MLAPP-C2

2.概率论

2.1 前言

2.2. 关于概率论的简要综述

2.2.1 离散随机变量

2.2.2 基本定理

2.2.3 贝叶斯法则

2.2.4 独立与条件独立

2.2.5 连续随机变量

2.2.6 分位数

2.2.7 期望和方差

2.3 一些常见的离散分布

2.3.1 二项式分布和伯努利分布

2.3.2 多项式和multinoulli分布

2.3.3 泊松分布

2.3.4 经验分布

2.4 一些常见的连续变量分布

2.4.1 高斯（正态）分布

2.4.2 退化概率密度分布

2.4.3 学生t分布

2.4.4 拉普拉斯分布

2.4.5 伽玛分布

2.4.6 贝塔分布

2.4.7 帕雷托分布

2.5 联合概率分布

2.5.1 协方差与相关系数

2.5.2 多变量高斯分布

2.5.3 多变量学生t分布

2.5.4 狄利克雷分布

2.6 随机变量的变换

2.6.1 线性变化

2.6.2 一般变换

2.6.3 中心极限定理

2.7 蒙特卡洛近似

2.7.1 例子：变量变换法，蒙特卡洛法

2.7.2 例子：通过蒙特卡洛方法对进行近似

2.7.3 蒙特卡洛近似的精度

2.8 信息论

2.8.1 熵

2.8.2 KL散度

2.8.3 互信息

2.概率论

2.1 前言

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. 关于概率论的简要综述

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。

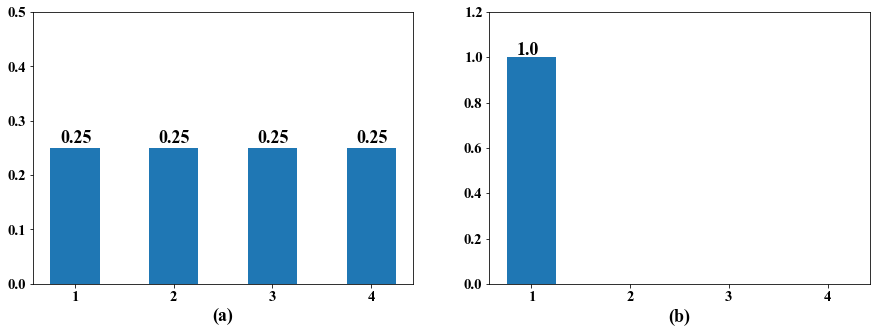


图2.1 离散变量的概率质量分布。图片由程序**discreteProbDistFig**生成。

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*可能是一个表述“明天会下雨”。*p*(*A*)满足0≤*p*(*A*)≤1，其中如果*p*(*A*)=0，则意味着事件*A*不可能发生，*p*(*A*)=1意味着事件*A*肯定发生。我们使用表示非*A*事件发生的可能性，满足。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（即只有两种状态）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（关于可能无限集的例子：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*X*的取值所构成的就是一个可数无限集）。我们表示事件*X*=*x*发生的概率为*p*(*X*=*x*)，或者直接写成*p*(*x*)。这里符号*p*()称为**概率质量函数**（probability mass function），满足性质0 ≤ *p*(*x* )≤ 1，。图2.1展示了定义在一个有限**状态空间**（state space）上的两种概率质量函数。左边属于均匀分布，*p*(*x*)=1/4，右边为一个退化分布，其中 () 为二元**指示函数**（indicator function），这个分布意味着*X*永远等于1。

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：

 （2.1）

式中符号分别表示并集和交集。

2.2.2.2 联合概率

我们定义事件*A*和*B*一起发生的概率为：

 （2.2）

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：

 （2.3）

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）：

 （2.4）

式中，我们模仿了Matlab中的一种符号写法*1*:*D*表示集合{1,2,…,*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*的**条件概率**（conditional probability）为：

 （2.5）

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）或者贝叶斯理论（Bayes’ Theorem）。

 （2.6）

Sir Harold Jeffreys指出“贝叶斯理论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 例子：医疗诊断

作为一个使用该法则的例子，考虑如下的医疗诊断问题。假设你是一个40多岁的女人，为了检查自己是否患有乳腺癌，去医院作一个叫mammogram的测试。如果测试结果是阳性的，那么你患有乳腺癌的概率是多大呢？显然这个答案取决于你对这个测试的信赖程度有多大。假设你已经知道这个测试的灵敏度为0.8，即如果你患病，那么这个测试是阳性的可能性为0.8。即：

 （2.7）

其中事件*x*=1表示测试结果为阳性，*y*=1表示患有乳腺癌。很多人可能会因此认为他们有80%的可能性患病。但这是错误的！因为忽略了患有乳腺癌的先验概率，幸运的是这个概率非常低：

 （2.8）

我们还需要考虑测试结果为假阳性的情况（即没有患病，但测试结果为阳性）。不幸的是，这样的假阳例是非常有可能的：

 （2.9）

根据贝叶斯法则，我们计算测试结果为阳性时，患病的可能性为：

 （2.10）

换句话说，如果检测为阳性，那么仅有3%的可能性患有癌症。

2.2.3.2 例子：生成式分类器

我们可以将医疗诊断的问题拓展到对任意特征向量**x**进行分类的问题：

 （2.11）

上式被称为**生成式分类器**（generative classifier）。因为它指定了数据**x**生成的方式——使用**类条件概率**（class-conditional density）*p*(**x**|*y*=*c*)和类先验分布*p*(*y*=*c*)。我们将在第3和4章进行深入讨论。一个代替方案是直接对类后验分布*p*(*y*=c|**x**)进行建模，这被称为**判别式分类器**（discriminative classifier）。我们将在8.6节讨论两种方式的优势和劣势。

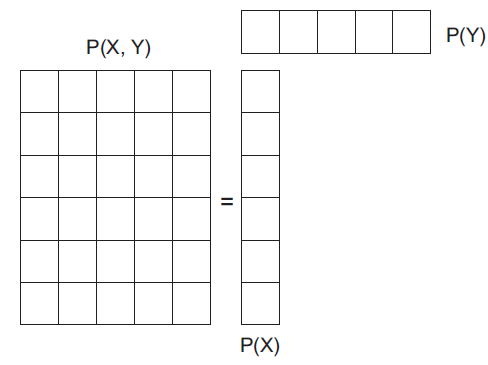


图2.2 计算*p*(*x*,*y*)=*p*(*x*)*p*(*y*)，其中。图中，X和Y为离散随机变量；*X*有6个可能状态，*Y*有5个可能状态。一般情况下的联合概率分布需要（6×5）-1=29个参数。通过假设（非条件）独立，我们仅仅需要（6-1）+（5-1）=9个参数。

2.2.4 独立与条件独立

如果变量*X*和*Y*的联合概率分布可以表示为两个边缘概率分布的乘积（如图2.2所示），则称*X*和*Y*是**非条件独立**（unconditionally independent）或**边缘独立**（marginally independent），表示为：

 （2.12）

通常情况下，我们称多个变量是相互独立的，如果它们的联合概率分布可以写成边缘概率的乘积。

不幸的是，非条件独立是很罕见的，因为大部分变量可能会彼此互相影响。然而，这种影响往往是通过某些中间变量实现而非直接影响。我们称*X*和*Y*在给定变量*Z*的情况下是**条件独立**（conditionally independent）的，如果条件联合分布可以写成条件边缘分布的乘积：

 （2.13）

当我们在第10章学习概率图模型时，我们将发现，上式可以使用图*X*-*Z*-*Y*表示。比如说，如果已经知道今天正在下雨（事件*Z*）这件事，那么明天下雨的概率（事件*X*）与今天的地是否是湿的（事件*Y*）无关。直觉上，那是因为*Z*直接引起了*X*和*Y*，所以如果我们知道了*Z*，那么事件*Y*的发生对我们判断事件*X*没有任何影响。我们将在第10章拓展这个概念。

另一个关于条件独立性的特征是：

**定理2.2.1.** 当且仅当存在函数*g*和*h*使得

 （2.14）

对于所有的*x*,*y*,*z*成立，且*p*(*z*)>0。

**附：证明上述定理：**

对式（2.14）两边分别对*x*和*y*进行积分，得到：



对上面两式相乘，得到：



对上式中的*x*,*y*进行积分，得到：



所以我们有：



故成立。

条件独立性假设允许我们基于局部信息建立大规模的概率模型。我们在书中将看到更多的例子。当我们在3.5节讨论朴素贝叶斯分类器，在17.2节讨论马尔科夫模型，在第10章讨论概率图模型时，将大量使用条件独立性假设。

