MLAPP 读书笔记 - 02 概率

> A Chinese Notes of MLAPP，MLAPP 中文笔记项目
> https://zhuanlan.zhihu.com/python-kivy

记笔记的人：cycleuser

2018年05月07日11:07:07

2.1 概率论简介

> 概率论就是把常识降维表达成计算而已。---皮埃尔 拉普拉斯（Pierre Laplace）1812

前面的章节里，可以看出概率论在机器学习里面扮演了很重要的角色。所以本章就详细讲一下概率论。不过本章内容不可能面面俱到而且也不会过分强调细节，所以你最好还是找一本参考书来看看啥的。本章会涉及到后文要用到的很多关键概念和思路。

在讲解具体内容之前，先停下来想一下，什么是概率？我们都很熟悉这种说法：一枚硬币人头朝上的概率是0.5.但这句话到底是什么意思？实际上对概率有两种不同解读。

第一种是频率论阐述（frequentist interpretation），这种观点认为概率表现力时间长期法师的频率。例如上一句中所说，只要足够多次地抛硬币，人头朝上的概率就是一半。

另一种是对概率的贝叶斯阐述（Bayesian interpretation）。这种观点是概率是用来衡量某种事物的不确定性（uncertainty），与信息更相关，而不是试验的重复（Jaynes 2003）。按照这种观点，上面说硬币的那句话的意思就是我们相信下次抛硬币两面朝上的概率各半。

贝叶斯解释的阐述的一个很大的好处是可以用于对具有不确定性的时间进行建模，而不必要进行长期频率测试。比如估算到 2020年的时候极地冰盖的融化量。这事情可能发生也可能不发生，但不能重复啊。我们也本应对某事的不确定性进行定量衡量；基于我们队这件事发生概率的认知，就可以采取近似行动，这部分参考本书5.7 讲解了在不确定情况下最优选择的相关讲解。在机器学习方面一个例子就是电子邮件判断是否为垃圾邮件。再比如就是雷达上发现一个小点，我们可能要计算该位置对应物体的概率分布，推断是鸟、飞机还是导弹。所有这类场景中，重复试验的思路都是不可行的，但贝叶斯阐述则是可用的，而且也符合直觉。因此本书对概率的解读就采用了贝叶斯阐述。好在概率论的基础内容都是相似的，也不受到所选阐述方式的影响。

> 此处查看原书中图2.1

2.2 概率论简单回顾

这部分就是简单回顾一下概率论的基础内容，读者如果对概率论生疏了可以看看，如果还很熟悉这些基本内容 就没必要看了，略过即可。

2.2.1 离散随机变量

表达式p(A)是指 A 事件发生（为真）的概率。A 可以是逻辑判断表达式，比如：“明天会下雨”。根据概率定义就可以知道 $0 \leq p(A) \leq 1$，如果p(A)=0则意味着绝对不会发生，如果p(A)=1则意味着必然发生。用$p(\bar{A})$表示事件 A 不发生的概率；很显然，$p(\bar{A}) = 1 - p(A)$。当事件 A 为真的时候通常还简写做 A=1，反之写为 A=0。这就跟布尔运算那个类似了。

对这个二值事件的记号进行扩展，可以定义一个离散随机变量（discrete random variable）X，这个随机变量可以从任意的一个有限元素集合或者无限但可列的集合 *X* 中取值。将 X = x 的概率记作$p(X = x)$，或者缩写成$p(x)$。这里的$p()$也叫做概率质量函数（probability mass function，缩写为 pmf）。跟上面的一样，也要满足 $0 \leq p(x) \leq 1$ 和$\sum_{x \in X} p(x)$。如图2.1所示的就是两个密度函数，定义在有限状态空间$x = \{1, 2, 3, 4, 5\}$。左边的是一个均匀分布，$p(x) = 1/5$，右面的则是一个退化分布（degenerate distribution），$p(x) = \prod(x = 1)$，其中的$\prod()$是二值指标函数（binary indicator function）。这个分布表示的是 X 就总是1，也就是说是固定值。

2.2.2 基本规则

这部分讲的是概率论基本规则。

2.2.2.1 两个事件的结合概率

给定两个事件，A 和 B，可以用如下方式来定义结合概率：

$$p(A \vee B) = p(A) + p(B) - p(A \wedge B) \quad (2.1)$$

$$p(A \vee B) = p(A) + p(B) \text{若两个事件互斥（mutually exclusive）} \quad (2.2)$$

译者注：其实2.2毫无必要，因为两个事件互斥的时候也就是2.1里面的最后一项$p(A \wedge B) = 0$所以根本没必要单独定义。

2.2.2.2 联合概率

两个事件 A 和 B 的联合概率定义如下所示：

$$p(A, B) = p(A \wedge B) = p(A \mid B)p(B) \quad (2.3)$$

这也叫做乘法规则（product rule）。对两个事件的联合概率分布$p(A, B)$，可以以如下方式定义边

缘分布(marginal distribution)：

$p(A) = \sum_b p(A, B) = \sum_b p(A \mid B = b)p(B = b)$ (2.4)

上式中对所有可能的状态 B 来进行求和。对 $p(B)$的定义与之相似。也有时候也叫做加法规则（sum rule）或者全概率规则（rule of total probability）

乘法规则可以多次使用，就得到了链式规则（chain rule）：

$p(X_{1:D}) = p(X_1)p(X_2 \mid X_1)p(X_3 \mid X_2, X_1)p(X_4 \mid X_1, X_2, X_3)...p(X_D \mid X_{1:D-1})$（2.5）

上面的1:D表示的是有序集合{1,2,...,D}。

2.2.2.3 条件概率

若事件 B 为真，在此条件下事件 A 的条件概率如下所示：

$p(A \mid B) = p(A, B)/p(B) if p(B) > 0$ （2.6）

2.2.3 贝叶斯规则

结合条件概率的定义以及乘法规则和加法规则，就能推出贝叶斯规则（Bayes rule），也称为贝叶斯定理（Bayes Theorem）：

$$p(X = x \mid Y = y) = \frac{p(X=x, Y=y)}{p(Y=y)} = \frac{p(X=x)p(Y=y \mid X=x)}{\sum_{\dot{x}} p(X=\dot{x})p(Y=y \mid X=\dot{x})} \quad (2.7)$$

2.2.3.1 样例：医疗诊断

假如一位四十岁的女性，决定通过乳腺 X光检测（mammogram）做乳腺癌检测。如果结果是阳性，那么有多大概率患上？很明显这依赖于检测本身的可靠性。假设这个检测的敏感度（sensitivity）是80%，也就是如果一个人患上了，那么被检测出来是阳性的概率为0.8.即：

$p(x = 1 \mid y = 1) = 0.8$(2.8)

其中的 x=1意思就是检测结果阳性，而 y=1的意思是确实患病。
据此很多人就认为患病概率也就是80%。这是错误的！
这个计算方法忽略了患病的先验概率（prior probability），即：
$p(y = 1) = 0.004$(2.9)

把这个先验概率忽略掉，就是基本比率谬误（base rate fallacy）。此外还要考虑的就是测试本身的假阳性（false positive）或者假警报（false alarm）的概率：

$p(x = 1 \mid y = 0)$(2.10)

上面三个项目都结合起来，应用贝叶斯规则，可以计算正确概率了：

$$
\begin{aligned}
p(y=1|x=1) &= \frac{p(x=1|y=1)p(y=1)}{p(x=1|y=1)p(y=1)+p(x=1|y=0)p(y=0)} \\
&= \frac{0.8\times 0.004}{0.8\times 0.004+0.1\times 0.996} \\
&= 0.031
\end{aligned}
\tag{2.11、2.12}
$$

上式中$p(y=0)=1-p(y=1)=0.996$。也就是说即便检测结果是阳性，患病概率也只有3%而已。

2.2.3.2 样例：生成分类器

对上面医疗诊断的例子进行泛化，就可以对任意类型的特征向量 x 来进行分类了，如下所示：

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

这就是生成分类器（Generative classifiers），使用类条件概率密度$p(x|y=c)$和类先验概率密度$p(y=c)$来确定如何生成数据。更多细节在本书第3、4章。另一种方法是直接拟合类后验概率密度$p(y=c|x)$，这就叫辨别式分类器（discriminative classifier）。这两种方法的优劣在本书8.6有详细讲解。

2.2.4 独立分布和有条件独立分布

> 此处查看原书中图2.2

X 和 Y 为无条件独立（unconditional independent）或者边缘独立（marginally independent），记作 $X\perp Y$，如果用两个边缘的积来表示，则如图2.2所示：

$$
X\perp Y \iff p(X,Y)=p(X)p(Y) \tag{2.14}
$$

总的来说，如果联合分布可以写成边缘的积的形式，就可以说一系列变量之间相互独立（mutually independent）。

不过很可惜，这种无条件独立的情况是很少见的，大多数情况下变量之间都互相关联。好在一般这种影响都可以通过其他变量来传导的，而不是直接的关联。如果X 和 Y 对于给定的 Z 来说有条件独立，则意味着能够将条件联合分布写成条件边缘的积的形式，就说：

