## 2.2概率论基本知识

概率论是用于表示不确定性陈述的数学框架。它不仅提供了量化不确定性的方法，也提供了用于导出新的不确定性陈述的公理。在人工智能领域，我们主要以两种方式来使用概率论。首先，概率法则告诉我们AI系统应该如何推理，所以我们设计一些算法来计算或者近似由概率论导出的表达式。其次，我们可以用概率和统计从理论上分析我们提出的AI系统的行为。

从信息论到机器学习，从模式识别到数据挖掘，概率与统计的概念和原理活跃于计算机科学的各个领域，发挥着重要的作用。它是众多科学和工程学科的基本工具。

为什么在机器学习中会用到概率呢？这是因为，机器学习必须始终处理不确定量，有时也可能需要处理随机（非确定性）量。不确定性和随机性可能来自多个方面。研究人员至少从20世纪80年代开始就对使用概率论来量化不确定性提出了令人信服的论据。几乎所有的活动都需要能够在不确定性存在时进行推理。概率可以被看作是用于处理不确定性的逻辑扩展。

### 2.2.1概率、随机变量

概率是对随机事件发生的可能性的度量，一般以一个在0到1之间的实数表示一个事件发生的可能性的大小。越接近1，该时间更可能发生；越接近0，则该事件更不可能发生。这是一个客观论证，而不是主观验证。

随机变量，通俗来讲，它是一种会随机改变的不确定量。就其本身而言，一个随机变量只是对可能的状态的描述；它必须伴随着一个概率分布来指定每个状态的可能性。

随机变量可以是离散的或者连续的。离散型随机变量拥有有限或者可数无限多的状态。注意，这些状态不一定非要是整数，它们也可能只是一些被命名的状态并且没有数值。连续型随机变量都伴随着实数值。

### 2.2.2概率分布

概率分布也是概率论里的一个基本概念。概率分布是用来描述随机变量或一簇随机变量在每一个可能取到的状态的可能性大小。我们描述在描述概率分布时，首先要确定这个随机变量是离散的还是连续的。

比如掷骰子或者投硬币之类的就是离散型问题。离散型变量的概率分布可以用概率分布律函数来描述。我们通常用P来表示概率分布律函数。通常，每一个随机变量都会有一个不同的概率分布律函数，并且读者必须根据随机变量来推断所使用的概率分布律函数，而不是根据函数的名称来推断，例如：P(x)通常和P(y)不一样。

概率分布律函数将随机变量能够取得的每个状态映射到随机变量该状态的概率。X=x的概率用P(x)来表示，概率为1表示X=x是确定的，概率为0表示X=x是不可能发生的。我们把变量x遵循的分布表示为：X~P(x)。

一个函数P如果想要成为随机变量x的概率分布函数，那么它必须满足下面这几个条件：

* P的定义域必须是x所有可能状态的集合。
* 不可能发生事件的概率为0，并且没有比这概率更低的状态了。类似的，能够确保一定发生事件的概率为1，而且没有比这概率更高的状态了。
* .我们把这条性质称之为归一性。如果没有这条性质，那么当我们计算很多事件的其中之一发生的概率时可能会得到大于1的概率。

当我们研究的对象是连续型随机变量时，我们用概率密度函数而不是概率分布律函数来描述它的概率分布。一个函数p如果想要成为概率密度函数，必须满足下面几个条件：

* P的定义域必须是x所有可能状态的集合。
* .注意，我们并不要求P(x)。
* 。

概率密度函数p(x)并没有直接对特定的状态给出概率，相对的，它给出了落在一定面积无限小的区域内的概率。我们可以对概率密度函数求积分来获得点集的真实分布律。特别地，x落在集合S中的概率可以通过p(x)对这个集合求积分来得到。在单变量的例子中，x落在区间[a,b]的概率是。

接下来我们来认识一下累积分布函数（或者简称为分布函数），它的数学定义如下，,其中X是一个实数值随机变量。我们可以通过累积分布函数与概率密度函数的图像来表示实数值的概率分布。如果条件允许，我们还可以用数式替代图像来表示分布。它们的形式由具体的分布决定。而且，在很多实际应用中，概率密度函数比累积分布函数应用范围更广。