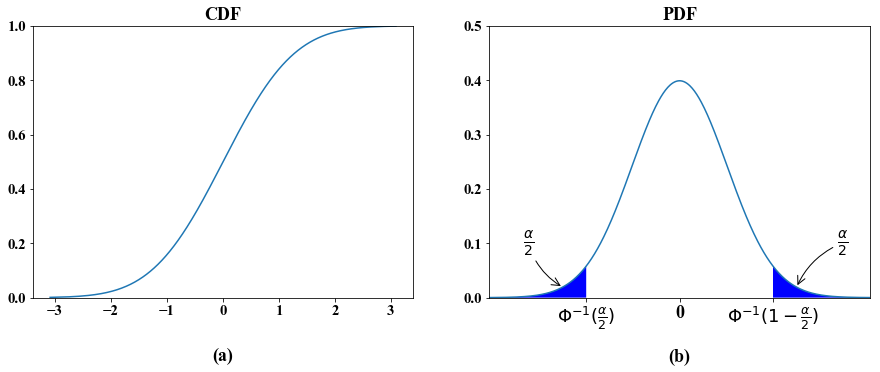


图2.3 （a）标准正态分布的累积质量函数cdf。（b）相应的概率密度函数。其中阴影部分包含的概率质量。所以剩余的部分包含的概率质量。图片由程序**quantileDemo**生成。

2.2.5 连续随机变量

截至目前，我们仅仅讨论了离散随机变量的相关性质，接下来，我们将研究连续随机变量的相关属性。

假设*X*是某个不确定的连续值。该*X*属于区间[a,b]的概率可以通过如下的方式进行计算。定义事件*A*=(*X*≤a)，*B*=(*X*≤b)，*W*=(a＜*X*≤b)。显然我们有，因为*A*和*W*是互相独立的，根据加法法则，有：

 （2.15）

所以：

 （2.16）

定义函数，这被称为*X*的**累计分布函数**（cumulative distribution function，cdf）。这是一个单调非递减函数。图2.3(a)展示了该函数的基本情况。使用这个函数，我们有：

 （2.17）

现在定义（假设该导数存在），这被称为**概率密度函数**（probability density function，pdf），如图2.3(b)所示。给定概率密度函数，我们可以计算一个连续随机变量在某个区间内的概率为：

 （2.18）

当积分区间取微元，则有：

 （2.19）

我们要求*f*(*x*)≥0，但对于任意给定的*x*，*f*(*x*)>1是有可能的，只要概率密度积分为1。（**对于连续变量我们通常只使用它的概率密度函数*f*(x)，为方便起见，我们用*p*(x)代替*f*(x)，但必须注意的是这里的*p*(x)是指在x处的概率密度。**）比如在***均匀分布***（uniform distribution）Unif(a,b)中：

 （2.20）

如果*a*=0,*b*=1/2,那么显然对于任意*x*∈[0,1/2]，*p*(x)=2。

2.2.6 分位数

因为累积分布函数*F*是单调递增函数，所以存在反函数*F*-1。如果*F*是关于*X*的累积分布函数，那么的取值满足（这里使用大写的*P*表示概率质量），被称为函数*F*的**分位数**（quantile）。被称为分布的**中位数**（median），即在该中位数的左右两侧，概率质量分别为0.5。和被称为下分位数和上分位数。

利用累积分布函数的反函数可以计算**尾域概率**（tail area probailities）。比如说，如果为高斯分布的累积分布函数，那么所有在分位点左侧的点涵盖了概率质量的，如图2.3(b)所示。根据对称原理，分位点的右侧的点包含概率质量的。所以，区间（，）内包含概率质量的1-。如果设置=0.05，那么中心95%的概率质量被如下区间包含：

（，）=（-1.96,1.96） （2.21）

如果分布是，则中心95%概率质量的分布区域为()。这个区间通常被近似为。

2.2.7 期望和方差

在概率分布中，最熟悉的统计量就是**期望**（mean），用符号表示。对于离散随机变量，定义为，对于连续随机变量，定义为（如果无界，则期望没有定义）。

**方差**（variance）是对概率分布离散程度的一种衡量，由符号表示。定义为：

 （2.22）

根据上式我们得到一个有用的结果：

 （2.23）

**标准差**（standard deviation）被定义为：

 （2.24）

这里需要注意的是标准差与随机变量*X*本身的量纲是一致的。

2.3 一些常见的离散分布

本章，我们将回顾一些定义在离散状态空间（可数和可数无限的）上的参数概率分布。

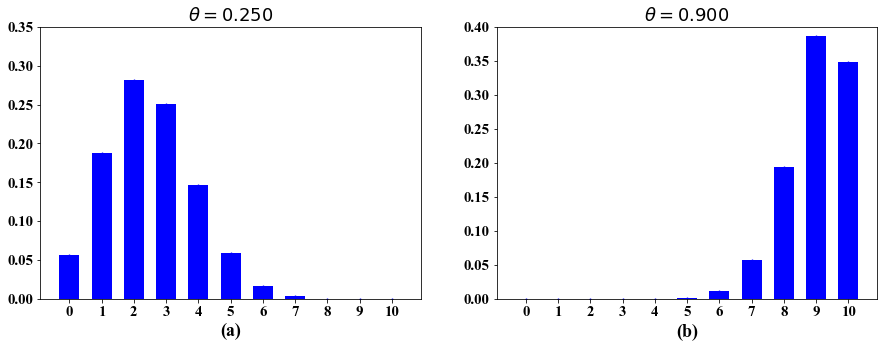


图2.4. 伯努利分布的图形，其中n=10，。图形由程序**binomDistPlot**生成。

2.3.1 二项式分布和伯努利分布

假设我们投掷一个硬币n次。令*X*∈{0,…,n}为硬币朝上的次数。如果硬币朝上的概率为，我们称*X*服从**二项式**（binomial）分布，符号上记作，其概率质量函数定义为：

 （2.25）

其中：

 （2.26）

为在*n*个项目中选择*k*个的方式的数量（被称为**二项式系数**binomial coefficient）。图2.4展示了多项式分布的一些例子。该分布的期望和方差为：

 （2.27）

现在假设我们只投掷一次硬币，令*X*∈{0,1}为二元随机变量，且硬币朝上的概率为。我们称*X*服从**伯努利**（Bernoulli）分布。符号上记作，其概率质量函数定义为：

 （2.28）

（**注：读者应该习惯于这种利用指示函数进行概率函数表达的方式，这在后面的内容中经常出现**）显然，伯努利分布只是二项式分布在*n*=1时的特例。

2.3.2 多项式和multinoulli分布

二项式分布可以用来对投掷硬币（只有两种输出结果）这种情况进行建模。如果我们投掷一个具有*K*个面的骰子n次，此时输出结果（面*k*朝上的次数）服从多项式分布。定义为一个随机向量，其中*xj*表示面 *j* 朝上的次数。随机向量**x**的概率质量函数为：

 （2.29）

式中表示面 *j* 朝上的概率，**多项式系数**（multinomial coefficient）为：

 （2.30）

现在假设*n*=1，即我们投掷一个*K*面骰子1次。在这种情况下，**x**将会是一个只有元素0和1的向量（位向量），且只有一个分量取值为1。特别的，如果骰子的面*k*朝上，那么**x**的第*k*个元素为1。在这种情况下，向量**x**可以作为一个标量*x*处理。比如说，假设骰子具有3个面，那么事件“第3个面朝上”用向量**x**表示为（0,0,1），如果使用标量表示，即*x*=3。我们称向量**x**为标量*x*的**哑编码**（dummy encoding），两者的关系表示为。比如说，如果*K*=3，我们可以将状态1,2,3分别编码为（1,0,0），（0,1,0）和（0,0,1）。这种编码方式又被称为**one-hot encoding，**因为在*K*个状态中只有一个状态处于激活（hot）。在这种情况下，概率质量函数为：

 （2.31）

图2.1显示了一些例子。这种特殊的情况又被称为**范畴**（categorical）或者**离散**（discrete）分布（Gustavo Lacerda建议称其为multinoulli distribution，以此类比于Binomial/Bernoulli分布，本书也将采用这种术语）。我们使用如下符号进行表达：