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

在本书第10章会谈到图模型，到时候会把这个假设写成图 X-Z-Y 的形式，更直观表现了 X 和 Y 之间的独立性由 Z 传导。例如，确定当天是否下雨（事件 Z），则明天是否下雨（事件 X）和当天土地是否湿润（事件 Y）之间独立。

直观来说，这是因为 Z 可以同时导致 X 和 Y，所以如果知道了 Z，就不需要知道 Y 就可以预测 X，反之亦然。第10章会详细介绍这部分内容。

条件独立分布的另外一个特点是：

定理 2.2.1

$X \perp Y \mid Z$则意味着存在着函数 g 和 h ，对全部的x,y,z，在p(z)>0的情况下满足下列关系：

$$p(x, y \mid z) = g(x, z)h(y, z) \quad (2.16)$$

练习2.8有详细证明过程。

条件概率假设让我们可以从小处着手构建大兴的概率模型。本书中有很多这样的样例，尤其是本书的3.5当中就提到了朴素贝叶斯分类器（naive Bayes classifiers），在本书17.2还有马尔科夫模型（Markov models），在本书第10章有图模型（graphical models），所有这些模型都充分利用了条件概率的性质。

2.2.5 连续随机变量

签名谈到的都是具有不确定性的离散量。接下来就要将概率论扩展应用到具有不确定性的连续量。

设 X 是一个不确定的连续量。X 在某个空间$a \leq X \leq b$的概率可以用如下方式计算。
先定义如下事件：

$$A = (X \leq a)$$
$$B = (X \leq B)$$
$$W = (a \leq X \leq b)$$

很明显有
$B = A \bigvee W$，由于 A 和 W 是相互独立的，所以可以用加法规则：
$$p(B) = p(A) + p(W) \quad (2.17)$$
显然有：
$$p(w) = p(B) - p(A) \quad (2.18)$$

> 此处查看原书中图2.3

定义一个函数$F(q) * = p(X \leq q)$，这就是 X 的累积密度函数（cumulative distribution function，缩写为 cdf）。很明显这个 cdf 是一个单调递增函数（monotonically increasing function）。如图2.3（a）所示。来利用这个记号则有：

$$p(a < X \leq b) = F(b) - F(a) \quad (2.19)$$

接下来假设这个函数 F(x)可导，则定义函数$f(x) = \frac{d}{dx}F(x)$，这个函数就叫概率密度函数（probability density function，缩写为 pdf）。参考图2.3（b）。有了 pdf，就可以计算一个有限区间上的连续变量的概率了：

$$P(a < X \leq b) = \int_a^b f(x)dx \quad (2.20)$$

随着这个区间逐渐缩小，直到趋向于无穷小，就可以得到下面的形式：

$$P(x < X \leq x + dx) \approx p(x)dx \quad (2.21)$$

我们要满足$p(x) \geq 0$，但对于任意的 x，$p(x) \geq 1$也有可能，只要密度函数最终积分应该等于1就行了。举个例子，下面的正态分布Unif(a,b）：

$$Unif(x \mid a, b) = \frac{1}{b-a}\prod(a \leq x \leq b) \quad (2.22)$$

如果设置 $a = 0, b = \frac{1}{2}$，则有了对于在$x \in [0, \frac{1}{2}]$之间取值的任意 x 都有 $p(x) = 2$，

2.2.6 分位数

由于累积密度函数（cdf） F 是一个单调递增函数，那就有个反函数，记作F^-。如果 F 是 X 的累积密度函数（cdf），那么$F^{-1}(\alpha)$就是满足概率$P(X \leq x_\alpha) = \alpha$的值；这也叫做 F 的 α分位数（quantile）。$F^{-1}(0.5)$就是整个分布的中位数（median），左右两侧的概率各自一半。而$F^{-1}(0.25)$和$F^{-1}(0.75)$则是另外两个分位数。

利用这个累积密度函数（cdf）的反函数还可以计算尾部概率（tail area probability）。例如，如果有高斯分布 $N(0, 1)$，ϕ 是这个高斯分布的累积密度函数（cdf），这样在$\phi^{-1}(\alpha/2)$ 左边的点就包含了$\alpha/2$概率密度，如图2.3（b）所示。与之对称，在$\phi^{-1}(1-\alpha/2)$ 右边的点也包含了$\alpha/2$概率密度。所以呢，在$(\phi^{-1}(\alpha/2), \phi^{-1}(1-\alpha/2))$ 就包含了$1-\alpha$的概率密度。如果设置$\alpha = 0.05$，那么中间这个区间就占了全部概率的95%了。

$$(\phi^{-1}(0.025), \phi^{-1}(0.975)) = (-1.96\ ,\ \ 1.96) \quad (2.23)$$

如果这个正态分布是$N(\mu, \sigma^2)$，那么其95%的区间就位于$(\mu - 1.96\sigma\ ,\ \ \mu + 1.96\sigma)$。有时候就简写一下成了$\mu \pm 2\sigma$。

2.2.7 均值（Mean）和方差（variance）

对正态分布来说，大家最常见常用的性质估计就是均值（mean），或者称之为期望值（expected value），记作μ。对于离散 rv （译者注：这个 rv 很突兀，之前没出现，也没解释是啥, 推测是 random variables）的情况，可以定义成$E[X] * = \sum_{x \in X} xp(x)$；对于连续 rv 的情况，可以定义为\E$[X] * = \int_X xp(x)dx$。如果这个积分是无穷的，则均值不能定义，更多例子后文会有。

然后就是方差（variance）了，这个表征的是分布的“散开程度（spread）”，记作σ^2。定义如下：

$$var[X] * = E[(X-\mu)^2] = \int (x-\mu)^2 p(x)dx \tag{2.24}$$

$$= \int x^2 p(x)dx + \mu^2 \int p(x)dx - 2\mu \int xp(x)dx = E[X^2] - \mu^2 \tag{2.25}$$

从上面的式子就可以推导出：

$E[X^2] = \mu^2 + \sigma^2$(2.26)

然后就可以定义标准差（standard deviation）了：

$std[X] * = \sqrt{var[X]}$(2.27)

标准差和 X 单位一样哈。

2.3 常见的离散分布

本节介绍的是一些常用的定义在离散状态空间的参数分布，都是有限或者无限可列的。

2.3.1 二项分布和伯努利分布

设咱们抛硬币 n 次，设$X \in \{0, \ldots, n\}$ 是人头朝上的次数。如果头朝上的概率是θ，这就可以说 X 是一个二项分布（binomial distribution），记作$X \sim Bin(n, \theta)$。则 pmf（概率质量函数）可以写作：

$Bin(k \mid n, \theta) * = \binom{n}{k} \theta^k (1-\theta)^{n-k}$ （2.28）

上式中的

$\binom{n}{k} * = \frac{n!}{(n-k)!k!}$(2.29)

是组合数，相当于国内早期教材里面的C_n^k，从 n 中取 k 个样的组合数，也是二项式系数（binomial coefficient）。如图2.4所示就是一些二项分布。

> 此处查看原书中图2.4

这个分布的均值和方差如下所示：

$mean = \theta, var = n\theta(1-\theta)$（2.30）

换个思路，如果只抛硬币一次，那么$X \in \{0, 1\}$就是一个二值化的随机变量，成功或者人头朝上的概率就是θ。这时候就说 X 服从伯努利分布（Bernoulli distribution），记作$X \sim Ber(\theta)$，其中的 pmf 定义为：

$Ber(x \mid \theta) = \theta^{\prod(x=1)} (1-\theta)^{\prod(x=0)}$(2.31)

也可以写成：

$$Ber(x \mid \theta) = \begin{cases} \theta & \text{if x =1} \\ 1-\theta & \text{if x =0} \end{cases}$$

(2.32)

很明显，伯努利分布只是二项分布中 n=1 的特例。

2.3.2 多项式（multinomial）分布和多重伯努利（multinoulli）分布

二项分布可以用于抛硬币这种情况的建模。要对有 K 个可能结果的事件进行建模，就要用到多项分布（multinomial distribution）。这个定义如下：设$x = (x_1, \ldots, x_K)$是一个随即向量，其中的x_j是第 j 面出现的次数个数。这样 x 的概率质量函数 pmf 就如下所示：

$$Mu(x \mid n, \theta) * = \binom{n}{x_1, \ldots, x_K} \prod_{j=1}^{K} \theta_j^{x_j}$$(2.33)

