2.概率论

2.1 前沿

> Probability theory is nothing but common sense reduced to calculation. —— Pierre Laplace

在前面的章节中，我们已经了解了概率论在机器学习中所扮演的有用的角色。本章，我们将讨论更多关于概率论的细节。我们并没有足够的篇幅展开相关领域的深层次讨论——读者可以自行参考更多的相关书籍。但在后面的章节中，我们将简明扼要的介绍许多可能会用到的关键思想。

在我们探讨更多技术方面的细节之前，请容许我们思考一个问题：什么是概率？我们对于诸如“一个硬币面朝上的概率为0.5”的表述已经非常熟悉。但这句话到底意味着什么？关于概率至少有两种不同的解释。一种是**频率**（frequentist）学派的解释。在这种观点中，概率代表事件在长时间实验的情况下出现的频率。比如，在前面的例子中，我们是指如果我们投掷一枚硬币很多次，那么我们相信有一半的时间硬币的正面朝上。

另一个关于概率的解释为**贝叶斯**（Bayesian）学派的解释。在这种观点中，概率是用来量化我们对于某些事件的**不确定度**（uncertainty），所以本质上它与信息而非重复的实验相关。从贝叶斯学派的观点看待上述的例子，意味着我们相信在下一次投掷硬币时，硬币正面朝上的可能性为0.5。

贝叶斯解释的一个重要优势在于，它可以用来衡量那些无法进行重复试验的事件的不确定度。比如说我们希望估计到2020年冰川融化的概率。这个事件本身可能只发生一次甚至不会发生，也就是说它是不能被重复的试验。然而，我们可以量化针对该事件发生的不确定度（比如说基于我们采取的一些抑制全球变暖的行为，我们可以认为这个事件发生的可能性会变小）。再比如第1章中提及的垃圾邮件分类任务，我们可能已经收到一个特定的邮件信息，我们希望计算它是垃圾邮件的可能性。或者我们在雷达屏幕上观察到一个移动光点，我们希望计算这个飞行物身份的概率分布（它是一只鸟还是一个飞机呢？）。在上述的所有案例中，尽管没有一个事件是可以重复试验的，但贝叶斯观点却是有效且具备天然可解释性的。**所以在本书中我们将采用贝叶斯观点对概率的解释。**不过幸运的是，无论我们采取哪一种观点去看待概率，概率论的基本规则都是一样的。

2.2 关于概率论的简单综述

本章介绍关于概率论基础的一些知识，仅仅针对那些对相关知识已经生疏的读者。关于更多的相关细节，可以参考其他的相关书籍。已经对这块知识比较熟悉的读者可以直接跳过本章。

2.2.1 离散随机变量

表达式$p(A)$表示事件A发生的概率。比如A可能表示"明天会下雨"。其中$p(A)$满足$0 \leq p(A) \leq 1$，如果$p(A) = 0$，表示事件A不可能发生，$p(A) = 1$意味着事件A肯定发生。我们使用$p(\bar{A})$表示事件非A发生的可能性，满足$p(\bar{A}) = 1 - p(A)$。通常情况下，我们将"A发生"这个事件写作$A = 1$，"A不发生"写作$A = 0$。

通过定义**离散随机变量**（discrete random variable）X，我们可以扩展二元事件（即事件只存在两种状态）的符号表达，该离散变量取值于一个有限集或者可数无限集 χ(译者注：关于可数无限集的例子：比如做一个抛掷硬币的试验，直到第一次出现正面时抛掷硬币的次数χ的取值所构成的就是一个可数无限集)。我们将事件$X = x$发生的概率表示为$p(X = x)$，或者直接写成$p(x)$。其中符号$p()$称为**概率质量函数**(probability mass function),满足性质$0 \leq p(x) \leq 1, \sum_{x \in \chi} p(x) = 1$。图2.1展示了定义在一个有限**状态空间**(state space)$\chi = \{1, 2, 3, 4\}$上的两种概率质量函数。其中左图属于均匀分布,$p(x) = 1/4$，右图为一个退化分布$p(x) = \mathbb{I}(x = 1)$，其中$\mathbb{I}()$为二元**指示函数**(indicator function)，这个分布意味着X永远等于1，换句话说，它是一个常数。

2.2.2 基本定理

本章，我们将介绍概率论的基本定理。

2.2.2.1 两个事件并集发生的概率

给定两个事件A和B，定义事件A或B发生的概率为：

ParseError: KaTeX parse error: Undefined control sequence: \or at position 18: …egin{align} p(A\or B) &= p(A)+p(B)…