 （2.32）

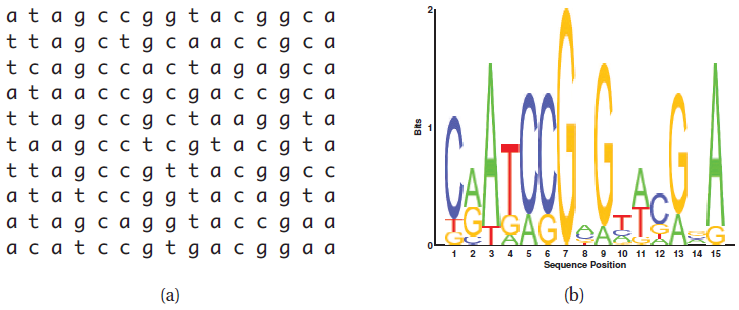


图2.5 （a）一些被对齐后的DNA序列。（b）相应的DNA图像。其中垂直方向代表2-*H*，其中*H*为对应列的分布的熵（测量单位为bit）。所以那些确定性分布（熵为0）的高度为2，均匀分布（熵为2）的高度为0。（注：关于熵的概念我们将在后面介绍）

2.3.2.1 应用：DNA序列图案

多项式分布的一个有趣的应用是在**生物序列分析**（biosequence analysis）。假设我们有一些DNA序列，如图2.5(a)所示，图中有10行（序列数目）15列（基因组上的位置）。

一种可视化图中数据的方法称为***序列标识图***（sequence logo），如图2.5（b）所示。图中字母A，C，G和T的字体大小正比于它们的经验概率，出现频次最高的字母出现在顶端。位置*t*的经验分布表示为，定义为：

 （2.33）

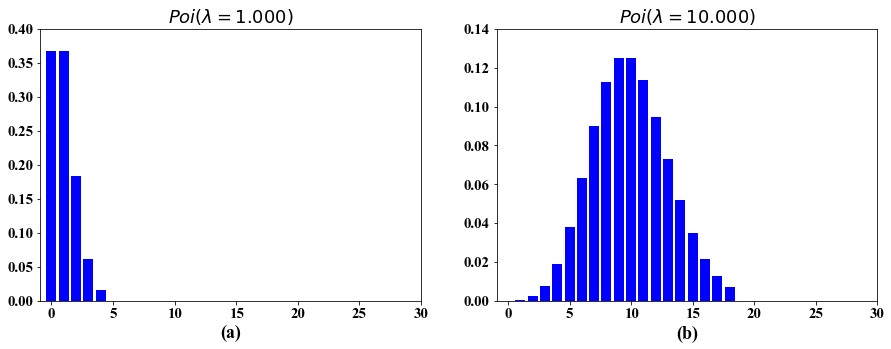


图2.6 泊松分布的说明，其中。值得注意的是我们在图中截断了x轴，因为泊松分布的定义域是整个非负整数集合。图形由程序**poissonPlotDemo**绘制。

2.3.3 泊松分布

我们称*X*∈{0,1,2，…}服从泊松分布，如果它的概率质量函数定义为：

 （2.34）

式中。上式第一项为归一化项，确保分布的和为1。

泊松分布通常应用于对小概率事件（比如放射性衰变和交通事故）的发生次数进行建模。图2.6展示了泊松分布的结果。

**附：泊松分布与二项式分布的关联**

如2.3.1节所述，二项式分布表示为：



当时，即重复一个成功率很低的试验近无数次，令保持不变。在这种情况下，二项式分布的形式为：



其中：



故：



2.3.4 经验分布

给定数据集*D*={*x*1,…,*xN*}，定义**经验分布**（empirical distribution）为：

 （2.35）

其中称为**狄拉克测度**（Dirac measure），定义为：

 （2.36）

2.4 一些常见的连续变量分布

本章我们将展示一些常用的单变量（一维）连续概率分布。

2.4.1 高斯（正态）分布

在统计和机器学习领域最常见的分布是**高斯**（Gaussian）分布或者**正态**（normal）分布。它的概率密度函数（注意在离散变量中，我们考察的是概率质量函数）定义为：

 （2.37）

其中为期望（众数），为方差，为归一化常数。

符号表示。如果，我们称*X*服从标准正态分布。图2.3（b）绘制了该概率密度函数，该曲线又被称为**钟型曲线**（bell curve）。

高斯分布的**精度**（precision）为方差的倒数，即。一个精度很高的高斯分布在图像上表现为图像高度集中在期望周边。

值得注意的是，因为我们讨论的是概率密度函数，所以我们允许*p*(*x*)>1。为了说明这一点，考虑当时的概率密度。，所以如果，我们有*p*(*x*)>1。

高斯分布的累积分布函数定义为：

 （2.38）

图2.3（a）绘制了当时的累积分布函数。

式（2.38）本身没有封闭解，但在很多软件中都有内置的实现。我们可以通过使用**误差函数**（error function，erf）进行计算：

 （2.39）

式中，且

 （2.40）

高斯分布是统计学中使用最广泛的分布，主要原因有以下几个。首先，高斯分布仅有2个参数，该参数可解释性高，且包含分布的最基本的性质（期望和方差）。第二，根据中心极限定理（见2.6.3），独立变量的和近似服从高斯分布，所以对于残差或者噪音数据来说，使用高斯分布进行建模比较合理（因为残差或者噪音一般是由多种不相关因素叠加形成）。第三，在给定数据期望和方差的前提下，高斯分布基于的额外假设最少（熵最大）（见9.2.6），促使该分布在很多情况下是一个很好的默认选择。最后，它具备简单的数学形式，从而可以用一种简单而且有效的方式实现。

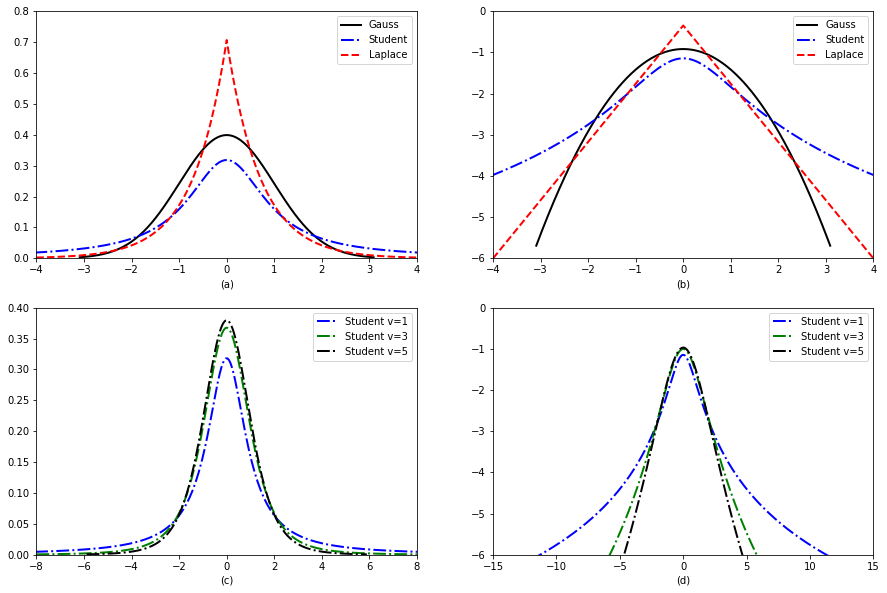


图2.7 （a）分布，和的概率密度函数。对于高斯分布和拉普拉斯分布而言，期望和方差分别为0和1.当时，学生分布的期望和方差没有定义。（b）三个分布的对数形式。值得注意的是，学生分布并不总是log-concave的，这一点与Laplace分布（总是log-concave的）是有差别的。然而三个分布都是单峰的。

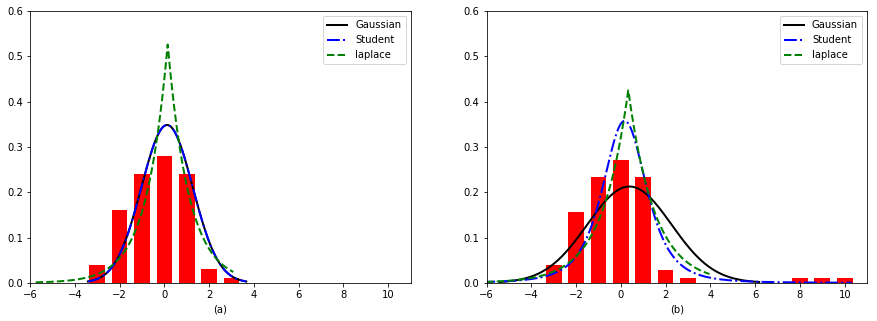


图2.8. 噪音点对高斯、学生和拉普拉斯分布拟合的影响。（a）没有噪音点；（b）含噪音点：高斯分布相较于其他两个分布，受噪音的影响更大。图像由程序**robustDemo**生成。

