斯判定准则，所得分类器$h^*(\boldsymbol{x})$为贝叶斯最优分类器，此时的总体风险$R(h^*)$称为贝叶斯风险。

书中提到：“$1-R(h^*)$反映了分类器所能达到的最好性能，即通过机器学习所能产生的模型精度的理论上限”，个人理解这里是不是想表达$1-R(h^*)$是贝叶斯决策论理论框架下所能达到的模型精度的理论上限？毕章这句话出现在本章，因此应该仅适用于本章的理论背景，因为通常 SVM 的分类精度比朴素贝叶斯分类器（详见 7.3 节）要好很多，当然朴素贝叶斯分类器并不能代表贝叶斯决策论，因为它是在一些假设条件下求出的后验概率。

4、式(7.4)的解释

该式所表达的即为 0/1 损失，也就是说分类正确时$(i=j)$损失为 0，否则损失为 1。

5、式(7.5)的推导

将式(7.4)代入式(7.1)，得

$$
\begin{aligned}
R(c_i \mid \boldsymbol{x}) &= \sum_{j=1}^{N} \lambda_{ij} P(c_j \mid \boldsymbol{x}) \\
&= \sum_{j=1}^{i-1} P(c_j \mid \boldsymbol{x}) + \sum_{j=i+1}^{N} P(c_j \mid \boldsymbol{x}) \\
&= \sum_{j=1}^{i-1} P(c_j \mid \boldsymbol{x}) + P(c_i \mid \boldsymbol{x}) + \sum_{j=i+1}^{N} P(c_j \mid \boldsymbol{x}) - P(c_i \mid \boldsymbol{x}) \\
&= \sum_{j=1}^{N} P(c_j \mid \boldsymbol{x}) - P(c_i \mid \boldsymbol{x}) \\
&= 1 - P(c_i \mid \boldsymbol{x})
\end{aligned}
$$

其中第 2 个等号利用了式(7.4)，最后一个等号利用了$\sum_{j=1}^{N} P(c_j \mid \boldsymbol{x}) = 1$ 。

6、式(7.6)的解释

式(7.3)的贝叶斯最优分类器是最小化式(7.5)，因此等价于最大化$P(c \mid \boldsymbol{x})$。

7、判别式模型与生成式模型

可以简单理解为判别式模型直接建模后验概率$P(c \mid \boldsymbol{x})$来预测类别c，而生成式模型则是对联合概率$P(\boldsymbol{x}, c)$建模来预测类别c。具体来说：

对于判别式模型来说，就是已知$\boldsymbol{x}$的条件下判别其类别标记，即求后验概率$P(c \mid \boldsymbol{x})$，前面介绍的决策树、BP 神经网络、支持向量机都属于判别式模型的范畴，其中尤其以 3.3 节介绍的对率回归最为直接，详见式(3.23)和式(3.24)。

对于生成式模型来说，理解起来比较抽象。思考两个问题：

(1)对于数据集D来说，其中的样本是如何生成的呢？

答：当然是按照联合概率分布$P(\boldsymbol{x}, c)$采样而得，也可以描述为根据$P(\boldsymbol{x}, c)$生成的。

(2)若已知样本$\boldsymbol{x}$和联合概率分布$P(\boldsymbol{x}, c)$，如何预测类别c呢？

答：在样本$\boldsymbol{x}$已知的情况下，可以分别求出$P(\boldsymbol{x}, c_1), P(\boldsymbol{x}, c_2), \ldots, P(\boldsymbol{x}, c_N)$，即生成样本$(\boldsymbol{x}, c_1), (\boldsymbol{x}, c_2), \ldots, (\boldsymbol{x}, c_N)$的概率（特征$\boldsymbol{x}$已知，可变部分只有类别$c$），当然是选择与$\boldsymbol{x}$联合概率最大的类别标记，即

$$h^*(\boldsymbol{x}) = \arg\min_{c \in \mathcal{Y}} P(\boldsymbol{x}, c)$$

因此，之所以称为“生成式”模型，是因为所求概率$P(\boldsymbol{x}, c)$是样本生成的概率。

8、式(7.7)和式(7.8)的解释

这两个式子本质上相同，因为学过概率论的人都知道以下概率关系：

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

前面在分析生成式模型的概念时提到，若已知样本$\boldsymbol{x}$和联合概率分布$P(\boldsymbol{x}, c)$，预测类别c时选择与$\boldsymbol{x}$联合概率最大的类别标记，而这实际上与选择使后验概率$P(c \mid \boldsymbol{x})$最大的类别标记等价；因为给定样本$\boldsymbol{x}$，概率$P(\boldsymbol{x})$为常量（而且概率肯定为大于零），因此，最大化$P(\boldsymbol{x}, c)$和$\frac{P(\boldsymbol{x},c)}{P(\boldsymbol{x})}$等价（常量并不影响极值点的存在位置，例如$f(\boldsymbol{x}) = x^2$与$g(\boldsymbol{x}) = 3x^2$极小值点均为$\boldsymbol{x} = 0$），即最大化后验概率$P(c \mid \boldsymbol{x})$与最大化联合概率$P(\boldsymbol{x}, c)$等价。