2.2.2.2 联合概率

我们定义事件A和B同时发生的概率为：
ParseError: KaTeX parse error: Undefined control sequence: \and at position 11: p(A,B)=p(A\and B)=p(A|B)p(B) \…
上式通常又被称为**乘法法则**(product rule)。给定两个事件的**联合概率分布**（joint distribution）$p(A, B)$，定义**边缘分布**(marginal distribution)如下:

$$p(A) = \sum_b p(A, B) = \sum_b p(A|B = b)p(B = b) \tag{2.4}$$

上式我们针对事件B所有的可能状态进行求和。类似地，我们也可以定义$p(B)$，这通常被称为**求和法则**（sum rule）或者叫**全概率法则**（rule of total probability）。

我们可以多次使用乘法法则，进而引出概率论中的**链式法则**（chain rule）：

$$p(X_{1:D}) = p(X_1)p(X_2|X_1)p(X_3|X_2, X_1)...p(X_D|X_{1:D-1}) \tag{2.5}$$

式中，我们模仿了Matlab中的一种符号写法$1:D$表示集合$\{1, 2, \ldots, D\}$。

2.2.2.3 条件概率

我们定义在事件B发生的前提下，事件A发生的概率为**条件概率**（conditional probability）：

$$p(A|B) = \frac{p(A, B)}{p(B)} \text{ if } p(B) > 0 \tag{2.6}$$

2.2.3 贝叶斯法则

根据求和法则和求积法则，结合条件概率的定义，可以得到**贝叶斯法则**（Bayes' rule）或者**贝叶斯定理**（Bayes' Theorem）。

$$p(X = x|Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)} = \frac{p(X = x)p(Y = y|X = x)}{\sum_{x'} p(X = x')p(Y = y|X = x')} \tag{2.7}$$

Sir Harold Jeffreys指出"**贝叶斯定理之于概率论等价于勾股定理之于几何学**" 。我们在下面将介绍两个贝叶斯理论的应用，但在全书后面的内容中，我们将会碰到很多相关的例子。

2.2.3.1 案例：医疗诊断

考虑一个关于使用贝叶斯法则的案例：医疗诊断问题。假设你是一个40多岁的女性，你决定去做一个名叫**manmogram**检测,以判断自己是否患有乳腺癌。如果检测结果显示为阳性，那么你患病的概率有多大？这个问题的答案显然与检测方法的可靠性有关。假设你被告知检测方法的**敏感性**(sensitivity)为80%，也就是说，如果你患有癌症，那么测试结果显示为阳性的可能性为0.8。换句话说：

$$p(x = 1|y = 1) = 0.8 \tag{2.8}$$

其中$x = 1$表示检测结果为阳性，$y = 1$表示你确实患有乳腺癌。许多人因此就认为他们患有癌症的可能性为80%。但这是错误的！因为他们忽略了患有乳腺癌的先验概率，幸运的是，这个概率相当低：

$$p(y = 1) = 0.004 \tag{2.9}$$

忽略先验概率的情况被称为**基率谬论**(base rate fallacy)。我们同样需要考虑测试结果可能是**假正例**(false positive)或者**误警报**(false alarm)的情况。不幸的是，类似这样的假正例发生的可能性还很高(在现有的筛选技术下)：

$$p(x=1|y=0)=0.1 \tag{2.10}$$

将上述三项通过贝叶斯法则进行合并，我们可以计算出正确的答案：

ParseError: KaTeX parse error: Multiple \tag

其中$p(y=0)=1-p(y=1)=0.996$。换句话说，如果你的检测结果为阳性，你也只有大概3%的可能性患有乳腺癌。(该案例中的数据来自于相关文献，基于该项分析，美国政府决定不再推荐女性在40岁时进行每年的mammogram检测：因为假正例会导致不必要的担心和压力，并最终导致不必要的，昂贵的，并且可能存在潜在伤害的后续检测。5.7节将介绍当我们在不确定的情况下权衡利弊的最优方法)。

2.2.3.2 案例：生成式分类器

我们可以将医疗诊断的例子一般化，从而实现对任意形式的特征向量$\mathbf{x}$进行分类：

$$p(y=c|\mathbf{x})=\frac{p(y=c)p(\mathbf{x}|y=c)}{\sum_{c'}p(y=c'|\theta)p(\mathbf{x}|y=c')} \tag{2.13}$$

上式被称为**生成式分类器**(generative classifier)，因为它通过使用**类条件概率密度**$p(\mathbf{x}|y=c)$和类先验分布$p(y=c)$来指定如何生成样本。我们会在第3章和第4章详细讨论这一类模型。与该模型不同的时，判别式分类器直接对类后验概率分布$p(y=c|\mathbf{x})$进行训练。我们会在8.6节讨论两种方法的优缺点。