2.4.2 退化概率密度分布

当，高斯分布在图形上将变得无限高瘦，且分布集中在期望周边：

 （2.41）

其中被称为**狄拉克脉冲函数**（Dirac delta function），定义为：

 （2.42）

同时：

 （2.43）

脉冲函数的一个有用性质是**平移性质**（sifting property）：

 （2.44）

2.4.3 学生*t*分布

高斯分布的一个问题在于它对异常值比较敏感，因为概率密度的对数值与数据点到中心距离的平方呈线性关系。一个更加健壮的分布被称为**学生t分布**（Student’s *t* distribution），简称为学生分布（尽管更常见的叫法为*t*-分布）。它的概率密度定义为：

 （2.45）

其中被称为期望，被称为尺度参数，被称为**自由度**（degrees of freedom）。图2.7绘制了相关的图形。该分布具备如下性质：

 （2.46）

其中方差只在时有定义，期望只在时有定义。

为了说明学生分布的鲁棒性，考虑图2.8.左侧，我们利用一个高斯分布和一个学生分布对一个没有异常值的数据分布建模。图形右侧，我们添加了一些异常值，不难发现高斯分布发生了较大的变化，然而学生分布受影响很小。这是因为学生分布在曲线尾部依然保持着较高的概率值（至少当比较小时是这样的）。

如果，该分布被称为**柯西**（Cauchy）或**洛伦兹**（Lorentz）分布。这种情况是需要被注意的，因为此时的分布尾部较重（即概率值依然很高），导致求解期望的积分无法收敛。

为了保证方差的有限性，我们要求。通常情况下使用，这个设置在很多问题上都有比较好的性能。当时，学生分布将很快地接近高斯分布（从图2.7（c）可以看出），从而失去它的鲁棒性。

2.4.4 拉普拉斯分布

另一个尾部较重的分布是**拉普拉斯分布**（Laplace distribution），该分布又被称为**双边指数**（double sided exponential）分布。其概率密度函数定义为：

 （2.47）

其中被称为位置参数，*b*>0被称为尺度参数。图2.7绘制了该概率密度函数。该函数具有如下性质：

 （2.48）

该分布对噪音的鲁棒性如图2.8所示。**相比于高斯分布，该分布在x=0处分配的概率密度更高。这个性质将促使模型变得稀疏化**，关于这一点我们将在13.3节看到。

2.4.5 伽玛分布

**伽玛**（gamma）分布是一个关于随机变量*T* > 0的灵活分布。该分布由两个参数确定，分别称为**形状**（shape）*a* > 0和**速率**（rate）*b* > 0：

 （2.49）

式中被称为伽玛函数：

 （2.50）

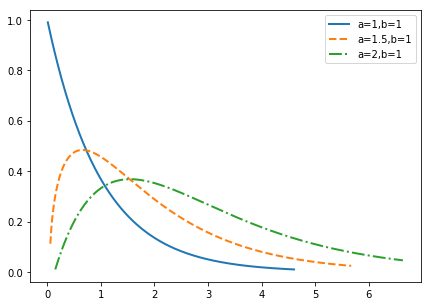


图2.9 不同形式的gamma分布，图形由程序**gammaRainfallDemo**生成。

图2.9绘制了不同的伽玛分布。该分布具备如下的性质：

 （2.51）

有些分布是Gamma分布的特例，我们将在后面进行讨论，主要包括：

* **指数分布（**Exponential distribution**）** 该分布定义为，其中被称为速率参数。指数分布（也称为负指数分布）是描述泊松过程中的事件之间的时间的概率分布，泊松过程是指事件以恒定平均速率连续且独立地发生的过程。
* **厄兰分布**（Erlang distribution）当Gamma分布中的参数*a*为整数时，对应的分布为厄兰分布，通常情况下，固定*a*=2，得到单参数厄兰分布，，其中为速率参数。
* **卡方分布**（Chi-squared distribution） 该分布定义为，它描述了高斯随机变量平方和的分布。更加明确的表达为：如果，那么随机变量。

另一个有用的结果为：如果，则，其中IG被称为**逆伽玛**（inverse gamma）分布，定义为：

 （2.52）

逆伽玛分布的性质包括：

 （2.53）

其中期望值只在*a* > 1时有定义，方差只在*a* > 2时有定义。

我们将会在后面介绍这些分布的应用。

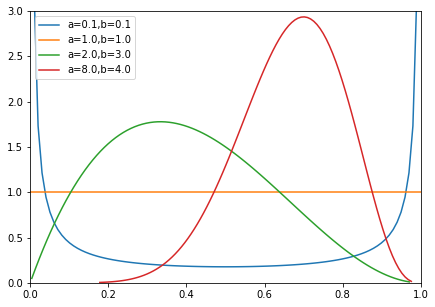


图2.10. 一些beta分布，图形由程序**betaPlotDemo**生成

2.4.6 贝塔分布

贝塔分布的定义域为[0,1]，其概率密度函数为：

 （2.54）

式中*B*(*a*,*b*)为贝塔函数，定义为：

 （2.55）

图2.10展示了一些beta分布。我们要求*a*,*b*>0以使得该分布可积（确保*B*(*a*,*b*)存在）。如果*a*=*b*=1，我们可以得到均匀分布。如果*a*和*b*皆小于1，我们得到一个双峰分布，峰值出现在0和1处；如果*a*和*b*都大于1，该分布为单峰分布。该分布具有如下性质：

 （2.56）

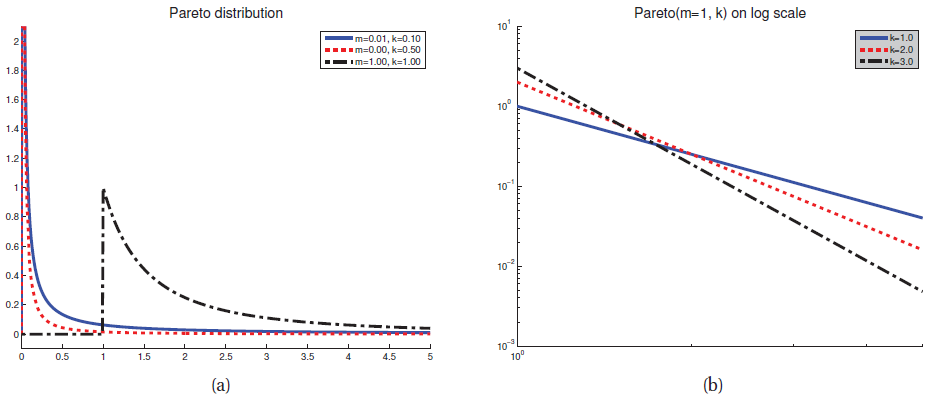


图2.11 （a）帕累托分布。（b）log-log尺度下的概率密度函数。

2.4.7 帕雷托分布

**帕累托分布**（Pareto distribution）定义为：

 （2.57）

这个密度函数要求*x*必须大于某个常数m，但也不能太大，其中*k*值控制着这个“太大”的幅度。当时，该分布逼近。图2.11(a)展示了该分布的图形。该分布具有如下的性质：

 （2.58）

当m=0时，该分布退化为，其中*a*= - (*k*+1)。如果我们在对数坐标系中绘制该分布，则显示为一条直线，该直线定义为，如图2.11(b)所示，这被称为**幂律分布定理**（power law）。当我们对那些拥有**长尾**（long tails）或者**重尾**（heavy tails）效应的数据进行建模时，这个定理十分有用。举例来说，据观察，在英语中使用频次最高的单词（“the”）发生的概率大概是使用频次第二高的单词（“of”）的2倍，而单词“of”出现的概率大概是使用频次第四高的单词的2倍。如果我们绘制出单词出现频率与单词使用次数排名的关系曲线，将得到**幂律分布定理**，又被称为**齐夫定理**（Zipf’s law）。财富的分配往往也服从这样的倾斜的分布，尤其是在美国这样的富裕国家。