由式(7.8)，估计$P(c \mid \boldsymbol{x})$的问题转化为如何基于训练数据D来估计$P(c)$和$P(\boldsymbol{x} \mid c)$。

基于训练数据D来估计$P(c)$相对容易，以第 76 页的表 4.1 西瓜数据集 2.0 为例：

$$\text{P(好瓜=是)} = \frac{8}{17}, \quad \text{P(好瓜=否)} = \frac{9}{17}$$

基于训练数据D来估计$P(\boldsymbol{x} \mid c)$则比较困难，仍第 76 页的表 4.1 西瓜数据集 2.0 为例：

若估计当$\boldsymbol{x}$ =(色泽=青绿，根蒂=蜷缩，敲声=浊响，纹理=清晰，脐部=凹陷，触感=硬滑)，c =(好瓜=是)的概率$P(\boldsymbol{x} \mid c)$，则在前 8 个“好瓜=是”的样本中，仅编号为 1 的瓜的特征等于$\boldsymbol{x}$，此时若将$P(\boldsymbol{x} \mid c)$估计为$\frac{1}{8}$，这实际上是不可靠的，因为此时的样本集合太小了；再例如若$\boldsymbol{x}$ =(色泽=青绿，根蒂=蜷缩，敲声=浊响，纹理=清晰，脐部=凹陷，触感=软粘)，则在前 8 个“好瓜=是”的样本中根本没出现此$\boldsymbol{x}$，但“未被观测到”与“出现概率为零”通常是不同的（参见西瓜书本节最后一段话）。

这两个式子就是**贝叶斯定理**，它将估计**后验概率**$P(c \mid \boldsymbol{x})$转化为估计**先验概率**$P(c)$和**似然概率**$P(\boldsymbol{x} \mid c)$，而这正是朴素贝叶斯分类器、半朴素贝叶斯分类器等的理论基础。

9、先验概率和条件概率、后验概率和似然概率

对于条件概率$P(A \mid B)$，表示事件A在另外一个事件B已经发生条件下发生的概率，读作“在B的条件下A的概率”（参见百度百科词条 条件概率）；而$P(A)$称为先验概率，即在无任何已知条件下事件A的概率。

若给定B，则$P(A \mid B)$称为后验概率；此时，对于多个候选A，选择使后验概率$P(A \mid B)$最大的那一个，称为最大后验(maximum a posteriori, MAP)估计。

若给定A，则$P(A \mid B)$称为似然概率；此时，对于多个候选B，选择使似然概率$P(A \mid B)$最大的那一个，称为最大似然(maximum likelihood, ML)估计。（即 7.2 节）

后验概率$P(A \mid B)$可以理解为已知B之后（即后验），A发生的(条件)概率；而似然概率$P(A \mid B)$可以理解为已知A之后，猜猜它更像是（即似然）在哪个B的条件下得到的。

一般来说后验概率是 P(果|因)，似然概率是 P(因|果)，而并不像本条解释中的A和B随心指定已知哪个。机器学习任务中，已知样本特征向量$\boldsymbol{x}$来预测样本类别y，因此$P(y \mid \boldsymbol{x})$是后验概率，$P(\boldsymbol{x} \mid y)$是似然概率。

另外，$P(A \mid B)$和$P(B \mid A)$的差别是很大的，正如百度百科词条[条件概率]所述，很多人经常会犯此错误，误以为$P(A \mid B)$大致等于$P(B \mid A)$，此即条件概率的谬论。例如，A表示事件喝了口开水，B表示事件被烫着了，则$P(A \mid B)$表示该人被烫着了的条件下喝了口开水的概率（已知被烫着了，求此时喝了口开水的概率），而$P(B \mid A)$表示该人喝了口开水的条件下被烫着的概率（已知喝了口开水，求此时被烫着的概率），显然这两个概率是不同的。

7.2 极大似然估计

极大似然估计(Maximum Likelihood Estimation, MLE)实际上在概率论课程中就介绍过过了，本人还记得本科时的概率论与数理统计课程期末考试就有一道 MLE 的试题。

频率主义学派和贝叶斯学派的参数估计解决方案分别是极大似然估计和贝叶斯估计，参见[张连文. 贝叶斯网络引论[M]. 科学出版社，2006. (http://www.cse.ust.hk/bnbook/)]的第 144-150 页（7.2 节单参数最大似然估计和 7.3 节单参数贝叶斯估计，尤其是第 149 页的例 7.2 和例 7.3；该书已绝版，pdf 电子版参见给出的链接）。