### 2.2.3联合概率、边缘概率和条件概率

对于现代的概率统计来说，分析多个随机变量之间的相互关系是一个关键。“在星期五购买一次性纸尿布的顾客很可能也会买啤酒”这类的观点是不是看起来很熟悉？为了讨论这类问题，我们必须分析多个随机变量之间的相互关系。我们的目标首先是要理解联合概率、边缘概率与条件概率这三个概念。它们是讨论随机变量之间关系的基本道具。最近活跃于各领域的贝叶斯公式也是这组概念的一种应用。此外，独立性的定义也基于这组概念。在日常的概率问题中，我们在使用独立性时无需特地声明，不过由于本书的要求更高，因此必须明确定义独立性的含义。在分析随机变量之间的关系时，某一变量是否独立往往是解决问题的关键。

假设有随机变量X与Y，此时P(X=a,Y=b)用于表示X=a且Y=b的概率。这类包含多个条件且所有条件同时成立的概率称为联合概率。要注意，联合概率并不是其中某个条件的成立概率，而是所有条件同时成立的概率。与之对应地，P(X=a)或P(Y=b)这类仅与单个随机变量有关的概率称为边缘概率。联合概率的一览表称为联合分布，边缘概率的一览表称为边缘分布。

联合概率与边缘概率的关系如下。





接下来，我们来引入条件概率这个概念。在实际生活中，许多有价值的变量都能以条件概率这一概念来描述。举个例子，“包含‘免费’这一单词的邮件很可能是广告”，这种A条件下事件B的概率称为条件概率。下面是条件概率的定义。



我们从面积的角度看，可以将理解为“区域a中b的比例”，代表的是“区域a中b的面积”，而P(X=a)代表的是“区域a的总面积”，那么就是“区域a中b的面积除以区域a的总面积”了。

条件概率经常是理工科一些问题的焦点。这是因为在研究理工科问题时，我们常会采用控制变量法分析变量之间的关系，讨论变量X取特定值时变量Y的取值情况。如果没有误差，我们可以用函数Y=f(X)来表示它们的关系。不过在现实中，我们很难去除所有影响因素，测到绝对精确的值。于是，即使X的测定值不变，Y的测定值也会发生细微变化。此时，我们就需要研究在X为某个特定值时Y的概率分布。也就是说，我们将研究条件概率，而不是函数Y=f(X)。

也是正因如此，要在实际问题中运用概率理论，我们必须掌握条件概率的计算方法。接下来我们来熟悉一下条件概率中常见的基本关系式。

我们可以像下面这样，通过边缘分布和条件分布来表示联合分布。



我们也可以想了解条件概率的定义一样从面积的角度来理解这个式子，这里就不一一赘述。

现在，我们将应用条件概率来解决逆问题。简单地说，逆问题是指那些需要从结果反推原因的问题。通常，原因X无法被直接观察、测量，此时，我们常会通过其结果Y来反推X。比如，在文字识别问题时，需要根据扫描仪读取的图像数据Y来推测用户书写的文字X。请注意，X相同，Y也可能会不同。由于绝大多数情况都存在噪声与误差，因此我们不能简单地使用函数Y=f(X)来模拟问题。为此，我们需要借助概率来处理这些噪声与误差，通过随机变量X,Y来描述这两者之间的相互关系。即在已知时来计算。由此，如果我们还知道P(X=a)，就可以用贝叶斯公式来实现这一目的：



贝叶斯公式可以从条件概率的定义直接推导出来，但我们最好记住这个公式以及它的名字，因为在之后的算法中会经常用到。

### 2.2.4独立性

现在，我们来讨论多个随机变量之间关系中更为根本的问题。如果一个问题当中存在多个随机变量，我们首先会怀疑这些随机变量之间是否真的存在关联。因此，独立性的概念是很多应用问题中的关键。