2.5 联合概率分布

截止目前，我们将精力都集中在了单变量概率分布。本节，我们将开始讨论一些具有挑战性的问题，针对多个相关变量的联合分布建立概率密度函数。

**联合概率分布**（joint probability distribution）的形式为，其中*D*>1为变量的数量，该分布对变量之间的关系进行建模。如果所有的变量都是离散的，该联合概率分布可以表示为一个高维的数组，其中一个变量代表一个维度。然而，该联合分布所需要确定的参数的数量为，其中*K*为每个变量可取值的数量。

通过使用条件独立性假设，我们可以使用较少的参数来定义高维的联合概率分布，关于这一点我们将在第10章进行解释。对于连续随机变量，一个可选的方式是将联合概率密度函数限定为一些特定的函数形式，我们将在下文介绍其中的相关内容。

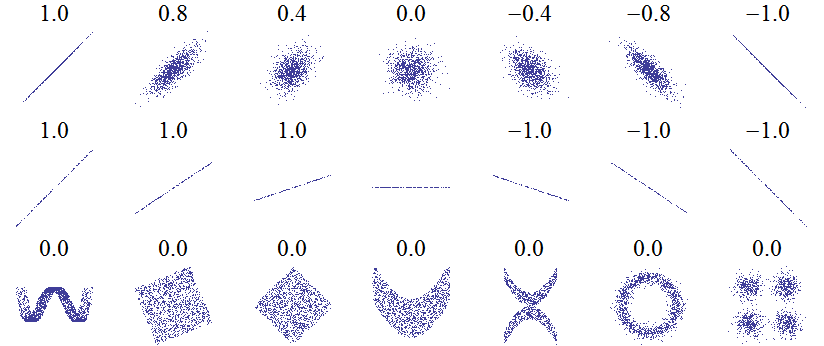


图2.12 几个(*x*,*y*)数据点集合，以及每个集合中*x*和*y*的相关系数。值得注意的是相关系数反映了线性关系的噪度和方向（最顶行），而非这个线性关系的斜率（中间行），更不能反映数据之间的非线性关系（最底行）。值得注意的是，中间的数据集的斜率为0，但在这种情况下，相关系数是没有定义的，因为变量Y的方差为0。图片源自：https://upload.wikimedia.org/wikipedia/commons/0/02/Correlation\_examples.png

2.5.1 协方差与相关系数

协方差反映了两个随机变量*X*和*Y*之间的**线性相关程度**。定义为：

 （2.59）

如果**x**是一个*d*维的随机向量，它的**协方差矩阵**（covariance matrix）定义为如下对称正定矩阵：

 （2.60）

协方差可以是0和无穷大之间的任意值。变量*X*和*Y*之间的**相关系数**（correlation coefficient）定义为：

 （2.61）

其相关系数矩阵定义为：

 （2.62）

相关系数满足-1≤corr[*X*,*Y*]≤1。所以在相关系数矩阵中所有的对角元素都为1，其他元素在-1与1之间。当且仅当，*Y*=*aX*+*b*时，corr[*X*,*Y*]=1，此时*X*与*Y*呈线性关系。直觉上，有些人会觉得相关系数应该与直线的斜率相关，比如*Y*=*aX*+*b*中的系数*a*。然而，该系数实际上等于。一个理解相关系数的更好的方式是将它理解为变量之间的线性相关程度，如图2.12所示。

如果*X*和*Y*是独立变量，即，此时我们有，进而有corr[*X*,*Y*]=0，即它们是不相关的。然而，上述命题的逆命题却不成立，即不相关（相关系数为0）的变量不一定独立。比如说，令，*Y*=*X*2。显然*Y*与*X*是相关的，然而根据计算我们有corr[*X*,*Y*]=0。（**读者可以试着计算一下两个变量之间的相关系数**）图2.12展示了相关的案例。存在一些数据集，*X*与*Y*存在很明显的相关性，然而其相关系数却等于0。一个更好地衡量随机变量之间的相关性的测度为**互信息**，我们将在2.8.3节介绍。

**附：证明-1≤corr[X,Y]≤1。**

考虑关于随机变量*X*,*Y*的两个函数*g*(*X*),*h*(*Y*)，构造含变量*t*的函数*u*(*t*)。

（2.63）

考虑到恒成立，所以恒成立。故上式关于*t*的二次函数的判别式，即：，所以下式恒成立：

 （2.64）

令，我们有：

 （2.65）

所以，我们有：

 （2.66）

2.5.2 多变量高斯分布

**多变量高斯**（multivariate Gaussian）分布或者**多变量正态**（multivariate normal，MVN）分布是连续变量中使用最广泛的联合概率密度函数。我们将在第4章详细讨论MVNs，这里我们仅给出一些定义和图形上的展示。

对于*D*维随机向量而言，其MVN定义为：

 （2.67）

其中为期望向量，为*D*×*D*的协方差矩阵。有些时候我们经常使用**精度矩阵**（precision matrix），定义为协方差矩阵的逆矩阵。归一化常数确保概率密度函数的积分为1。

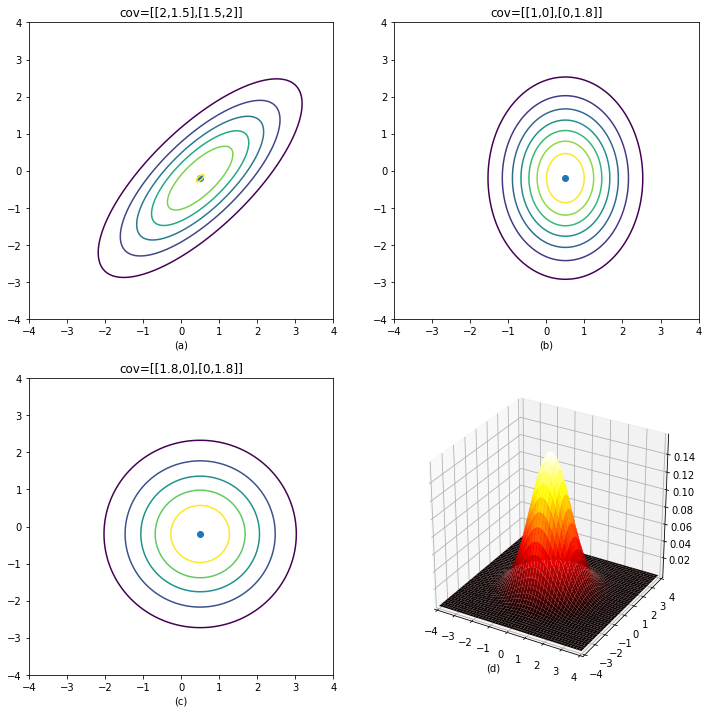


图2.13 多变量高斯分布的2维等高图。（a）一个完全协方差矩阵的等高线是倾斜的（即存在线性关系）；（b）对角协方差矩阵对应的等高线与x,y轴平行；（c）各向同性的协方差矩阵对应的等高线是一个圆；（d）高斯分布的曲面图。图形由程序**gaussPlot2Demo**生成。

图2.13绘制了一些在二维空间中的MVN，分别对应不同的协方差矩阵。一个完全协方差矩阵具有*D*(*D*+1)/2个参数（除以2是因为对称）（**读者可以考虑一下为什么参数的数量是这个**）。一个对角协方差矩阵拥有*D*个参数。一个**各向同性**（isotropic）协方差矩阵只有一个自由参数（其中为*D*维单位向量）。

2.5.3 多变量学生*t*分布

正如单变量高斯分布中的情况，其对应的**多变量学生t分布**（multivariate Student t）提高了函数的鲁棒性，定义为：

（2.68）

其中被称为尺度矩阵（scale matrix）（因为它不能完全算是协方差矩阵），。该密度函数的尾部较高斯分布更肥。值越小，尾部越肥。当，其分布趋向于高斯分布（这与单变量的情况一致）。该分布具有如下性质：