对于极大似然法，关键在于找出似然概率的表达式。式(7.9)给出了最原始的似然表达式，注意该式中$\boldsymbol{\theta}_c$是待求模型参数，极大似然要做的是寻找当前数据集更像是由哪组模型参数$\boldsymbol{\theta}_c$生成的；连乘容易造成下溢，这是由于式(7.9)的概率均小于 1，例如 100 个 0.1 连乘，则乘积非常小，此时计算机可能无法表达如此小的数（仅为举例，请具体问题具体分析），此即下溢（下溢的反义词为上溢，例如 100 个 10 连乘，则乘积非常大）；解决下溢的常用方法就是取对数，将连乘操作变为连加操作，此即式(7.10)；得到了（对数）似然的表达式之后，在$\boldsymbol{\theta}_c$的可行域内最大化似然函数即得极大似然估计，此即式(7.11) 。

例：投掷硬币 5 次，结果依次是正面、正面、反面、正面、反面，试基于此观察结果估计硬币正面朝上的概率θ。（改编自[张连文. 贝叶斯网络引论[M]. 科学出版社, 2006]例 7.1）

解：设正面朝上的概率为θ，各次投掷结果相互独立，则似然为

$$\begin{aligned} L(\theta) =& \theta \cdot \theta \cdot (1-\theta) \cdot \theta \cdot (1-\theta) \\ =& \theta^3 (1-\theta)^2 \end{aligned}$$

对数似然为（实际上，本例不需要去求对数似然）

$$LL(\theta) = \ln L(\theta) = 3\ln\theta + 2\ln(1-\theta)$$

求导并令导数等于零

$$\begin{aligned} \frac{\partial LL(\theta)}{\partial \theta} =& \frac{\partial (3\ln\theta + 2\ln(1-\theta))}{\partial \theta} \\ =& \frac{3}{\theta} - \frac{2}{1-\theta} \\ =& \frac{3-5\theta}{\theta(1-\theta)} = 0 \end{aligned}$$

解得$\theta = \frac{3}{5}$。也就是说，基于观测结果对概率θ的极大似然估计就是正面出现的比例。

7.3 朴素贝叶斯分类器

本节内容通俗易懂，且以西瓜数据集 3.0 为例展示了朴素贝叶斯分类器的训练细节，涵盖了连续属性和离散属性的处理方法，还讲述了拉普拉斯修正方法，可谓面面俱到。

个人感觉，本节的核心包括三点：(1)朴素贝叶斯分类器的概念；(2)朴素贝叶斯分类器的连续属性和离散属性条件概率估计方法；(3)拉普拉斯修正。

朴素贝叶斯分类器即式(7.15)，相比于式(7.14)少了$P(\boldsymbol{x})$，这是因为$P(\boldsymbol{x})$与类标记c无关，而常量并不影响极值点的存在位置，例如$f(x)=x^2$与$g(x)=3x^2$极小值点位置相同，均为$x=0$。式(7.15)是多个概率连乘的形式，可以取对数转化为连加的形式以防止下溢，详见本文档 7.2 节的解释中的第三段。

离散变量的概率估计较为简单，直接按频率计数即可，如式(7.16)的先验概率估计和式(7.17)的条件概率估计，注意式中的$|\cdot|$表示求集合元素的个数。而连续变量必须假设一种概率分布，常用的就是高斯分布；注意：此处概率密度函数并非概率，即取值范围不是 0 到 1 之间，对于高斯概率密度函数来说，式(7.18)的$p(x_i \mid c) \in (0, \frac{1}{\sqrt{2\pi}\sigma_{c,i}}]$，其中当$x_i=\mu_{c,i}$时取得最大值$\frac{1}{\sqrt{2\pi}\sigma_{c,i}}$，进一步地当$\sigma_{c,i}<\frac{1}{\sqrt{2\pi}}$时，$\frac{1}{\sqrt{2\pi}\sigma_{c,i}}>1$，即会出现$p(x_i \mid c)>1$的情形。

在离散变量的概率估计中，由于训练集样本的不充分性，可能会出现分子为零的情形，但“未被观测到”与“出现概率为零”是不同的，因此要进行平滑(smoothing)，常用的是拉普拉斯修正，如式(7.19)和式(7.20)所示。从式(7.19)和式(7.20)中可以看出，拉普拉斯修正实质上假设了属性值与类别均匀分布，这类似于[张连文. 贝叶斯网络引论[M]. 科学出版社, 2006]例 7.3 引入的先验，只是平滑值没有那么大而已（例 7.3 中，分子 100 和分母 200）。

注意，本节例子计算有笔误，第 152 页的第 9 个等式$P_{凹陷|是}=\frac{5}{8}$。截止到目前（第 28 次印刷），西瓜书官方勘误修订仅在第 8 次印刷时修正了第 3 个等式$P_{蜷缩|是}$，但第 9 个等式仍未修正（若修正第 84 页的西瓜数据集 3.0 也可，但亦未修正）。

最后再次强调，式(7.17)所得$P(x_i \mid c) \in [0,1]$为条件概率，但式(7.18)所得$p(x_i \mid c)$为条件概率密度而非概率，其值并不在局限于区间$[0,1]$之内。