其中的θ_j是第 j 面出现的概率，另外那个组合数的计算如下所示：

$$\binom{n}{x_1, \ldots, x_K} * = \frac{n!}{x_1! x_2! \ldots x_K!}$$(2.34)

这样得到的也就是多项式系数（multinomial coefficient），将一个规模为$n = \sum_{k=1}^{K}$的集合划分成规模从x_1到x_K个子集的方案数。

接下来设 n=1。这就好比是讲一个 K 面的骰子只投掷一次，所以 x 就是由 0 和 1 组成的向量。其中只有一个元素会是1.具体来说就是如果 k 面朝上，就说第 k 位为1。这样就可以把 x 看做一个用标量分类的有 K 个状态的随机变量，这样 x 就是自己的虚拟编码（dummy encoding），即：$x = [\prod(x = 1), \ldots, \prod(x = K)]$。例如，如果 K=3，那么状态1、2、3对应的虚拟编码分别是(1,0,0),(0,1,0),(0,0,1)。这样的编码也称作单位有效编码（one-hot encoding），因为只有一个位置是1.这样对应的概率质量函数 pmf 就是：

$$Mu(x \mid 1, \theta) = \prod_{j=1}^{K} \theta_j^{\prod(x_j = 1)}$$(2.35)

可以参考图2.1的（b-c）作为一个例子。这是类别分布（categorical distribution）或者离散分布（discrete distribution）的典型案例。Gustavo Lacerda 建议大家称之为多重伯努利分布（multinoulli distribution），这样与二项分布/伯努利分布的区别关系相仿。本书就采用了这个术语，使用下列记号表示这种情况：

$$Cat(x \mid \theta) * = Mu(x \mid 1, \theta)$$(2.36)

换句话说，如果$x \sim Cat(\theta)$，则$p(x = j \mid theta) = \theta_j$。参考表2.1。

2.3.2.1 应用：DNA 序列模体

> 此处查看原书中图2.5

生物序列分析（biosequence analysis）是一个典型应用案例，设有一系列对齐的 DNA 序列，如图2.5（a）所示，其中有10行（序列），15列（沿着基因组的位置）。如图可见其中几个位置是进化的保留位，是基因编码区域的位置，所以对应的列都是“纯的”，例如第7列就都是 G。

如图2.5（b）所示，对这种数据的可视化方法是使用序列标识图（sequence logo）。具体方法是把字幕ACGT 投到对应位置上，字体大小与其实验概率（empirical probability）成正比，最大概率的字幕放在最顶部。

对计数向量归一化，得到在位置t, $\hat{\theta}_t$的经验概率分布，可以参考本书的公式3.48：

$$N_t = (\sum_{i=1}^{N} \prod(X_{it} = 1), \sum_{i=1}^{N} \prod(X_{it} = 2), \sum_{i=1}^{N} \prod(X_{it} = 3), \sum_{i=1}^{N} \prod(X_{it} = 4)) \quad (2.37)$$

$$\hat{\theta}_t = N_t / N \quad (2.38)$$

这个分布就被称作一个模体（motif）。可以计算每个位置上最大概率出现的字母，得到的就是共有序列（consensus sequence）。

2.3.3 泊松分布（Poisson Distribution）

如果一个离散随机变量$X \in \{0, 1, 2, \ldots\}$服从泊松分布，即$X \sim Poi(\lambda)$，其参数$\lambda > 0$，其概率质量函数 pmf 为：

$$Poi(x \mid \lambda) = e^{-\lambda} \frac{\lambda^x}{x!} \quad (2.39)$$

第一项是标准化常数（normalization constant），使用来保证概率密度函数的总和积分到一起是1.

泊松分布经常用于对罕见事件的统计，比如放射性衰变和交通事故等等。参考图2.6是一些投图样本。

2.3.4 经验分布（empirical distribution）

某个数据集，$D = \{x_1, \ldots, x_N\}$，就可以定义一个经验分布，也可以叫做经验测度（empirical measure），形式如下所示：

$$p_{emp}(A) * = \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i}(A) \quad (2.40)$$

其中的$\delta_x(A)$是狄拉克测度（Dirac measure），定义为：

$$\delta_x(A) = \begin{cases} 0 \text{ if } x \notin A \\ 1 \text{ if } x \in A \end{cases}$$

(2.41)

一般来说可以给每个样本关联一个权重（weight）
$p(x) = \sum_{i=1}^{N} w_i \delta_{x_i}(x)$（2.42）

其中要满足$0 \leq w_i \leq 1$以及$\sum_{i=1}^{N} w_i = 1$。可以想象成一个直方图（histogram），其中每个点x_i位置都有各自的峰（spike），而w_i决定了峰值i 的高低。这个分布中，所有不在数据集中的点就都设为0了。

2.4 一些常见的连续分布

接下来介绍一些常用的单变量一维连续概率分布。

2.4.1 高斯（正态）分布

不管是统计学还是机器学习里面，最广泛使用的都是高斯分布了，也叫做正态分布。其概率密度函数 pdf 为：
$N(x \mid \mu, \sigma^2) * = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$(2.43)

上式中的$\mu = E[X]$是均值（mean）也是模（mode）,$\sigma^2 = var[X]$ 是方差（variance）。$\sqrt{2\pi\sigma^2}$是归一化常数（normalization constant），用于确保整个密度函数的积分是1，具体可以参考练习2.11。

可以写成$X \sim N(\mu, \sigma^2)$来表示$p(X = x) = N(x \mid \mu, \sigma^2)$。$X \sim N(0, 1)$就是标准正态分布（standard normal distribution）。图2.3（b）是这样一个标准正态分布的概率密度函数图像，也被称作钟形曲线。

所谓高斯分布的精确度（precision）就是方差的倒数$\lambda = 1/\sigma^2$。精确度高的意思也就是方差低，而整个分布很窄，对称分布在均值为中心的区域。

要注意这是一个概率密度函数（pdf），所以完全可以$p(x) > 1$。比如在中心位置，$x = \mu$,这样就有$N(\mu\mu, \sigma^2) = (\sigma\sqrt{2\pi})^{-1} e^0$,所以如果$\sigma < 1/\sqrt{2\pi}$,这样就有$p(x) > 1.$

高斯分布的累积分布函数(cdf)为:

$\phi(x; \mu, \sigma^2) * = \int_{-\infty}^{x} N(z \mid \mu, \sigma^2) dz$(2.44)

图2.3(a) 所示为当$\mu = 0, \sigma^2 = 1$时候的 cdf 函数曲线.这个函数的积分没有闭合形式表达式,不过在多数软件包里面都内置了.另外还能以误差函数(error function,缩写为 erf)的形式来计算:

$$\phi(x; \mu, \sigma) = \frac{1}{2}[1 + erf(z/\sqrt{2})] \quad (2.45)$$

其中的$z = (x - \mu)/\sigma$,误差函数为:

$$erf(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2} dt \quad (2.46)$$

高斯分布是统计学里面用的最广的分布,有几个原因.首先是这两个参数很好解释,分别对应着分布中的两个基础特征,均值和方差.其次中心极限定理(central limit theorem, 参考本书2.6.3)也表明独立随机变量的和旧近似为高斯分布,所以高斯分布很适合用来对残差或者噪音建模.然后高斯分布有最小假设数(least number of assumptions),最大熵(maximum entropy),适合用于有特定均值和方差情境下建立约束,如本书9.2.6所述,这就使得高斯分布是很多情况下很不错的默认选择.另外,高斯分布的数学形式也很简单,容易实现,效率也很高.高斯分布更广泛应用参考 Jaynes 2003 第七章.

2.4.2 退化概率分布函数(Degenerate pdf)

如果让方差趋近于零,即$\sigma^2 \to 0$,那么高斯分布就变成高度为无穷大而峰值宽度无穷小的形状了,中心当然还是在μ位置:

$$\lim_{\sigma^2 \to 0} N(x \mid \mu, \sigma^2) = \delta(x - \mu) \quad (2.47)$$

这个新的分布函数δ就叫做狄拉克函数(Dirac delta function).其定义为:

$$\delta(x) = \begin{cases} \infty \text{ if } x = 0 \\ 0 \text{ if } x \neq 0 \end{cases} \quad (2.48)$$

这样进行积分也有

$$\int_{-\infty}^{\infty} \delta(x) dx = 1 \quad (2.49)$$

这个狄拉克函数的有个特点就是筛选特性(sifting property),从一系列求和或者积分当中筛选了单一项目:

$$\int_{-\infty}^{\infty} f(x)\delta(x - \mu) dx = f(\mu) \quad (2.50)$$

只有当$x - \mu = 0$的时候这个积分才是非零的.

高斯分布有个问题就是对异常值很敏感,因为从分布中心往外的对数概率衰减速度和距离成平方关系(since the logprobability only decays quadratically with distance from the center).