 （2.69）

2.5.4 狄利克雷分布

Beta分布在多变量情况下的泛化版本为**狄利克雷**（Dirichlet）分布，其定义域为**概率单纯形**（probability simplex）：

 （2.70）

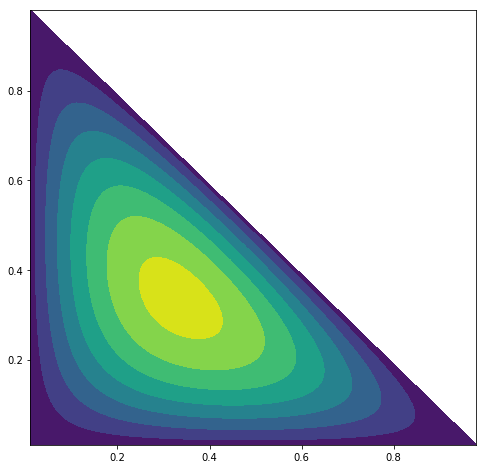
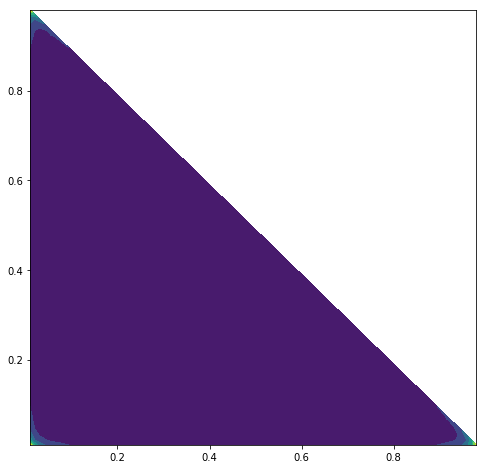
概率密度函数定义为：

 （2.71）

其中为beta函数在*K*个变量情况下的泛化：

 （2.72）

其中。

（a） （b） （c）

图2.14 三种不同值下对应的狄利克雷分布，图形由程序**dirichlet3dPlot**生成。图(a)：；图(b)：；图(c)：。注：图c中三个角存在较大的概率分布。

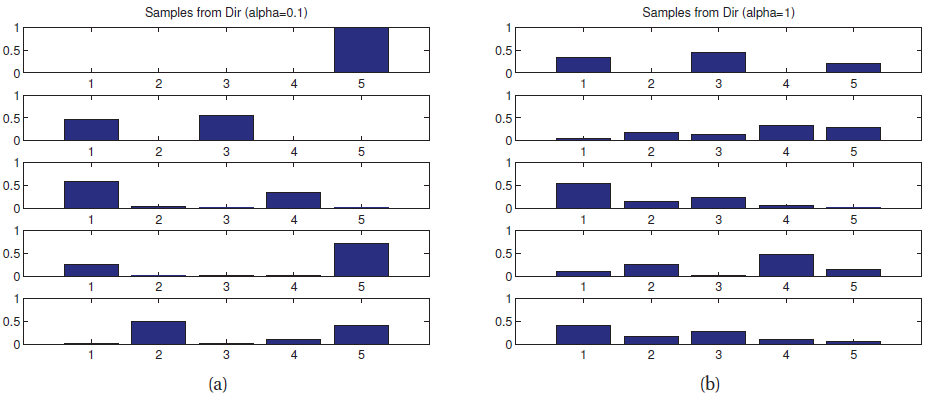


图2.15 从5维狄利克雷分布中采样得到的概率单纯形。（a）。在这种情况下采样得到的概率单纯形较为系数，很多结果出现的概率为0；（b），在这种情况下，采样的结果显得更加均匀。

图2.14显示了当*K*=3时的狄利克雷分布。图2.15显示了一些采样概率单纯形。不难发现控制着整个分布的强弱（有多尖），控制着峰值发生的地方。比如Dir(1,1,1)对应均匀分布，Dir(2,2,2)为以（1/3，1/3，1/3）为中心的覆盖更广的分布，Dir(2,2,2)为以（1/3，1/3，1/3）为中心的更加紧凑。如果对于所有的*k*, ，在概率序列的角落处出现峰值。

该分布具有如下性质：

 （2.73）

我们经常使用对称狄利克雷先验，即。在这种情况下，期望变成1/*K*，方差变成，所以增加可以提高分布的精度（即减小方差）。

2.6 随机变量的变换

如果是某个随机变量，**y**=*f*(**x**)。那么变量**y**的分布应该是什么呢？这是本节我们要讨论的问题。

2.6.1 线性变化

假设函数*f* ()为线性函数：

 （2.74）

在这种情况下，我们可以很容易地得到关于**y**的期望和方差。首先，期望为：

 （2.75）

其中。如果*f* ()是一个标量函数，，相应的期望为：

 （2.76）

对于方差，我们有：

 （2.77）

其中。如果*f*()是一个标量值，结果为：

 （2.78）

我们将在后面的章节中广泛地使用这些结果。不过值得注意的是，只有当**x**服从高斯分布时，变量**y**的分布才能只由期望和方差两个参数唯一确定。一般情况下，我们必须使用下文描述的方式去导出关于变量**y**的完整分布，而不是仅仅求解期望和方差就可以了。

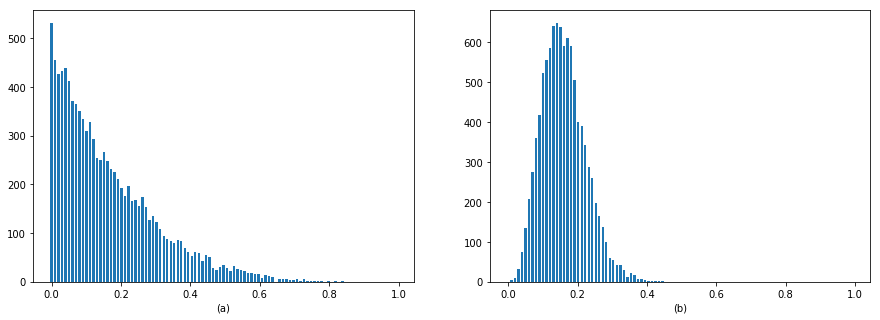


图2.17 中心极限定理的图形表示。我们绘制了的统计直方图，其中。当，该分布趋向于高斯分布。（a）*N*=1；（b）*N*=5。

2.6.2 一般变换

如果*X*是离散变量，我们可以通过简单的求和方式求解关于*y*的概率质量函数，具体的计算方式为：

 （2.79）

比如当*X*为偶数，函数*f* (*X*)=1，否则为0。在集合{1,…,10}上均匀分布，所以。注意在这个例子中，函数*f*是一个多对1的函数。

如果*X*是连续变量，我们不能使用（2.79）进行求解，因为此时的是概率密度函数，而非概率质量函数，前者不具备可加性。因此，我们使用累积密度函数进行推导：

 （2.80）

对于那些单调的可逆函数而言，我们有：

 （2.81）

对上式进行求导，我们有：

 （2.82）

式中。

关于上式我们可以考虑如下的理解：落在区间的观测值数量等价于落在的区间的数量，即。所以。

2.6.3 中心极限定理

现在考虑*N*个随机变量，其概率密度函数（不一定是高斯分布）为*p*(*x*i)，每个随机变量的期望和方差分别为和。我们假设每个变量是**独立**（independent）和**同分布**的（identically distributed）（简写为iid）。令为随机变量的和。这是一个关于随机变量简单但应用十分广泛的变换。随着*N*的增大，该和的分布服从高斯分布：

 （2.83）

所以下面定义的随机变量逼近标准正态分布：

 （2.84）

式中为样本均值。上式被称为中心极限定理（central limit theorem）。

图2.17给出了中心极限定理的例子，图中我们计算了从beta分布中随机采样的样本的均值。不难发现，样本的均值迅速收敛到一个高斯分布。

2.7 蒙特卡洛近似

在2.6.2节，我们讨论了一种计算关于随机变量函数分布的方法，当这个函数是单调的时候，可以通过简单的公式变换求得变换后变量的概率分布。但在一般情况下，这个函数不一定是单调的。此时，一种简单但十分有效的方式如下：首先我们从分布中采样得到*S*个样本，称为*x*1,…,*x*S。（有很多方式可以用来产生这些样本，对于高维数据，一个著名的方式被称为**马尔科夫蒙特卡洛法**，我们将在24章展开介绍。）给定这些样本，我们可以通过使用关于的经验分布对的分布进行估计。这被称为**蒙特卡洛**（Monte Carlo）近似，这是一个以豪华的赌场而闻名于世的欧洲城市。蒙特卡洛技术在统计物理学中被首次使用——尤其是在原子弹的发展过程中——但是如今在统计和机器学习领域被广泛应用。