7.4 半朴素贝叶斯分类器

一般来说，朴素贝叶斯分类器就足够了，而且更复杂的模型并不一定会得到更好的泛化性能。但科研无止境，且本节的几个思想很值得在科研中借鉴。

1、式(7.21)的解释

在朴素贝叶斯分类器中，假设每个属性独立地对分类结果发生影响，因此

$$P(\boldsymbol{x} \mid c)=\prod_{i=1}^{d} P(x_i \mid c)$$

“独依赖估计”策略假设每个属性在类别之外最多仅依赖一个其他属性，因此

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

其中pa_i为属性x_i所依赖的属性。前者只依赖类别c，而后者同时依赖c和pa_i。

根据贝叶斯公式，类似于式(7.14)到式(7.15)，舍掉分母常数项$P(\boldsymbol{x})$，即得式(7.21)

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

在上一节中详细展示了$P(x_i \mid c)$，即先挑出类别为c的样本，若是离散属性则按计数法估计$P(x_i \mid c)$，若是连续属性则求这些样本的均值和方差，按高斯分布估计$p(x_i \mid c)$。现在估计$P(x_i \mid c, pa_i)$，则先挑出类别为c、属性x_i所依赖的属性值为pa_i的样本，剩下步骤与估计$P(x_i \mid c)$时相同。

2、图 7.1 的解释

图中箭头指向表示依赖关系，如所有属性$x_1, x_2, \ldots, x_d$均依赖于类别变量y，因此存在由y指向$x_1, x_2, \ldots, x_d$的箭头。图(a)中的 NB 各属性之间不存在依赖关系，因此无箭头；图(b)假设所有属性均依赖于x_1（x_1是超父属性），因此存在由x_1指向$x_2, x_3, \ldots, x_d$的箭头；图(c)中属性之间依赖关系是树形结构（不存在回路，则每个结点有且仅有一个父结点）。

3、式(7.22)的解释

该式写为如下形式可能更容易理解：

$$I(x_i, x_j \mid y) = \sum_{n=1}^{N} P(x_i, x_j \mid c_n) \log \frac{P(x_i, x_j \mid c_n)}{P(x_i \mid c_n)P(x_j \mid c_n)}$$

其中$i, j = 1, 2, \ldots, d$且$i \neq j$，N为类别个数。该式共可得到$d(d-1)/2$个$I(x_i, x_j \mid y)$，即每对(x_i, x_j)均有一个条件互信息$I(x_i, x_j \mid y)$。

4、TAN 算法的解释

该算法共分四步，第 1 步即式(7.22)，为每对(x_i, x_j)得到一个条件互信息$I(x_i, x_j \mid y)$；第 2 步以$I(x_i, x_j \mid y)$为权重，构建完全图，即每个属性为结点，结点之间边的权重由第 1 步计算所得；第 3 步调用最大带权生成树算法，去除第 2 步完全图中的部分边，生成一颗树，并且该树所有边的权重之和应该是所有可能生成树中最大的，即最大带权的生成树，此时的树仍是无向图，挑选根变量使之变为有向图（从根结点开始，由父结点指向孩子结点，至于如何挑选根变量，书中没说，可以将树中所有结点均作为根结点，最后将学习结果做一次集成即可，集成学习参见第 8 章）；第 4 步加入结点y，即得类似于图 7.1(c)所示 TAN。

可能并不是所有人都熟悉最大带权生成树，但学过数据结构课程的人应该都知道其相反的概念最小生成树：一个有n个结点的连通图的生成树是原图的极小连通子图，且包含原图中的所有n个结点，并且有保持图连通的最少的边。最小生成树可以用 kruskal（克鲁斯卡尔）算法或 prim（普里姆）算法求出。（摘自百度百科词条最小生成树）

5、式(7.23)的推导

贝叶斯定理式(7.8)将联合概率$P(\boldsymbol{x}, c)$换为等价形式$P(\boldsymbol{x} \mid c)P(c)$；实际上，将向量$\boldsymbol{x}$拆开，把$P(\boldsymbol{x}, c)$写为$P(x_1, x_2, \ldots, x_d, c)$形式，此时

$$\begin{aligned} P(\boldsymbol{x}, c) =& P(x_1, x_2, \ldots, x_d, c) \\ =& P(x_1, x_2, \ldots, x_d \mid c)P(c) \\ =& P(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d \mid c, x_i)P(c, x_i) \end{aligned}$$

其中第 2 个等号即$P(\boldsymbol{x}, c) = P(\boldsymbol{x} \mid c)P(c)$，而第 3 个等号是是第 2 个等号的推广；更通

俗地来说，$P(A,B)=P(A\mid B)P(B)$，对于第 2 个等号，$A=\boldsymbol{x},B=c$，而对于第 3 个等号，$A=(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_d),B=(c,x_i)$。

类似于式(7.14)的属性条件独立性假设，则

$$P(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_d\mid c,x_i)=\prod_{\substack{j=1\\j\neq i}}^{d}P(x_j\mid c,x_i)$$

实际上，根据式(7.25)，若不考虑平滑项，$P(x_j\mid c,x_i)=1$，因此在连乘中不起作用，可去除上式中的$j\neq i$约束（当$j=i$时，式(7.25)的$|D_{c,x_i,x_j}|$与$|D_{c,x_i}|$相等），即

$$P(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_d\mid c,x_i)=\prod_{j=1}^{d}P(x_j\mid c,x_i)$$