2.2.4 独立性和条件独立性

如果两个随机变量X和Y的联合概率分布可以表示为边缘分布的乘积，则称X和Y是**无条件独立**(unconditionally independent)或者**边缘独立**(marginally independent)的，

ParseError: KaTeX parse error: Undefined control sequence: \rightleftarrow at position 11: X \perp Y \rightleftarrow p(X,Y)=p(X)p(Y)…

一般情况下，我们称一系列变量之间是互相独立的，如果其联合概率分布等于各个边缘分布的乘积。不幸的是，无条件独立很少会发生，因为大部分变量会相互影响。然而，这种影响可以通过其他变量间接产生。因此我们称已知Z的情况下,X和Y是**条件独立**(conditionally independent,CI)的，其成立的充要条件是条件联合分布等于条件边缘分布的乘积：

$$X \perp Y|Z \leftrightarrow p(X,Y|Z)=p(X|Z)p(Y|Z) \tag{2.15}$$

我们在第10章会讨论图模型，届时我们可以发现这种关系可以表示为一种图结构 $X-Z-Y$，该图反映出X和Y之间的所有关联性都需要通过Z。举例来说，如果已经知道今天是否会下雨(Z)，那么明天将会下雨的概率(事件X)与今天的地是否潮湿(事件Y)之间是独立的。直觉上，因为Z同时导致了X和Y

，所以如果已经知道了Z，那么为了了解X，我们并不需要关于Y的信息，反之亦然。我们将会在第10章进行展开讨论。
条件独立的另一个特性为：

定理2.2.1 $X \perp Y|Z$ 当且仅当存在函数g和h满足：

$$p(x,y|z) = g(x,z)h(y,z) \tag{2.16}$$

对于所有的x, y, z成立，且$p(z) > 0$.

条件独立性假设允许我们通过一些局部信息去构建大规模的概率模型。本书中将介绍大量的相关案例，特别是在3.5节，我们将会讨论朴素贝叶斯分类器，在17.2节，会讨论马尔可夫模型，在第10章讨论图模型;所有这些模型都充分应用了条件独立性的性质。

2.2.5 连续随机变量

截至目前，我们只是讨论了关于离散变量的情况。本节将介绍有关连续变量的相关内容。

假设X是某个未知的连续变量。变量X满足$a \leq X \leq b$的概率可以通过如下的方式进行计算。 接下来，我们定义三个事件 $A = (X \leq a), B = (X \leq b), W = (a < X < b)$. 显然，我们有 $B = A \vee W$， 因为事件A和事件B是互斥的，根据概率论的球和法则，我们有：

$$p(B) = p(A) + p(W) \tag{2.17}$$

所以

$$p(W) = p(B) - p(A) \tag{2.18}$$

接下来定义函数$F(q) \triangleq = p(X \leq)$. 该函数被称为**累积分布函数**(cumulative distribution function, cdf)，属于非单调递减函数。 图2.3(a)给出了示意图。基于该定义，我们有：

$$p(a < X \leq b) = F(b) - F(a) \tag{2.19}$$

定义$f(x) = \frac{d}{dx}F(x)$(不妨假设导数确实存在)，该式被称为**概率密度函数**(probability density function, pdf)。 图2.3(b)给出了示意图。 在已知概率密度函数的情况下，我们可以计算一个连续变量属于某个有限区间的概率:

$$p(a < X \leq b) = \int_a^b f(x)dx \tag{2.20}$$

如果该区间足够小，我们可以有:

$$p(x \leq X \leq x + dx) \approx p(x)dx \tag{2.21}$$

其中$p(x) \geq 0$,但对于任意给定的x, $p(x) > 1$是存在可能的，只要积分等于1即可。举例来说，考虑一个**均匀分布**(uniform distribution)$Unif(a, b)$:

$$Unif(x|a, b) = \frac{1}{b-a}\mathbb{I}(a \leq x \leq b) \tag{2.22}$$

如果我们令$a = 0, b = \frac{1}{2}$，我们有$\forall x \in [0, \frac{1}{2}], p(x) = 2$.

2.2.6 分位数(Quantiles)

如果累积分布函数F是一个单调递增函数，那么它必然存在一个逆函数，定义为F^{-1}。如果F是关于变量X的cdf，则$F^{-1}(\alpha) = x_\alpha$满足$p(X \leq x_\alpha) = \alpha$，$x_alpha$被称为函数$F$的$\alpha$**分位数**。常用的分位数包括，$F^{-1}(0.5)$为分布的**中**(median)位数，也就是说一半的概率质量分布在左边，另一半分布在右边。