使用蒙特卡洛方法可以近似地求解关于随机变量任意函数的期望值。我们简单的采集样本，然后基于样本求解函数的算术平均值。这个过程如下：

 （2.85）

式中。上式被称为蒙特卡洛积分。与数值积分（基于目标函数在固定的网格点上的值进行计算）相比，该积分只在那些概率为非负的点进行计算。

通过改变函数*f* ()的具体形式，我们可以估计很多感兴趣的值，比如：

* 
* 
* 
* 

我们将在下文给出一些例子，在后面的章节中我们也将看到更多的内容。

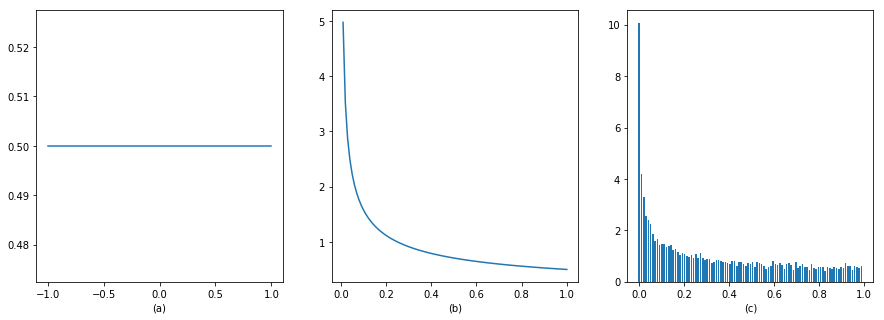


图2.18 计算的概率分布，其中*p*(*x*)服从均匀分布（左图）。关于y的分布的解析解如图（b）所示，采用蒙特卡洛法得到的关于y的概率分布如图（c）所示。图形由程序**changeOfVarsDemo1d**生成。

2.7.1 例子：变量变换法，蒙特卡洛法

在2.6.2节，我们讨论了如何通过解析的方式计算一个随机变量函数的分布。一种更加简单的方式是使用蒙特卡洛近似法。比如说，假设，且。通过从上采样很多的样本点，计算这些样本点的平方值以及最终的经验分布，基于这个经验分布我们可以对进行估计。图2.18给出了一个说明。我们将在后面的章节中广泛使用这种方法。

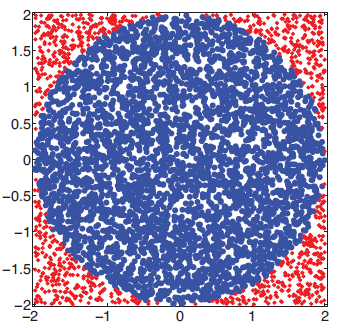


图2.19 通过蒙特卡洛积分法对进行估计。圆内的点为蓝色，外部为红色的点。

2.7.2 例子：通过蒙特卡洛方法对进行近似

蒙特卡洛近似方法在很多地方都有应用，且不局限在统计学一个领域。假设我们想估计值。我们已经知道一个半径为*r*的圆的面积为，但同时也等于如下的定积分：

 （2.86）

因此。下面让我们使用蒙特卡洛近似方法估计积分值。令为指示函数，和为定义在区间[-*r*,*r*]上的均匀分布，所以我们有。同时有：

 （2.87）

我们发现，其标准差为0.09（2.7.3节对标准差进行讨论）。我们在图2.19中绘制了那些被接受（圆内）和被拒绝（圆外）的点。

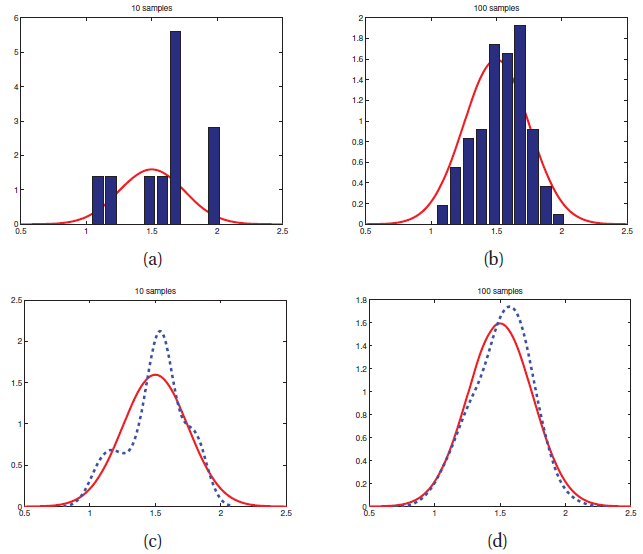


图2.20 从高斯分布中采样得到10个和100个样本。红色实线为真实的概率密度。图（a）和图（b）表示样本的分布直方图。图（c）和图（d）中蓝色点图表示根据样本点导出的核密度估计，红色实线为真实的概率密度函数。

2.7.3 蒙特卡洛近似的精度

蒙特卡洛近似法的精度随着样本数量的增加而提高。图2.20说明了这个问题，在图形的上方，我们绘制了从高斯分布中采样得到的样本的直方图。在图形下方，我们绘制了关于这些样本分布的光滑版本，这个版本使用核密度估计方法得到（章节14.7.2介绍）。接着在稠密的网格点上对这个光滑的分布进行评估。值得注意的是这个光滑的版本仅仅是为了绘图，它不是用于蒙特卡洛估计本身。

如果我们令真实的期望，通过蒙特卡洛近似得到的估计值为，在独立样本的情况下，我们有：

 （2.88）

其中：

 （2.89）

这是中心极限定理的结果。当然上式中我们并不知道，但它也可以通过蒙特卡洛方法进行估计：

 （2.90）

然后我们有：

 （2.91）

上式中被称为（数值或经验）**标准差**（standard error），它是我们对估计值的不确定度。（我们将在6.2节对标准差展开更多的讨论）

如果我们希望得到一个估计值，该估计值在真实值的范围内的概率值至少为95%，我们需要产生的样本数量*S*满足。我们可以将1.96近似为2，此时。

2.8 信息论

**信息论**（Information theory）所关心的是数据如何以一个紧凑的形式进行表达（也就是我们说的**数据压缩**data compression或者**信源编码**source coding）,以及将数据进行传输和存储，且传输和存储的方式对噪音应该具备鲁棒性（又被称为**误差纠正** error correction或者**信道编码**channel coding）。首先，这些看起来与概率论和机器学习所关心的内容相去甚远，然而事实上，它们之间有一个很紧密的关联。为了说明这一点，需要注意的是紧凑地表示数据需要将短码字分配给高度可能的位串，并将较长的码字保留给不太可能的位串。这与自然语言中情况十分相似，常见的单词（比如“a”，“the”，“and”）一般比罕见的单词短得多。此外，解码通过含噪信道发送的消息需要具有人们倾向于发送的各种消息的良好概率模型。**在各种情况下，我们需要一个模型可以预测哪种数据是有可能的，哪种是不可能的。而这个问题也是机器学习的核心问题。**（机器学习与信息论之间的更多关联可以参考其他书籍）。

显然我们不能在这里深度讨论信息论的细节，然而我们可以介绍一些在本书后面可能会用到的一些基本概念。

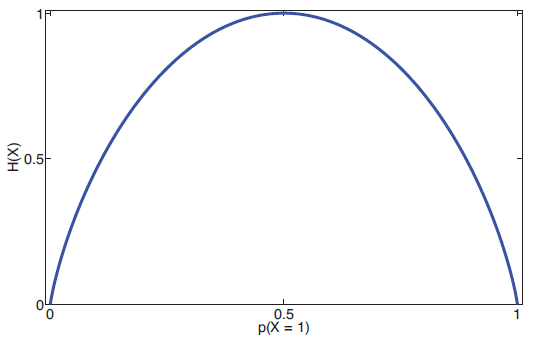


图2.21 服从伯努利分布的随机变量的熵与概率值的关系。最大熵为1。

2.8.1 熵

一个服从分布*p*的随机变量*X*的熵，用(*X*)或者(*p*)表示，它是不确定度的一个度量。特别的，对于一个含有*K*种状态的离散变量，它的熵定义为：

 （2.92）

我们通常使用2作为对数基，此时编码的信号单位为**比特**（**bits**），如果使用*e*作为对数基，那么相应的单位为**nats**。比如说，如果，其直方图分布为，得到=2.2855。拥有最大熵的离散变量的分布为均匀分布（9.2.6节给出证明）。因此，对于一个含有*K*个状态的变量而言，当时，熵最大为。相反，任意的脉冲函数具备的熵最小，因为它将所有的概率质量集中在了一个状态，没有任何不确定度。在图2.5（b）中我们绘制了DNA序列图，每一个条形块的高度为2-，其中为分布的熵，2为可能的最大熵。所以高度为0对应均匀分布，高度为2对应一个确定性分布。