综上所述：

$$\begin{aligned}P(c\mid\boldsymbol{x})&=\frac{P(\boldsymbol{x},c)}{P(\boldsymbol{x})}=\frac{P(c,x_i)P(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_d\mid c,x_i)}{P(\boldsymbol{x})}\\&\propto P(c,x_i)P(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_d\mid c,x_i)\\&=P(c,x_i)\prod_{j=1}^{d}P(x_j\mid c,x_i)\end{aligned}$$

上式是将属性x_i作为超父属性的，AODE 尝试将每个属性作为超父来构建 SPODE，然后将那些具有足够训练数据支撑的 SPODE 集成起来作为最终结果。具体来说，对于总共d个属性来说，共有d个不同的上式，集成直接求和即可，因为对于不同的类别标记c均有d个不同的上式；至于“足够训练数据支撑的 SPODE”条件，注意式(7.24)和式(7.25)的使用到了$|D_{c,x_i}|$与$|D_{c,x_i,x_j}|$，若D_{x_i}集合中样本数量过少，则$|D_{c,x_i}|$与$|D_{c,x_i,x_j}|$将会更小，因此在式(7.23)中要求D_{x_i}集合中样本数量不少于m'。

6、式(7.24)和式(7.25)的解释

这两个式子本身就是频率计数，它们都应用了类似于式(7.19)和式(7.20)的拉普拉斯修正。其中，式(7.24)分母之所以加$N\times N_i$是因为在样本集合D（分母）中类别c与属性x_i（分子）的可能组合个数共有$N\times N_i$种，式(7.25)分母之所以加N_j是因为样本集合D_{c,x_i}（分母）中属性x_j（分子）的可能取值个数共有N_j种（其实这个与是否在样本集合D_{c,x_i}中没关系）。具体计算参见书中接下来的例子，琢磨一下即可，此处只想通过这两个式子进一步阐明拉普拉斯修正的使用方法。

7.5 贝叶斯网

贝叶斯网是一种经典的概率图模型，西瓜书第 14 章专门介绍概率图模型。可以发现，本节篇幅相比于其它节的篇幅显得格外长，共占了约 6 页的版面，这对于西瓜书每章控制在约 25 页来说可以算是长篇大论了。然而，用 6 页内容介绍贝叶斯网模型，这甚至算不上蜻蜓点水。但也许正如西瓜书在第十次印刷之际作者写的《如何使用本书》中所述，“不要指望本书是无所不包、‘从入门到精通’的书籍”，“本书的主要目的就是为读者提供一张‘初级地形图’，给初学者‘指路’”，“根据本书提供的‘地形图’，读者若渴望对某个知识点进

一步探究，‘按图索骥’应该无太大困难”。

学习贝叶斯网，除少数专门理论研究人员除外，更多的人是在自己的问题上将贝叶斯网作为一个工具来使用。因此，学习贝叶斯网时重在理解其概念，然后能够使用网上公开的软件包即可，《贝叶斯网络综合应用》(https://blog.csdn.net/jbb0523/article/details/79438202)给出了一个贝叶斯网络使用例子，可供学习参考。

本节内容比较生涩难懂，但又无法特别展开，因此仅介绍一些公式、插图等内容。

1、式(7.26)的解释

注意本式上方的一段话：“给定父结点集，贝叶斯网假设每个属性与它的非后裔属性独立”，该式正是基于此而来。本式给出的是贝叶斯网所有结点的联合概率分布，以图 7.2 为例，结点x_1无父结点，因此在式(7.26)中为$P(x_1)$；结点x_2亦无父结点，因此在式(7.26)中为$P(x_2)$；结点x_3父结点为x_1，因此在式(7.26)中为$P(x_3 \mid x_1)$；结点x_4父结点为x_1, x_2，因此在式(7.26)中为$P(x_4 \mid x_1, x_2)$；结点x_5父结点为x_2，因此在式(7.26)中为$P(x_5 \mid x_2)$，故图 7.2 的联合概率定义为：

$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_2)P(x_3 \mid x_1)P(x_4 \mid x_1, x_2)P(x_5 \mid x_2)$$

2、式(7.27)的解释

该式不能基于概率论去推导，而是应该基于式(7.26)去推导。式(7.27)针对图 7.3 中的 V 型结构，这也是一个贝叶斯网，因此根据式(7.26)，V 型结构三个节点的联合概率为

$$P(x_1, x_2, x_4) = P(x_1)P(x_2)P(x_4 \mid x_1, x_2)$$

这正是式(7.27)中第 2 个等号的由来。第 1 个等号就是概率论中的边际化，即要消去联合概率中的某个变量，只需对该变量积分（或求和）即可。而$\sum_{x_4} P(x_4 \mid x_1, x_2) = 1$，因此有第 3 个等号。