有一个更健壮的分布,就是所谓的 T 分布或者也叫学生分布(student distribution),其概率密度函数如下所示:

$$T(x|\mu,\sigma^2,v) \propto [1+\frac{1}{v}(\frac{x-\mu}{\sigma})^2]^{-(\frac{v+1}{2})} \quad (2.51)$$

上式中的μ是均值,$\sigma^2>0$是范围参数(scale parameter),$v>0$称为自由度(degrees of freedom). 如图2.7就是该函数的曲线.为了后文用起来方便,这里特地说一下几个属性:

$$mean=\mu, mode=\mu, var=\frac{v\sigma^2}{v-2} \quad (2.52)$$

这个模型中,当自由度大于2 $v>2$的时候方差才有意义,自由度大于1 $v>1$均值才有意义.

此处参见原书图2.7

此处参见原书图2.8

T 分布的稳定性如图2.8所示,左侧用的是没有异常值的高斯分布和T 分布,右侧是加入了异常值的.很明显这个异常值对于高斯分布来说干扰很大,而 T 分布则几乎看不出来有影响.因为 T 分布比高斯分布更重尾(heavier tails), 至少对于小自由度 v 的时候是这样,如图2.7所示.

如果自由度 v=1,则 T 分布就成了柯西分布(Cauchy distribution)或者洛伦兹分布(Lorentz distribution) .要注意这时候重尾(heavy tails)会导致定义均值(mean)的积分不收敛.

要确保有限范围的方差(finite variance), 就需要自由度 v>2.一般常用的是自由度v=4,在一系列问题中的性能表现也都不错(Lange 等1989).如果自由度远超过5,即v>>5,T 分布就很快近似到高斯分布了,也就失去了健壮性(robustness)了.

此处参见原书图2.9

2.4.3 拉普拉斯分布(Laplace distribution)

另外一个常用的重尾分布就是拉普拉斯分布,也被称为双面指数分布(double sided exponential distribution),概率密度函数如下所示:

$$Lap(x|\mu,b)*=\frac{1}{2b}\exp(-\frac{|x-\mu|}{b}) \quad (2.53)$$

上式中的μ是位置参数(location parameter), b>0 是缩放参数(scale parameter),如图2.7所示就是其曲线.这个分布的各个属性如下所示:

$$mean=\mu, mode=\mu, var=2b^2 \quad (2.54)$$

这个分布的健壮性(robustness)如图2.8中所示,另外从图中也可以发现拉普拉斯分布比高斯分布在0点有更多概率.这个特性在本书13.3的时候还要用到,很适合用于在模型中增强稀疏性(encourage sparsity).

2.4.4 γ分布

这个分布很灵活,适合正实数值的rv, x>0. 用两个参数定义,分别是形状参数(shape) a>0 和频率参数(rate) b>0:

$$Ga(T|shape=a,rate=b)*=\frac{b^a}{\Gamma(a)}T^{a-1}e^{-Tb} \quad (2.55)$$

上式中的$\Gamma(a)$是一个γ函数:
$\Gamma(x)\int_0^\infty u^{x-1}e^{-u}du$(2.56)

参考图2.9就是一些样例图像.这个分布的各个属性如下所示:

$$mean=\frac{a}{b},mode=\frac{a-1}{b},var=\frac{a}{b^2} \quad (2.54)$$

有一些分布实际上就是γ分布的特例, 比如下面这几个:

- 指数分布(Exponential distribution)定义是$Expon(x|\lambda)*=Ga(x|1,\lambda)$,其中的$\lambda$是频率参数(rate).这个分布描述的是泊松过程(Poisson process) 中事件之间的时间间隔.例如,一个过程可能有很多一系列事件按照某个固定的平均频率λ连续独立发生.
- 厄兰分布(Erlang Distribution)就是一个形状参数 a 是整数的γ分布,一般会设置 a=2,产生的是一个单参数厄兰分布,$Erlang(x|\lambda)=Ga(x|2,\lambda),$,$\lambda$也是频率参数.
- 卡方分布(Chi-squared distribution)定义为$\chi^2(x|\nu)*=Ga(x|\frac{\nu}{2},\frac{1}{2})$.这个分布是高斯分布随机变量的平方和的分布.更确切地说,如果有一个高斯分布$Z_i\sim N(0,1),$,那么其平方和$S=\sum_{i=1}^{\nu}Z_i^2$则服从卡方分布$S\sim\chi_\nu^2$.

另一个有用的结果是:如果一个随机变量服从γ分布:$X\sim Ga(a,b)$ 那么这个随机变量的导数就服从一个逆γ分布(inverse gamma),即$\frac{1}{X}\sim IG(a,b)$,这个在练习2.10里面有详细描述.逆γ分布(inverse gamma)定义如下:

$$IG(x|shape=a,scale=b)*=\frac{b^a}{\Gamma(a)}x^{-(a+1)}e^{-\frac{b}{x}} \quad (2.58)$$

这个逆γ分布的三个属性 如下所示:

$$mean=\frac{b}{a-1},mode=\frac{b}{a+1},var=\frac{b^2}{(a-1)^2(a-2)} \quad (2.59)$$

这个均值只在 a>1 的情况下才存在,而方差仅在 a>2 的时候存在.

2.4.5 β分布

此处参见原书图2.10

β分布支持区间[0,1],定义如下:

$$Beta(x|a,b) = \frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1} \quad (2.60)$$

上式中的$B(a,b)$是一个β函数,定义如下:

$$B(a,b) * = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \quad (2.61)$$

这个分布的函数图像可以参考图2.10.需要 a 和 b 都大于零来确保整个分布可以积分,这是为了保证$B(a,b)$存在.如果 a=b=1, 得到的就是均匀分布(uniform distirbution),如图2.10中红色虚线所示.如果 a 和 b 都小于1,那么得到的就是一个双峰分布(bimodal distribution),两个峰值在0和1位置上,如图2.10中的蓝色实线所示.如果 a 和 b 都大于1了,得到的就是单峰分布(unimodal distribution)了,如图2.10中的另外两条虚线所示.这部分内容在练习2.16里会用到.这个分布的属性如下:

$$mean = \frac{a}{a+b}, mode = \frac{a-1}{a+b-2}, var = \frac{ab}{(a+b)^2(a+b+1)} \quad (2.52)$$

2.4.6 柏拉图分布(Pareto distribution)

这个分布是用来对有长尾(long tails)或称为重尾(heavy tails)特点的变量进行建模的.例如,英语中最常出现的词汇是冠词 the, 出现概率可能是第二位最常出现词汇 of 的两倍还多, 而 of 也是第四位的出现次数的两倍,等等.如果把每个词汇词频和排名进行投图,得到的就是一个幂律(power law),也称为齐夫定律(Zipf's law).财富的分配也有这种特点,尤其是在美帝这种腐朽的资本主义国度.

柏拉图分布的概率密度函数(pdf)如下所示:

$$Pareto(x|k,m) = km^k x^{-(k+1)} \prod(x \geq m) \quad (2.63)$$

通过定义可知, x 必须必某一个常数 m 大,但又不用大特别多,而其中的 k 就是控制这个的,避免 x 太大.随着$k \to \infty$,这个分布就接近于狄拉克分布$\delta(x-m)$了.参考图2.11(啊) 就是一些此类分布函数的图像,如果用对数坐标来进行投图,就会形成一条直线,如图2.11(b) 所示那样.这个直线的方程形式就是$\log p(x) = a\log x + c$,其中的 a 和 c 是某个常数.这也叫幂律(power law).这个分布的属性如下所示:

$$mean = \frac{km}{k} - 1 \text{ if } k > 1, mode = m, var = \frac{m^2 k}{(k-1)^2(k-2)} \text{ if } k > 2 \quad (2.64)$$

此处查看原书图2.11

2.5 联合概率分布

签名的都是单变量的概率分布,接下来要看看更难的,就是联合概率分布(joint probability distributions),其中要涉及到多个相关联的随机变量,实际上这也是本书的核心内容.

联合概率分布的形式是$p(x_1, \ldots, x_D)$,这些随机变量属于一个规模为 D>1 的集合,对这些随机变量间 的(随机stochastic)关系进行建模,就要用联合概率分布.如果所有变量都是离散的,那就可以吧联合分布表示成一个大的多维数组,每个维度对应一个变量.若设每个变量的状态数目总共是 K, 那么这样要建模的话,需要定义的参数个数就达到了$O(K^D)$了.