独立的定义有多种等价表述方式。在讨论概率问题是，独立性是一种很重要的基础概念，因此大家需要熟悉各种不同的表达方式。请大家记住，以下表述的含义全都是相同的。

* X与Y独立
* 条件概率与条件无关：
* 添加或去除条件不影响概率：
* 联合概率是边缘概率的乘积：

### 2.2.5期望、方差和协方差

函数f(x)关于某分布P(X)的期望（expectation）或者期望值是指，当x由P产生时，f作用于x的平均值。对于离散型随机变量，可以通过求和得到：



对于连续型随机变量可以通过求积分得到：



方差（variance）衡量的是当我们对x依据它的概率分布进行采样时，随机变量x的函数值会呈现多大的差异：



当方差很小时，f(x)的值形成的簇比较接近它们的期望值。方差的平方根被称为标准差。

协方差在某种意义上给出了两个变量线性相关性的强度以及这些变量的尺度：



协方差的绝对值如果很大则意味着变量值变化很大并且他们同时距离各自的均值很远。如果协方差是正的，那么两个变量都倾向于同时取得相对较大的值。如果协方差是负的，那么其中一个变量倾向于取得相对较大的值的同时，另一个变量倾向于取得相对较小的值，反之亦然。其他的衡量指标如相关系数将每个变量的贡献归一化，为了只衡量变量的相关性，而不受变量大小的分别影响。

协方差和相关性是有联系的，但实际上是不同的概念。它们是有联系的，因为两个变量如果相互独立，那么它们的协方差为零，如果两个变量的协方差不为零，那么它们一定是相关的。然而，独立性又是和协方差完全不同的性质。两个变量如果协方差为零，它们之间一定没有线性关系。独立性的要求比协方差为零更高，因为独立性在此之上还排除了非线性的关系。两个变量可能相互依赖，但同时它们协方差为零。

### 2.2.6常用概率分布

伯努利分布，又名两点分布或0-1分布。这是一种最简单的离散型分布。假设随机变量X为投一枚硬币的结果，类似于“投硬币”的试验就是伯努利试验。如果试验E是一个伯努利试验，将E独立重复地进行n次，则称这一串重复的独立试验为n重伯努利分布。进行一次伯努利试验，成功(X=1)概率为p()，失败(X=0)概率为1-p，则称随机变量X服从伯努利分布。它的概率分布函数为：



1. 分布是n重伯努利试验成功次数的离散概率分布。如果试验E是一个n重伯努利试验，每次伯努利试验成功概率为p,X代表成功的次数，则X的概率分布是二项分布，记为，其概率质量函数为：



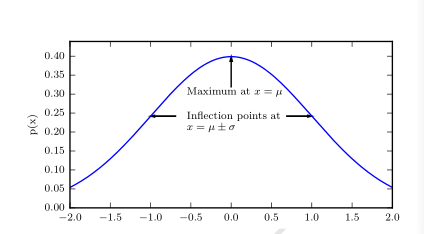


可以发现，伯努利分布其实就是二项分布在n=1时的特例。

现在，我们要来介绍一下正态分布，正态分布是对于实数上的分布最常用的分布，也称为高斯分布。



正态分布被两个参数控制，。参数给出了中心峰值的坐标，这也是分布的均值：E(X)=。分布的标准差用表示，方差用表示。下面是标准正态分布的概率密度函数，其中。



采用正态分布在很多应用中都是一个明智的选择。当我们缺乏对于某个实数分布的先验认识而不知道该选择怎样的形式时，正态分布是默认的比较好的选择，这里有两个原因。

第一，我们想要建模的很多分布真实情况都是比较接近正态分布的。中心极限定理告诉我们，很多独立随机变量的和近似服从正态分布。这意味着，在实际应用中，很多复杂系统都可以被成功的建模成正态分布的噪声，即使系统可以被分解成具有更多结构化行为的各个部分。

第二，在具有相同方差的所有可能的概率分布中，正态分布在实数上具有最大的不确定性。我们因此可以认为正态分布是对模型加入的先验知识量最少的分布。

参考资料：

《深度学习》

《程序员的数学2——概率统计》

数理统计中常用函数、概率分布函数总结 - CSDN博客 http://blog.csdn.net/DreamHome\_S/article/details/78275268