其中$\sum_{x_4} P(x_4 \mid x_1, x_2) = 1$这个结论是显而易见的，因为在给定$x_1, x_2$下，遍历了$x_4$所有状态。若想不明白，那就基于第 76 页的表 4.1 西瓜数据集 2.0 举个例子，可假设x_1, x_2分别为属性色泽中的青绿和属性根蒂中的蜷缩，x_4为纹理属性的可能取值（清晰、稍糊、模糊），显然：P(纹理=清晰|色泽=青绿, 根蒂=蜷缩) + P(纹理=清晰|色泽=青绿, 根蒂=蜷缩) + P(纹理=清晰|色泽=青绿, 根蒂=蜷缩) = 1 ，各概率计算参见 7.3 节各离散属性概率的估计。

本式推导的关键在于为什么$P(x_1, x_2, x_4) = P(x_1)P(x_2)P(x_4 \mid x_1, x_2)$，因为根据概率论有$P(x_1, x_2, x_4) = P(x_1, x_2)P(x_4 \mid x_1, x_2)$。

3、图 7.4 的解释

注意，摘除某个变量后，所有以其为端点的边全部删除。例如，摘除x_1后，与x_1相连的三条边全部删除，此时x_3与其余部分完全隔开。

由图 7.2 的贝叶斯网到图 7.4 的道德图，目标是为了分析有向图中变量间的条件独立性，使用的方法是“有向分离”。

英文 D-separation 和 moralization 在论文中经常见到，应该理解其含义。

4、“评分搜索”的解释

在 7.5.2 节，开篇第一段就提到了贝叶斯网络结构学习的常用办法：评分搜索。简单来说，将所有可能的贝叶斯网络结构当作定义域，每种结构都用事先定义好的评分函数映射得到一个评分，即$y = f(x)$形式，x为自变量（即某种贝叶斯网络结构），函数值y为该结构的

评分，贝叶斯网络结构学习的过程相当于在定义域内寻找函数最大值的过程。

接下来的式(7.28)到式(7.31)都在介绍评分函数，了解概念即可，不必细究。

5、式(7.32)的解释

贝叶斯网络学习包括结构学习和参数学习。前面都在介绍结构学习，参数学习（即图 7.2 中的 CPT 表）相对简单，简单情况下可自行计算，参见《贝叶斯网络参数学习（基于 FullBNT-1.0.4 的 MATLAB 实现）》（https://blog.csdn.net/jbb0523/article/details/78915828）。

6、贝叶斯网推断的解释

在 7.5.3 节基于吉布斯采样介绍贝叶斯网推断，更多推断内容参见 14.5 节。

若仅想得到各节点之间的关系，贝叶斯网络学习完成后就可以了；但更多的情况是要根据学习结果进行类似于对率回归、决策树、神经网络、支持向量机、朴素贝叶斯分类器等所做的预测，求出后验概率$P(y \mid \boldsymbol{x})$。

在贝叶斯网中，并不区分哪个节点是类别变量y，哪个节点是特征变量x_i。我们根据已知变量观测值（称为“证据”）去推测待查询变量的过程称为“推断”。对于普通的分类任务来说，已知变量一般就是特征$\boldsymbol{x}$，待查询变量一般就是类别y。

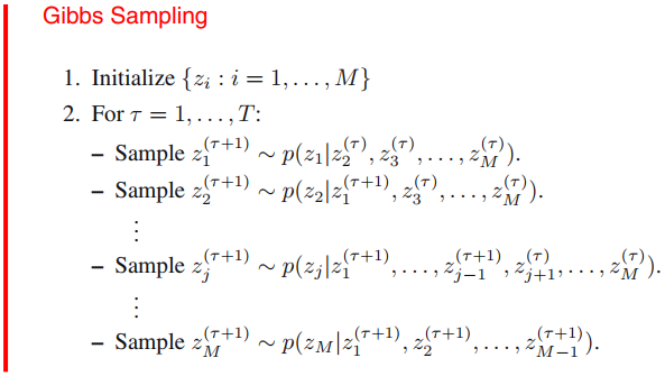
7、式(7.33)的解释

该式为吉布斯采样算法的一部分，稍后解释图 7.5。

此处想说的是概率$P(\mathbf{Q}=\mathbf{q} \mid \mathbf{E}=\mathbf{e})$类似于后验概率$P(y \mid \boldsymbol{x})$。

8、图 7.5 的解释

首先，去除图中的第 11 行到第 13 行，才是标准的吉布斯采样算法，该三行仅是利用吉布斯采样结果得到式(7.33)的后验概率而已。看一下机器学习经典专著《Pattern Recognition and Machine Learning》（简称 PRML）中的吉布斯采样算法：



可以看出，第 2 步的 **For** 循环即图 7.5 中第 3 行到第 14 行的 **For** 循环（去除图中的第 11 行到第 13 行），区别在于 PRML 中将图 7.5 中的第 4 行到第 10 行的 **For** 循环逐行给出的。