2.5.1 协方差和相关系数

协方差(covariance)是用来衡量两组变量之间(线性)相关的程度的,定义如下:

$$cov[X, Y] * = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y] \quad (2.65)$$

此处查看原书图2.12

如果 x 是一个 d 维度的随即向量,那么它的协方差矩阵(covariance matrix) 的定义如下所示,这是一个对称正定矩阵(symmetric positive definite matrix):

$$cov[x] * = E[(x - E[x])(x - E[x])^T] \quad (2.66)$$

$$= \begin{pmatrix} var[X_1] & cov[X_1, X_2] & \ldots & cov[X_1, X_d] \\ cov[X_2, X_1] & var[X_2] & \ldots & cov[X_2, X_d] \\ \ldots & \ldots & \ldots & \ldots \\ cov[X_d, X_1] & cov[X_d, X_2] & \ldots & var[X_d] \end{pmatrix} \quad (2.67)$$

协方差可以从0到∞之间取值.有时候为了使用方便,可以只用其中的上半部分.

两个变量X 和 Y 之间的皮尔逊相关系数(correlation coefficient)定义如下:

$$corr[X, Y] * = \frac{cov[X, Y]}{\sqrt{var[X] var[Y]}} \quad (2.68)$$

而相关矩阵(correlation matrix)则为:

$$R = \begin{pmatrix} cov[X_1, X_1] & cov[X_1, X_2] & \ldots & cov[X_1, X_d] \\ \ldots & \ldots & \ldots & \ldots \\ cov[X_d, X_1] & cov[X_d, X_2] & \ldots & var[X_d] \end{pmatrix}$$

(2.69)

从练习4.3可知相关系数是在[-1,1]这个区间内的,因此在一个相关矩阵中,每一个对角线项值都是1,其他的值都是在[-1,1]这个区间内.

另外还能发现的就是当且仅当有参数 a 和 b 满足$Y = aX + b$的时候,才有$corr[X, Y] = 1$,也就是说 X 和 Y 之间存在线性关系,参考练习4.3.

根据直觉可能有人会觉得相关系数和回归线的斜率有关,比如说像$Y = aX + b$这个表达式当中的系数 a 一样.然而并非如此,如公式7.99中所示,实际上回归系数的公式是$a = cov[X, Y]/var[X]$.可以将相关系数看做对线性程度的衡量,参考图2.12.

回想本书的2.2.4,如果 X 和 Y 相互独立,则有$p(X, Y) = p(X)p(Y)$,这样二者的协方差$cov[X, Y] = 0$,相关系数$corr[X, Y] = 0$, 很好理解,相互独立就是不相关了.但反过来可不成立,不相关并不能意味着相互独立.例如设$X \sim U(-1, 1), Y = X^2$. 很明显吧,这两个是相关的对不对,甚至 Y 就是 X 所唯一决定的,然而如练习4.1所示,这两个变量的相关系数算出来等于零啊,即$corr[X, Y] = 0$.图2.12有更多直观的例子,都是两个变量 X 和 Y 显著具有明显的相关性,而计算出来的相关系数却都是0.实际上更通用的用来衡量两组随机变量之间是否独立的工具是互信息量(mutual information),这部分在本书2.8.3当中有设计.如果两个变量真正不相关,这个才会等于0.

此处查看原书图2.13

2.5.2 多元高斯分布

多元高斯分布(multivariate Gaussian)或者所谓多元正态分布(multivariate normal,缩写为MVN),是针对连续随机变量的联合概率分布里面用的最广的.在第四章会对其进行详细说明,这里只说一些简单定义并且给个函数图像瞅瞅.

在 D 维度上的多元正态分布(MVN)的定义如下所示:

$$N(x \mid \mu, \Sigma) * = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} \exp[-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)] \quad (2.70)$$

上式中$\mu = E[x] \in R^D$是均值向量,而$\Sigma = cov[x]$一个$D \times D$的协方差矩阵.有时候我们会用到一个名词叫做精度矩阵(precision/concentration matrix),这个就是协方差矩阵的逆矩阵而已,也就是$\Lambda = \Sigma^{-1}$.前面那一团做分母的$(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}$也还是归一化常数,为了保证这个概率密度函数的积分等于1,更多参考练习4.5

图2.13展示了一些多元正态分布的密度图像,其中有三个是三个不同协方差矩阵的下的二维投影,另外一个是立体的曲面图像.一个完整的协方差矩阵有$D(D+1)/2$个参数,除以2是因为矩阵Σ是对称的.对角协方差矩阵的方向有 D 个参数,非对角线位置的元素的值都是0. 球面(spherical)或者各向同性(isotropic)协方差矩阵$\Sigma = \delta^2 I_D$有一个自由参数.

2.5.3 多元学生 T 分布

相比多元正态分布 MVN, 多元学生T 分布更加健壮,其概率密度函数为:

$$\Gamma(x|\mu,\Sigma,v)=\frac{\Gamma(v/2+D/2)}{\Gamma(v/2+D/2)}\frac{|\Sigma|^{-1/2}}{v^{D/2}\pi^{D/2}}\times[1+\frac{1}{v}(x-\mu)^T\Sigma^{-1}(x-\mu)]^{-(\frac{v+D}{2})} \quad (2.71)$$

$$=\frac{\Gamma(v/2+D/2)}{\Gamma(v/2+D/2)}|\pi V|^{-1/2}\times[1+(x-\mu)^T\Sigma^{-1}(x-\mu)]^{-(\frac{v+D}{2})} \quad (2.72)$$

其中的Σ叫做范围矩阵(scale matrix),而并不是真正的协方差矩阵,$V=v\Sigma$.这个分布比高斯分布有更重的尾部(fatter tails).参数v越小,越重尾;而当$v\to\infty$则这个分布趋向为高斯分布.这个分布的属性如下所示:

$$mean=\mu, mode=\mu, Cov=\frac{v}{v-2}\Sigma \quad (2.73)$$

2.5.4 狄利克雷分布

β分布扩展到多元就成了狄利克雷分布(Dirichlet distribution),支持概率单纯形(probability simplex),定义如下:

$$S_K=x:0\le x_k\le 1, \sum_{k=1}^K x_k=1 \quad (2.74)$$

其概率密度函数 pdf 如下所示:

$$Dir(x|\alpha)*=\frac{1}{B(\alpha)}\prod_{k=1}^K x_k^{\alpha_k-1}\prod(x\in S_K) \quad (2.75)$$

此处查看原书图2.14

此处查看原书图2.15

上式中的$B(\alpha_1,\ldots,\alpha_K)$是将$\beta$函数在 K 个变量上的自然推广(natural generalization),定义如下:

$$B(\alpha)*=\frac{\prod_{k=1}^K\Gamma(\alpha_k)}{\Gamma(\alpha_0)} \quad (2.76)$$

其中的$\alpha_0*=\sum_{k=1}^K\alpha_k$.

图2.14展示的是当 K=3的时候的一些狄利克雷函数图像,图2.15是一些概率向量样本.很明显其中$\alpha_0*=\sum_{k=1}^K\alpha_k$控制了分布强度,也就是峰值位置.例如Dir(1, 1, 1)是一个均匀分布,Dir(2, 2, 2) 是以为(1/3, 1/3, 1/3)中心的宽分布(broad distribution),而Dir(20, 20, 20) 是以为(1/3, 1/3, 1/3)中心的窄分布(narrow distribution).如果对于所有的 k 都有$\alpha_k<1$,那么峰值在单纯形的角落.

这个分布的属性如下:

$$E[x_k]=\frac{\alpha_k}{\alpha_0}, mode[x_k]=\frac{\alpha_k-1}{\alpha_0-K}, var[x_k]=\frac{\alpha_k(\alpha_0-\alpha_k)}{\alpha_0^2(\alpha_0+1)} \quad (2.77)$$

上式中的$\alpha_0=\sum_k\alpha_k$.通常我们用对称的狄利克雷分布,$\alpha_k=\alpha/K$. 这样则有方差$var[x_k]=\frac{K-1}{K^2(\alpha+1)}$.

这样增大α就能降低方差,提高了模型精度.

2.6 随机变量变换

如果有一个随机变量$x \sim p()$,还有个函数$y = f(x)$,那么 y 的分布是什么?这就是本节要讨论的内容.

2.6.1 线性变换

设$f()$是一个线性函数:
$y = f(x) = Ax + b$(2.78)

这样就可以推导 y 的均值和协方差了.首先算均值如下:
$E[y] = E[Ax + b] = A\mu + b$(2.79)

上市中的$\mu = E[x]$.这就叫线性期望(linearity of expectation).如果$f()$ 是一个标量值函数(scalar-valued function)$f(x) = a^T x + b$,那么对应结果就是:

$E[a^T x + b] = a^T \mu + b$(2.80)

对于协方差,得到的就是:

$cov[y] = cov[Ax + b] = A\Sigma A^T$(2.81)

其中的$\Sigma = cov[x]$, 这个证明过程留作联系.如果$f()$ 是一个标量值函数(scalar-valued function),这个结果就成了:
$var[y] = var[a^T x + b] = a\Sigma a^T$(2.82)

这些结果后面的章节都会多次用到.不过这里要注意,只有x 服从高斯分布的时候,才能单凭借着均值和协方差来定义 y 的分布.通常我们必须使用一些技巧来对 y 的完整分布进行推导,而不能只靠两个属性就确定.