当随机变量为2值变量时，即，我们可以令，。因此此时的熵为：

 （2.93）

上式被称为**二元熵函数**（binary entropy function），通常也被写成。图2.21绘制了该函数的示意图。不难发现，当时，熵的最大值为1，对应均匀分布。

2.8.2 KL散度

一种衡量两个分布之间的差异性的方式叫做Kullback-Leibler divergence（KL 散度）或者叫**相对熵**（relative entropy）。定义为：

 （2.94）

其中对于概率密度函数（连续变量），上式中的求和将改成对概率密度函数的积分。上式还可以写成：

 （2.95）

其中被称为**交叉熵**（cross entropy）：

 （2.96）

交叉熵是当我们使用模型*q*来编码服从分布*p*的源数据时所需要的平均比特数。因此章节2.8.1定义的常规熵为我们使用真实的模型进行编码时所需要的平均比特数，所以KL散度定义为这两者之间的差别。换句话说，由于我们使用新的模型*q*对真实的模型*p*进行编码，在这个过程中所需要的**额外**比特数的平均值就是KL散度。

上文中所提到的**额外**的比特数表明，且当且仅当*q*=*p*时等号成立。下面我们来证明这个重要的结论：

**定理2.8.1**. **信息不等式**（Information inequality）且当且仅当*p*=*q*时等号成立。

证明：为了证明上式，我们需要引入**琴森不等式**（Jensen’s inequality）。该不等式指出，对于任何凸函数*f*，我们有：

 （2.97）

式中并且。关于上式当n=2时显然成立（根据凸函数的定义），通过归纳法可以证明出*n*>2同样成立。

现在让我们证明主要的结论。令为分布*p*(*x*)的定义域。我们有：

 （2.98）

 （2.99）

 （2.100）

值得注意的是log为凹函数，所以琴森不等式的形式需要进行相应的变换。式2.99中等号成立的充要条件是对于某常数*c*而言。式2.100成立的充要条件为，进而我们有*c*=1。所以的充要条件是对于每一个*x*，我们有。

上述定理的一个重要结果是对于离散变量而言，最大熵分布为均匀分布。更加准确的说，，其中为变量*X*为状态数量，等式成立的充要条件为*p*(*x*)为均匀分布。为了说明这一点，令，我们有：

 （2.101）

上式被称为拉普拉斯**不充分理由原则**（principle of insufficient reason），该原则指出当我们没有任何其他理由支持某一个分布时，我们选择均匀分布。章节9.2.6将讨论如何在满足一些特定约束的情况下，利用最大熵原则，去构造一个分布。（比如说，高斯分布在满足一阶矩和二阶矩约束的前提下具备最大熵）。

2.8.3 互信息

考虑两个随机变量*X*和*Y*。假设我们希望知道一个变量在已知的情况下能够提供多少关于另一个变量的信息。我们可以计算相关系数，但这只是针对实数域的随机变量，更进一步地说，该参数对相关性的衡量十分有限，正如图2.12所示。一个更加通用的方式是衡量联合概率分布*p*(*X*,*Y*)与因子分解式*p*(*X*)*p*(*Y*)的相似度。这个量被称为**互信息**（mutual information，MI），定义如下：

 （2.102）

显然我们有，且等号成立的条件的充要条件为。也就是说只有两个变量互相独立的时候，它们之间的互信息才为0。

为了深入探讨互信息的意义，对上式进行重新书写。上式等价于：

 （2.103）

其中被称为**条件熵**（conditional entropy），定义为。因此互信息可以解释为当我们观察到变量*Y*时，关于变量*X*的不确定度的减少量，或者说当我们观察到变量*X*后对变量*Y*的不确定度的减少。我们会在本书后面的内容中遇到几个互信息的应用。

与互信息紧密相关的一个量为**点互信息**（pointwise mutual information，PMI）。对于两个事件（并非随机变量）*x*和*y*，PMI定义为：

 （2.104）

这衡量了事件同时发生与偶然发生之间的差异。显然，变量*X*和*Y*的MI是PMI的期望值。有趣的是，PMI可以重写成：

 （2.105）

上式表达了我们将先验分布*p*(*x*)（*p*(*y*)）更新为后验分布*p*(*x*|*y*)（*p*(*y*|*x*））所学习到的知识量。（**译者注：**比如说事件*x*原来发生的概率*p*(*x*)很低，但当事件*y*发生时，其概率值*p*(*x*|*y*)变得很大，那么也就是说事件*y*的发生与事件*x*的发生存在很大的关联性，也就是说通过事件*y*我们得到关于事件*x*的信息。而PMI对这个信息量进行了量化。）

2.8.3.1 连续随机变量的互信息

上节关于MI的定义局限在离散变量。对于连续随机变量，通常情况下是将它进行离散化，将连续随机变量的区间划分为块（bins），然后计算在每一个直方块中数据的数量，采用离散变量中计算互信息的方式进行计算。

不幸的是，所使用的块的数量，以及每一个块边界的所在位置对结果都有十分重要的影响。一种解决方案是对MI直接进行估计，而不需要对密度进行估计。另一个方法是尝试不同的块的大小和边界位置，然后计算在这之间最大的MI。这个被适当归一化的统计量被称为**最大信息系数**（maximal information coefficient，MIC）。更加精确的表达，定义：

 （2.106）

其中为大小为*x*×*y*的2维网格集合，*X*(*G*)，*Y*(*G*)代表变量基于该网格所进行的变量离散化。定义MIC为：

 （2.107）

其中*B*的大小取决于我们可以使用的块的数目，并且可以对分布进行可靠的估计。不难发现MIC的区间为[0,1]，其中0代表变量之间没有关系，1代表在没有噪音的情况下，存在任何形式的关系，且不局限与线性关系。

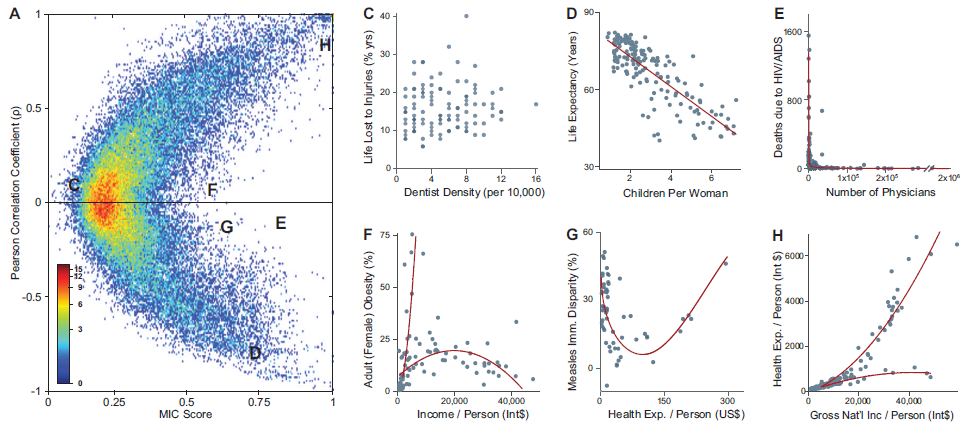


图2.22 左：相关系数与最大相关系数（MIC）之间的关系。右：特定的变量对所描述的散点图。红色实线为非参数光滑回归曲线（章节15.4.6）。

图2.22给出了在实际应用中的该统计量的图形。该数据由357个变量组成，这些变量衡量了各种社会、经济、健康和政治指标，数据由世界卫生组织（WHO）收集。在图形左侧，我们看到了所有63546对随机变量相关系数（CC）和最大信息系数（MIC）之间的关系散点图。在图形的右侧，我们绘制了特定的变量对，下面我们针对该图进行讨论：

* 标记为C的点CC值和MIC值都很低。说明这两个变量之间没有关系（即因伤而死亡的人数占比与人口中牙医的人口密度）
* 标记为D和H的点CC值和MIC值都很高，因为这些变量之间有近似的线性关系
* 标记为E，F和G的点CC值较低，但MIC值较高。说明变量之间不存在线性关系（有些情况下，比如E和F，其关系不一定是函数形式的，可能在变量之间存在一对多的关系。）

总之，我们发现基于互信息的统计量（比如MIC）可以用于发现变量之间有趣的关系，而这些关系使用更加简单的方法（比如相关系数）是不能发现的。基于这个原因，MIC已经被称为“21世纪的相关系数”。