在图 7.5 中，大写黑体字母表示变量，如$\mathbf{E}, \mathbf{Q}, \mathbf{Z}$，相应的小写字母$\mathbf{e}, \mathbf{q}, \mathbf{z}$表示变量的取值，例如第 76 页的表 4.1 西瓜数据集 2.0 中，属性纹理为变量，共有三个可能取值（清晰、稍糊、模糊）。

第 5 行是除去变量Q_i外的其他变量（每次仅更新一个变量），第 6 行则是相应的变量取值；第 7 行根据贝叶斯网B计算分布$P_B(Q_i \mid \mathbf{Z}=\mathbf{z})$，也就是根据除去变量$Q_i$外的所有其他变量来计算条件概率$P_B(Q_i \mid \mathbf{Z}=\mathbf{z})$（例如，假设此处$Q_i$表示属性纹理，则该概率表示根据其它变量取值计算条件概率分布，包含三个概率值，分别是纹理为清晰、稍糊、模糊时的条件概率）；第 8 行根据$P_B(Q_i \mid \mathbf{Z}=\mathbf{z})$采样获取$Q_i$取值（假如根据清晰、稍糊、模糊三

个概率选择其中一个，每个属性值被选到的概率对应其概率$P_B(Q_i \mid \mathbf{Z}=\mathbf{z})$)；第 9 行更新待查询变量中$Q_i$的取值（如果只有一个待查询变量$Q_i$，那么第 4 行到第 10 行的 **For** 循环实际也仅包含一次循环）。

第 4 行到第 10 行的 **For** 循环就是每次更新一个待查询变量，该 **For** 循环结束则所有待查询变量被更新一次;然后第 3 行到第 14 行的 **For** 循环结束则所有待查询变量被更新T轮。我们要估计后验概率$P(\mathbf{Q}=\mathbf{q} \mid \mathbf{E}=\mathbf{e})$，而吉布斯采样算法对$\mathbf{Q}$随机赋初值（第 2 行），那么每执行一次第 3 行到第 14 行的 **For** 循环则$\mathbf{Q}$就会被更新一次，迭代T次$\mathbf{Q}$会被更新T次，在这T次循环当中，$\mathbf{Q}$的取值可能会等于待查询的$\mathbf{q}$（例如$\mathbf{Q}$为是否好瓜、甜度如何，而$\mathbf{q}$为坏瓜、不甜），假设有n_q次等于待查询的$\mathbf{q}$，则待解后验概率$P(\mathbf{Q}=\mathbf{q} \mid \mathbf{E}=\mathbf{e})=\frac{n_q}{T}$。

吉布斯采样算法能有效的依据是，经过若干次第 3 行到第 14 行的 **For** 循环后，$\mathbf{Q}$各种可能取值的出现概率收敛于一个平稳分布，该分布恰是$P(\mathbf{Q}=\mathbf{q} \mid \mathbf{E}=\mathbf{e})$。例如（坏瓜、不甜）的后验概率为 0.2，则第 3 行到第 14 行的 **For** 循环执行 100 次当中，会有约 20 次得到的$\mathbf{q}$取值为（坏瓜、不甜）。

吉布斯采样是马尔可夫链蒙特卡罗（Markov Chain Monte Carlo, MCMC）方法的一种实现，是 MCMC 方法的代表算法 Metropolis-Hastings(简称 MH)的一种特殊形式，具体到 14.5.1 节再细聊这其中的关系。

7.6 EM 算法

本节介绍非常流行的 EM 算法，的确什么都说了，但初学者读完本节之后的感觉肯定是觉得什么都没说。EM 算法只是一个框架，所以本节的几个公式都是神龙见首不见尾高度抽象的表达式，要想理解掌握 EM 算法最好看几个具体例子，例如 9.4.3 节的高斯混合聚类。

简单来说，EM 算法就是 E 步和 M 步的交替迭代，至于 E 步和 M 步的具体形式那就要具体问题具体分析了。交替迭代的求解思想很常用，例如待求解问题有两组参数 A 和 B，但却不能同时求解，而且又相互约束，这时就可以初始化参数 A，求解参数 B，待求得参数 B 后再更新参数 A 的值，然后再求解参数 B，再更新参数 A……

交替迭代的关键问题是多次交替迭代后算法会不会收敛到某个解，而这就是大牛们的舞台了。作为一个普通人，一般都是通过更好的实验结果来佐证算法的收敛性，所以也只能把论文发表在普通的期刊上；而大牛们可以通过一系列假设，证明算法在某条件下是收敛的，然后再配以实验结果，最后把论文发表在顶级期刊上。

1、式(7.34)的解释

本式等号右侧$\ln P(\mathbf{X}, \mathbf{Z} \mid \Theta)$就是类似式(7.10)的似然概率，但等号左侧$LL(\Theta \mid \mathbf{X}, \mathbf{Z})$与式(7.10)表达方式不同：为了表示对$\Theta$做最大似然估计是在已知$\mathbf{X}, \mathbf{Z}$条件下，使用了类似于条件概率的表达方式。这里不用管它，就是一种表示符号而已，用$LL(\Theta)$亦可。