2.6.2 通用变换

如果 X 是一个离散随机变量, $f(x) = y$, 推导 y 的概率质量函数 pmf,只要对所有x 的概率值了加到一起就可以了, 如下所示:
$p_y(y) = \sum_{x:f(x)=y} p_x(x)$(2.83)

例如,若X是偶数则$f(X) = 1$,奇数则$f(X) = 0$,$p_x(X)$是在集合$\{1, \ldots, 10\}$上的均匀分布(uniform),这样$p_y(1) = x \in \{2, 4, 6, 8, 10\}, p_x(x) = 0.5, p_y(0) = 0.5$.注意这个例子中的函数 f 是多对一的函数.
如果 X 是连续的随机变量,就可以利用公式2.83,因为$p_x(x)$是一个密度,而不是概率质量函数了,也就不能把密度累加起来了. 这种情况下用的就是累积密度函数 cdf 了,协作下面的形式:

$$P_y(y) * = P(Y \le y) = P(f(X) \le y) = P(X \in \{x | f(x) \le y\}) \quad (2.84)$$

对累积密度函数 cdf 进行微分,就能得到概率密度函数 pdf 了:

$$P_y(y) * = P(Y \le y) = P(X \le f^{-1}(y)) = P_x(f^{-1}(y)) \quad (2.85)$$

求导就得到了:

$$p_y(y) * = \frac{d}{dy} P_y(y) = \frac{d}{dy} P_x(f^{-1}(y) = \frac{dx}{dy} \frac{d}{dx} P - X(x) = \frac{dx}{dy} px(x) \quad (2.86)$$

显然 $x = f^{-1}(y)$,可以把 dx看作是对 x 空间的一种体测量;类似的把dy当作对 y 空间体积的测量.这样$\frac{dx}{dy}$就测量了体积变化.由于符号无关紧要,所以可以取绝对值来得到通用表达式:

$$p_y(y) = p_x(x) | \frac{dx}{dy} | \quad (2.87)$$

这也叫变量转换公式(change of variables formula).按照下面的思路来理解可能更容易.落在区间$(x, x + \delta x)$的观测被变换到区间$(y, y + \delta y)$, 其中$p_x(x)\delta x \approx p_y(y)\delta y$.因此$p_y(y)\ approx p_x(x) | \frac{\delta x}{\delta y} |$.例如,假如有个随机变量$X \sim U(-1, 1), Y = X^2$.那么则有$p_y(y) = \frac{1}{2} y^{-\frac{1}{2}}$. 具体看练习2.10.

2.6.2.1 变量的多重变化(Multivariate change of variables)

前面的结果可以推到多元分布上.设f是一个函数,从R^n映射到R^n, 设$y = f(x)$. 那么就有这个函数的雅可比矩阵 J(Jacobian matrix):

$$J_{x \to y} * = \frac{\partial(y_1, \ldots, y_n)}{\partial(x_1, \ldots, x_n)} * = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \cdots & \cdots & \cdots \\ \frac{\partial y_n}{\partial x_1} & \cdots & \frac{\partial y_n}{\partial x_n} \end{pmatrix}$$

(2.88)

矩阵 J 的行列式|det J|表示的是在运行函数 f 的时候一个单位的超立方体的体积变化.
如果 f 是一个可逆映射(invertible mapping),就可以用逆映射$y \to x$的雅可比矩阵(Jacobian matrix)来定义变换后随机变量的概率密度函数(pdf)

$$p_y(y) = p_x(x) | det(\frac{\partial x}{\partial y}) | = p_x(x) | detJ_{y \to x} \quad (2.89)$$

在练习4.5,你就要用到上面这个公式来推导一个多元正态分布的归一化常数(normalization

constant).

举个简单例子,假如要把一个概率密度函数从笛卡尔坐标系(Cartesian coordinates)的$x = (x_1, x_2)$转换到一个极坐标系(polar coordinates)$y = (r, \theta)$, 其中有对应关系:$x_1 = r\cos\theta, x_3 = r\sin\theta$.这样则有雅可比矩阵如下:

$$J_{y \to x} = \begin{pmatrix} \frac{\partial x_1}{\partial r} & \frac{\partial x_1}{\partial \theta} \\ \frac{\partial x_2}{\partial r} & \frac{\partial x_2}{\partial \theta} \end{pmatrix} = \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix}$$

(2.90)

矩阵 J 的行列式为:

$$|detJ| = |r\cos^2\theta + r\sin^2\theta| = |r|$$(2.91)

因此:

$$p_y(y) = p_x(x)|detJ|$$(2.92)

$$p_{r,\theta}(r, \theta) = p_{x_1,x_2}(x_1, x_2)r = p_{x_1,x_2}(r\cos\theta, r\sin\theta)r$$(2.93)

以几何角度来看,可以参考图2.16,其中的阴影部分面积可以用如下公式计算:

$$P(r \le R \le r + dr, \theta \le \Theta \le \theta + d\theta) = p_{r,\theta}(r, \theta)drd\theta$$(2.94)

在这个限制范围内,这也就等于阴影中心部分的密度$p(r, \theta)$乘以阴影部分的面积,$rdrd\theta$.因此则有:

$$p_{r,\theta}(r, \theta)drd\theta = p_{x_1,x_2}(r\cos\theta, r\sin\theta)rdrd\theta$$$(2.95)

此处查看原书图2.16

此处查看原书图2.17

2.6.3 中心极限定理(Central limit theorem)

现在设想有 N 个随机变量,概率密度函数(pdf)为p(x_i),且不一定是正态分布,每个的均值和方差都分别是μ, σ^2.然后假设每个随机变量都市独立同分布的(independent and identically distributed,缩写成iid).设$S_N = \sum_{i=1}^{N} X_i$是所有随机变量的和.这是一个很简单的变换,但应用很广.随着 N 的增大,这个和的分布会接近:

$$p(S_N = s) = \frac{1}{\sqrt{2\pi N\sigma^2}}\exp(-\frac{(s-N\mu)^2}{2N\sigma^2}) \quad (2.96)$$

所以这个量的分布就是:

$$Z_N* = \frac{S_N - N_\mu}{\sigma\sqrt{N}} = \frac{\bar{X}-\mu}{\sigma/\sqrt{N}} \quad (2.97)$$

这个分布就会收敛到标准正态分布了,其中样本均值为:$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$.这就叫做中心极限定理,更多内容参考(Jaynes 2003, p222) 或者 (Rice 1995, p169).

图2.17即是一例,其中计算β分布变量均值,右图可见很快收敛到正态分布了.

2.7 蒙特卡罗近似方法(Monte Carlo approximation)

要计算一个随机变量的函数的分布,靠公式变换 通常还挺难的.有另外一个办法,简单又好用.首先就是从分布中生成 S 个样本,就把它们标为$x_1, \ldots, x_S$.生成样本有很多方法,对于高维度分布来说最流行的方法就是马尔科夫链蒙特卡罗方法(Markov chain Monte Carlo,缩写为 MCMC),这部分内容在本书24章再行讲解.
还说这个例子,对分布函数$f(X)$使用经验分布(empirical distribution)$\{f(x_s)\}_{s=1}^{S}$来进行近似.这就叫蒙特卡洛近似(Monte Carlo approximation), 之所以用这个名字是因为欧洲的知名赌城.这种方法首先是在统计物理性里面应用发展起来的,确切来说还是在原子弹研究过程中,不过现在已经广泛应用在统计和机器学习领域里面了.

此处查看原书图2.18

应用蒙特卡罗方法,可以对任意的随机变量的函数进行近似估计.先简单取一些样本,然后计算这些样本的函数的算术平均值(arithmetic mean).这个过程如下所示:

$$E[f(X)] = \int f(xp(x)dx \approx \frac{1}{S}\sum_{s=1}^{S} f(x_s) \quad (2.98)$$

上式中$x_s \sim p(X)$.这就叫做蒙特卡罗积分(Monte Carlo integration),相比数值积分(numerical integration)的一个优势就是在蒙特卡罗积分中只在具有不可忽略概率的地方进行评估计算,而数值积分会对固定网格范围内的所有点的函数进行评估计算.

通过调整函数$f()$,就能对很多有用的变量进行估计,比如:

* $\bar{x} = \frac{1}{S}\sum_{s=1}^{S} x_s \rightarrow E[X]$
* $\frac{1}{S}\sum_{s=1}^{S}(x_s - \bar{x})^2 \rightarrow var[X]$
* $\frac{1}{S}\#\{x_s \le c\} \rightarrow P(X \le c)$
* 中位数(median)$\{x_1, \ldots, x_S\} \rightarrow median(X)$

下面是一些例子,后面一些章节有更详细介绍.

2.7.1 样例:更改变量,使用 MC (蒙特卡罗)方法

在2.6.2，我们讨论了如何分析计算随机变量函数的分布$y = f(x)$.更简单的方法是使用蒙特卡罗方法估计.例如,若$x \sim Unif(-1, 1), y = x^2$.就可以这样估计$p(y)$:从 $p(x)$ 中去多次取样,取平方,计算得到的经验分布.如图2.18所示.后文中还要广泛应用这个方法.参考图5.2.

此处查看原书图2.19

2.7.2 样例:估计圆周率π,使用蒙特卡罗积分

蒙特卡罗方法还可以有很多种用法,不仅仅是统计学领域.例如我们可以用这个方法来估计圆周率π.我们知道圆的面积公式可以利用圆周率和圆的半径 r 来计算,就是πr^2,另外这个面积也等于下面这个定积分(definite integral):

$I = \int_{-r}^{r}\int_{-r}^{r} \prod(x^2 + y^2 \leq r^2)dxdy$(2.99)

因此有 $\pi = I/(r^2)$.然后就可以用蒙特卡罗积分来对此进行近似了.设$f(x, y) = \prod(x^2 + y^2 \leq r^2)$是一个指示器函数(indicator function),只要点在圆内,则函数值为1,反之为0,然后设$p(x), p(y)$都是在闭区间[-r,r]上的均匀分布(uniform distribution),所以有$p(x) = p(y) = \frac{1}{2r}$ 这样则有:

$$I = (2r)(2r)\int\int f(x, y)p(x)p(y)dxdy \qquad (2.100)$$

$$= 4r^2\int\int f(x, y)p(x)p(y)dxdy \qquad (2.101)$$

$$= 4r^2\frac{1}{S}\sum_{s=1}^{S} f(x_s, y_s) \qquad (2.102)$$

当标准差为0.09的时候,计算得到的圆周率为$\hat{\pi} = 3.1416$,参考本书2.7.3就知道什么是标准差了.接受/拒绝的点如图2.19中所示.

此处查看原书图2.20

2.7.3 蒙特卡罗方法的精确度

随着取样规模的增加,蒙特卡罗方法的精度就会提高,如图2.20所示,在图上部是从一个高斯分布中取样的直方图,底下的两个图使用了核密度估计(kernel density estimate, 参考本书14.7.2)得到的光化曲线.这种光滑分布函数在密集网格点上进行评估和投图.这里要注意一点,光滑操作只是为了投图看而已,蒙特卡罗方法估计的过程根本用不着光滑.

如果我们知道了均值的确切形式,即$\mu = E]f(X)]$,然后蒙特卡罗方法近似得到的是$\hat{\mu}$,那么对于独立取样则有:

$(\hat{\mu} - \mu) \rightarrow N(0, \frac{\sigma^2}{S})$(2.103)

其中:
$\sigma^2 = var[f(X)] = E[f(X)^2] - E[f(X)]^2$(2.104)

这是由中心极限定理决定的.当然了,上式中的σ^2是位置的,但也可以用蒙特卡罗方法来估计出来:
$\hat{\sigma}^2 = \frac{1}{S}\sum_{s=1}^{S}(f(x_s) - \hat{\mu})^2$(2.105)

然后则有:
$P\{\mu - 1.96\frac{\hat{\sigma}}{\sqrt{S}} \leq \hat{\mu} \leq \mu + 1.96\frac{\hat{\sigma}}{\sqrt{S}}\} \approx 0.95$(2.106)

上式中的$\frac{\hat{\sigma}}{\sqrt{S}}$就叫做数值标准差或者经验标准差(numerical or empirical standard error), 这个量是对我们对μ估计精度的估计.具体信息查看本书6.2有更多讲解.

如果我们希望得到的答案$\pm\epsilon$范围内的概率至少为95%,那就要保证取样数目 S 满足条件 $1.96\sqrt{\hat{\sigma}^2/S} \leq \epsilon$, 这里的1.96可以粗略用2替代,这样就得到了$S \geq \frac{4\hat{\sigma}^2}{\epsilon^2}$.

2.8 信息理论

信息理论(information theory)关注的是以紧凑形式进行数据呈现(这种紧凑形式也被称为数据压缩(data compression)或者源编码(source coding)),以及以能健壮应对错误的方式进行传输和存储(这个过程也叫做纠错(error correction) 或者信道编码(channel coding)).第一眼看上去好像这和概率论以及机器学习没什么关系,不过实际是有联系的.首先,对数据进行紧凑表达需要给高概率的字符串赋予短编码字,而将长编码字留给低概率字符串.就好比自然语言中,特别常用的词汇都往往比少见的词汇短很多,比如冠词 a/the 明显比闪锌矿 sphalerite 短很多.另外,在噪音频道上进行信息解码也需要对人发送的不同信息建立一个良好的概率模型.这就都需要一个能够预测数据可能性的模型,这也是机器学习里面的一个核心问题,关于信息理论和机器学习之间关系的更多内容请参考(MacKay 2003).

显然这里不可能说太多太深关于信息理论的内容,有兴趣的话去看(Cover and Thomas 2006).这里也就是介绍本书中要用到的一些基础概念了.

2.8.1 信息熵

随机变量 X 服从分布 p, 这个随机变量的熵(entropy)则表示为$H(X)$或者$H(p)$,这是对随机变量不确定性的一个衡量.对于一个有 K 个状态的离散随机变量来说,其信息熵定义如下:

$H(X) * = -\sum_{k=1}^{K} p(X = k)\log_2 p(X = k)$(2.107)

通常都用2作为对数底数,这样单位就是 bit (这个是 binary digits 的缩写).如果用自然底数 e, 单位就叫做 nats 了.
举个例子,$X \in \{1, \ldots, 5\}$,柱状分布(histogram distribution), 概率$p = [0.25, 0.25, 0.2, 0.15, 0.15]$,利用上面的公式计算得到$H = 2.2855$.
熵最大的离散分布就是均匀分布,可以参考本书9.2.6.因为对于一个 K 元(K-ary)随机变量,如果$p(x = k) = 1/K$,则信息熵最大,这时候的熵为$H(X) = \log_2 K$. 熵最小的分布就是所有概率质量都在单一状态的δ分布,这时候熵为0,因为只有一个状态有概率,没有任何不确定性.

在图2.5当中对 DNA 序列进行了投图,每一列的高度定义为$2 - H$,其中的 H 就是这个分布的熵,2是最大可能熵(maximum possible entropy).因此高度为0的就表示均匀分布,而高度为2就对应着确定性分布(deterministic distribution).

此处查看原书图2.21

对于二值化随机变量的特例,$X \in \{0, 1\}$,则有$p(X = 1) = \theta, p(X = 0) = 1 - \theta$,这样熵为:

$$H(X) = -[p(X = 1)\log_2 p(X = 1) + p(X = 0)\log_2 p(X = 0)] \quad (2.108)$$
$$= -[\theta \log_2 \theta + (1 - \theta)\log_2 (1 - \theta)] \quad (2.109)$$

这也叫做二值熵函数(binary entropy function),也写作$H(\theta)$,如图2.21所示,课件当$\theta = 0.5$的时候熵值最大为1,这时候是均匀分布了.

2.8.2 KL 散度

KL 散度(Kullback-Leibler divergence),也称相对熵(relative entropy),可以用来衡量p和q两个概率分布的差异性(dissimilarity).定义如下:

$$KL(p||q) * = \sum_{k=1}^{K} p_k \log \frac{p_k}{q_k} \quad (2.110)$$

上式中的求和也可以用概率密度函数的积分来替代.就可以写成:

$$KL(p||q) = \sum_k p_k \log p_k - \sum_k p_k \log q_k = -H(p) + H(p, q) \quad (2.111)$$

上式中的$H(p, q)$就叫做交叉熵(cross entropy):

$$H(p, q) * = -\sum_k p_k \log q_k \quad (2.112)$$

参考 (Cover and Thomas 2006) 可以证明,当使用模型 q 来定义编码本(codebook)的时候,来自分布 p 的待编码数据的平均比特数(average number of bits)就是交叉熵.正规熵(regular entropy)$H(p) = H(p, p)$,参考本书2.8.1的定义,也就是使用真实模型时候的比特数期望值,因此 KL 散度也就是不同概率分布之间的不同的量度.换个说法, KL 散度就是要对数据编码所需要的额外比特(extra bits)的平均数,因为这时候用分布 q 来对数据进行编码,而不是使用分布 p.

既然是额外的比特数,这种表述就很明显说明这个 KL 散度应该是非负的,即$KL(p||q) \geq 0$,等于0则意味着两个分布相等,即$p = q$.接下来对此进行一下证明.