2、式(7.35)的解释

在式(7.34)的解释中提到，对Θ做最大似然估计需要已知$\mathbf{X}, \mathbf{Z}$。但是$\mathbf{Z}$是隐变量，也就是未知的，因此只需考虑已知$\mathbf{X}$即可，即最大化似然$\ln P(\mathbf{X} \mid \Theta)$。

概率$P(\mathbf{X} \mid \Theta)$可以对概率$P(\mathbf{X}, \mathbf{Z} \mid \Theta)$做边际化(marginalization)而得，类似于式(7.27)解释中的$\sum_{x_4} P(x_4 \mid x_1, x_2)=1$，此处有$P(\mathbf{X} \mid \Theta)=\sum_{\mathbf{Z}} P(\mathbf{X}, \mathbf{Z} \mid \Theta)$。

3、式(7.36)的解释

数学期望（参见百度百科词条 数学期望）公式区分离散型变量和连续型变量，此处以离散型变量来解释式(7.36)。

设隐变量$\mathbf{Z}$可能的取值个数为N，分别记为$\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_N$。在已知$\mathbf{X}$、$\Theta = \Theta^t$以及$\mathbf{Z} = \mathbf{Z}_i$时，可以计算出条件概率$P(\mathbf{Z}_i \mid \mathbf{X}, \Theta^t)$；而若已知$\mathbf{X}$和$\mathbf{Z} = \mathbf{Z}_i$，则可由式(7.34)得到对数似然$LL(\Theta \mid \mathbf{X}, \mathbf{Z}_i) = \ln P(\mathbf{X}, \mathbf{Z}_i \mid \Theta)$，因此可得式(7.36)：

$$\begin{aligned}\mathbb{E}_{\mathbf{Z}|\mathbf{X},\Theta^t} LL(\Theta \mid \mathbf{X}, \mathbf{Z}) &= \sum_{i=1}^{N} P(\mathbf{Z}_i \mid \mathbf{X}, \Theta^t) LL(\Theta \mid \mathbf{X}, \mathbf{Z}_i) \\ &= \sum_{i=1}^{N} P(\mathbf{Z}_i \mid \mathbf{X}, \Theta^t) \ln P(\mathbf{X}, \mathbf{Z}_i \mid \Theta)\end{aligned}$$

可以发现，最后的表达式中$\mathbf{X}, \mathbf{Z}_i, \Theta^t$均为已知量，尤其是$\mathbf{X}$为已观测变量，$\mathbf{Z}_i$为隐变量$\mathbf{Z}$可能的取值（虽然不知道$\mathbf{Z}$是什么，但它的取值范围已知），是客观已知的量，与参数无关；而Θ^t是参数Θ的当前取值，因此上式可以表达为$Q(\Theta \mid \Theta^t)$，表示已知Θ^t条件下的表达式。

4、EM 算法的解释

我们希望使用最大似然估计由$\ln P(\mathbf{X}, \mathbf{Z} \mid \Theta)$得到$\Theta$，但由于隐变量$\mathbf{Z}$也是未知的，因此不能直接求解。

EM 算法是一种交替迭代的方法：先假设已知Θ，求出$\mathbf{Z}$的期望；然后利用估计出的$\mathbf{Z}$再做极大似然估计得到新的Θ；进而再求$\mathbf{Z}$的期望，再得到新的Θ……

这里自然会产生一个问题：是否可以先假设已知$\mathbf{Z}$，做极大似然估计得到Θ，然后再利用估计出的Θ更新$\mathbf{Z}$呢？本来二者都是未知的，必须先初始化其一，到底应该是先有鸡再有蛋还是应该先有蛋再有鸡呢？

如果仅从交替迭代优化的角度来说，肯定是可以的，个人还曾经试验过；另外，第 163 页左下角的边注中提到 EM 算法可看作坐标下降法的特例，以附录 B.5 符号为准，一般来讲$\boldsymbol{x}$的d维分量迭代顺序当然可交换。（仅为个人观点，勿被误导）

7.7 本章小节

本章内容很多，但读完本章后的感觉可能会是这样的：除了 7.3 节的朴素贝叶斯分类器看的很爽快，其它似乎什么也没看懂；7.1 节的贝叶斯定理和 7.2 节的极大似然估计本就是本科概率论中的内容，似乎什么都知道却也什么都不知道，熬过了 7.4 节的半朴素贝叶斯分类器和 7.5 节贝叶斯网，在“山重水复疑无路”绝望之际，惊喜发现接下来是大名鼎鼎的 EM 算法，当准备大干一场找回信心的时候，却发现 7.6 节就四个公式，然而却啥也看不懂。

然而多读几遍，发现很多概念慢慢地都理解了，诸如贝叶斯定理、贝叶斯风险、贝叶斯最优分类器、生成式模型和判别式模型、后验概率和似然概率、极大似然估计和贝叶斯估计、拉普拉斯修正(平滑)等等，都是一些很重要、很常见的概念。

如果看书或看文章看不懂，那就多看几遍，不用一直盯着，可以先做其它事情，没事儿了就来瞅瞅，突然有一天你会发现：哦，原来如此！