定理2.8.1 信息不等式(Information inequality)

$KL(p||q) \geq 0$ 当且仅当 $p = q$的时候, KL 散度为0.

证明

要证明这个定理,需要用到詹森不等式(Jensen's inequality).这个不等式是说,对于任意的凸函数(convex function) f,有以下关系:

$f(\sum_{i=1}^{n}\lambda_i(x_i)) \leq \sum_{i=1}^{n}\lambda_i f(x_i)$(2.113)

其中$\lambda_i \geq 0, \sum_{i=1}^{n}\lambda_i = 1$. 由于凸函数的定义,对于 n=2的时候很显然,对于 n>2 的情况也可以归纳证明(proved by induction).

对定理的证明参考了(Cover and Thomas 2006, p28).设$A = \{x : p(x) > 0\}$是$p(x)$ 的支撑集合(support, 译者注:纯白或许就当做定义域理解好了).则有:

$$-KL(p||q) = -\sum_{x\in A} p(x)\log\frac{p(x)}{q(x)} = \sum_{x\in A} p(x)\log\frac{q(x)}{p(x)} \qquad (2.114)$$

$$\leq \sum_{x\in A} p(x)\frac{q(x)}{p(x)} = \log\sum_{x\in A} q(x) \qquad (2.115)$$

$$\leq \log\sum_{x\in \chi} q(x) = \log 1 = 0 \qquad (2.116)$$

当 上面第一个不等式是应用了詹森不等式.因为$\log(x)$是个严格凸函数,所以在等式2.115里面,当且仅当对于某些 c 使$p(x) = cq(x)$成立的时候,等量关系成立.等式2.116中的等量关系当且仅当$\sum_{x\in A} q(x) = \sum_{x\in \chi} q(x) = 1$的时候成立,这时候 c=1. 所以对于所有的 x 来说,当且仅当$p(x) = q(x), KL(p||q) = 0$.

证明完毕.

这个结果的一个重要推论就是*足有最大熵的离散分布就是均匀分布*.更确切地说,$H(X) \leq \log|\chi|$, $|\chi|$是 X 的状态数,当且仅当$p(x)$是均匀分布的时候等号成立.设$u(x) = 1/|\chi|$,则有:

$$0 \leq KL(p||u) = \sum_x p(x)\log\frac{p(x)}{u(x)} \qquad (2.116)$$

$$= \sum_x p(x)\log p(x) - \sum_x p(x)\log u(x) = -H(X) + \log|\chi| \qquad (2.118)$$

上面这个就是公式化的拉普拉斯不充分理由原则(Laplace's principle of insufficient reason),说的是在没理由优先选择某个分布的时候,优先选择均匀分布(uniform distribution).关于如何建立满足特定约束条件(certain constraints) 的分布可以阅读本书9.2.6,其他方面尽可能最小化(as least-commital as possible).(正态分布满足一阶和二阶矩约束条件,但其他方面有最大熵.)

2.8.3 信息量(Mutual information)

设有两个随机变量 X 和 Y. 如果我们想知道一个变量告诉我们关于另一个变量的多少信息。就可以计算相关系数(correlation coefficient)了,可是相关系数只适用于实数值的随机变量.另外相关系数对不相关程度的衡量作用也很有限,如图2.12所示.所以更常用的方法是对比联合分布(joint distribution)$p(X, Y)$和因式分布(factored distribution)$p(X)p(Y)$的相关性.这就叫互信息量(mutual information) 或者简写做 MI, 定义如下:

$$I(X; Y) * = KL(p(X, Y)||p(X)p(Y)) = \sum_x \sum_y p(x, y) \log \frac{p(x,y)}{p(x)p(y)} \quad (2.119)$$

$I(X; Y) \geq 0$的等号当且仅当$p(X, Y) = p(X)p(Y)$的时候成立.也就是如果两个变量相互独立,则互信息量 MI 为0. 为了深入理解 MI 这个量的含义,咱们用联合和条件熵的方式来重新表述一下.参考练习2.12可以得到上面的表达式等价于下列形式:

$$I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \quad (2.120)$$

上式中的$H(Y|X)$就是条件熵(conditional entropy) 定义为$H(Y|X) = \sum_x p(x)H(Y|X = x)$.这样就可以把 X 和 Y 之间的互信息量 MI 理解成在观测了 Y 之后对 X 的不确定性的降低,或者反过来就是观测了 X 后对 Y 不确定性的降低.本书后面一些内容中还会用到这个概念.参考2.13和2.14来阅读互信息量 MI 和相关系数之间的联系.

另外一个和互信息量 MI 有很密切联系的量是点互信息量(pointwise mutual information,缩写为 PMI), 对于两个事件(而不是随机变量) x 和 y,其点互信息量 PMI 定义如下:

$$PMI(x, y) * = \log \frac{p(x,y)}{p(x)p(y)} = \log \frac{p(x|y)}{p(x)} = \log \frac{p(y|x)}{p(y)} \quad (2.121)$$

这个量衡量的是与偶发事件相比,这些事件之间的差异.很明显 X 和 Y 的互信息量 MI 就是点互信息量 PMI 的期望值.所以就可以把点互信息量 PMI 的表达式写为:

$$PMI(x, y) = \log \frac{p(x|y)}{p(x)} = \log \frac{p(y|x)}{p(y)} \quad (2.122)$$

这个量是通过将先验(prior)的 $p(x)$ 更新到后验(posterior)的$p(x|y)$得到的,也可以是将先验的 $p(y)$ 更新到后验的$p(y|x)$得到.

2.8.3.1 连续随机变量的互信息量

上一节中的互信息量 MI 定义是针对离散随机变量的.对于连续随机变量,可以先对其进行离散化

(discretize)或者量子化(quantize),具体方法可以使将每个随机变量归类到一个区间里面,将变量的变化范围划分出来,然后计算每一段的小区间中的分布数量(Scott 1979).然后就可以利用上文的方法公式来计算互信息量 MI 了(代码参考PMTK3的 mutualInfoAllPairsMixed, 样例可以参考 miMixedDemo).

此处查看原书图2.22

然而很不幸,分成多少个小区间,以及小区间边界的位置,都可能对计算结果有很大影响.一种解决方法就是直接对互信息量 MI 进行估计,而不去先进行密度估计(Learned-Miller 2004)).另一种办法是尝试很多不同的小区间规模和位置,然后计算得到的最大互信息量 MI.经过适当的标准化之后,这个统计量就被称为最大信息系数(maximal information coefficient,缩写为 MIC)(Reshed et al. 2011).更确切来说定义如下所示:

$$m(x,y) = \frac{\max_{G \in G(x,y)} I(X(G);Y(G))}{\log \min(x,y)} \quad (2.123)$$

上式中的$G(x,y)$是一个规模为$x \times y$ 的二维网状集合,而$X(G)$, $Y(G)$表示的是将变量在这个网格上进行离散化得到的结果.在区间位置(bin locations)上最大化的过程可以通过使用动态编程(dynamic programming)来有效进行(Reshed et al. 2011).这样定义了连续变量互信息量 MIC如下:

$$MIC* = \max_{x,y:xy<B} m(x,y) \quad (2.124)$$

上式中的 B 是一个与取样规模相关的约束条件,用于约束能使用且能可靠估计分布的区间个数.((Reshed et al. 2011) 建议的是$B = N^{0.6}$.显然 MIC 处于区间[0, 1]中,其中-表示两个变量没关系,而1表示二者有无噪音的相关性(noise-free relationship),这种相关性可以是任意形式的,不仅仅是线性相关.

图2.22给出的是一个实例.其中的数据集包含了357个变量,衡量一系列的社会/经济/健康/政治指标,由世界健康组织 WHO 手机.左边的途中看到了对于 65,566 个变量对的相关系数(CC)与互信息量(MIC)的关系图.有图则投了一些特定变量对的散点图,其中包括了:
* C 图中的是 CC 和 MIC 都低,相应的散点图很明显表明了这两组变量之间没有关系:因伤致死比例和人群中牙医密度.
* D 图和 H 图中是 CC 和 MIC 都高,呈现近乎线性的相关性.
* E/F/G 这三个图都是低 CC 高 MIC.这是因为这些变量之间的关系是非线性的,例如在 E 图和 F 图中,都是非函数对应关系,比如可能是一对多的对应关系.

总的来说, MIC 这个统计量是基于互信息量的,可以用于发现变量之间的有意义的关系,而这些关系可能是那些简单的统计量,比如相关系数之类无法反应的.由于这个优势, MIC 也被称作是21世纪的相关性衡量变量“a correlation for the 21st century” (Speed 2011).

练习 2

尼玛太长了,回头再说吧,先去弄下一章了.