概率概述

随机变量

随机变量x表示一个不确定的数量。该变量可以表示一个实验的结果或波动特性的真实量度。如果我们观察几个实例$\{x_i\}_{i=1}^I$，它可能在每个场合取不同的值。然而，一些值可能比其他值更容易出现。这种信息时由随机变量的概率分布$Pr(x)$决定的。

随机变量可以是离散的或连续的。离散变量从一组事先确定的集合中取值。这组值可能是有序的或者无序的。可能是有限的或者无限的。离散变量的概率分布可以可视化为一个直方图或Hinton图。每个结果都有与之相关的一个正概率，且所有结果的概率之和是1。

连续随机变量取实数值。这些取值可能是有限的或无限的。无限连续变量可能取遍整个实数范围，或者可能是仅有上界或下界的区间。连续变量的概率分布可以通过绘制概率密度函数(PDF)来可视化。一个结果的概率密度表示随机变量取该值的相对可能性。它可以取任何正值，但是PDF的积分总是1。

联合概率

假设两个随机变量x和y。若观察x和y的多个成对实例，结果中某些组合出现得较为频繁。这样的情况用x和y得联合概率分布表示，记作$Pr(x,y)$。在$Pr(x,y)$中的逗号可以理解为“和”，所以$Pr(x,y)$是x和y的概率。一个联合概率分布中的相关变量可能全是离散变量。或全是连续变量，抑或是兼而有之。不管怎样，所有结果的概率之和（离散变量的总和与连续变量的积分）总是1。

一般来说，与二元变量的概率分布相比，我们会对多元变量的联合概率分布更感兴趣。我们讲$Pr(x,y,z)$记为标量变量x,y,z的联合概率分布，也可以把$Pr(\mathbf{x})$当成所有多维元素$\mathbf{x}=[x_1,x_2,\ldots,x_k]^T$的联合概率。最后，我们用$Pr(\mathbf{x},\mathbf{y})$表示所有多维变量$\mathbf{x},\mathbf{y}$的联合概率分布。

边缘化

任意单变量的概率分布都可以通过在联合概率分布上求其他变量的和（离散）或积分（连续）而得到。例如，如果x和y是连续的，并且已知$Pr(x,y)$，那么通过如下计算就可以得到概率分布$Pr(x)$和$Pr(y)$：

$$Pr(x)=\int Pr(x,y)dy$$

$$Pr(y)=\int Pr(x,y)dx$$

所求出的分布$Pr(x)$和$Pr(y)$称为边缘分布，其他变量的积分或求和过程称为边缘化。联合分布$Pr(x,y)$中忽略变量y的影响，计算边缘分布$Pr(x)$的过程也可以简单地解释为：计算x地概率分布且忽略（或不考虑）y的值。

一般来说，可以通过边缘化所有其他的变量求出任何变量子集的联合概率。例如，给定变量w,x,y,z，其中w是离散的，z是连续的，可以用下面的式子求出$Pr(x,y)$：

$$Pr(x,y)=\sum_w\int Pr(w,x,y,z)dz$$

条件概率

给定y取y^*时x的条件概率，是随机变量x在y取固定值y^*时x的相对概率的取值。这个条件概率记为$Pr(x|y=y^*)$。其中“|”可以理解为“给定”。

条件概率$Pr(x|y=y^*)$可以由联合分布$Pr(x,y)$计算出来。特别是，计算联合分布中某个恰当的切片$Pr(x,y=y^*)$。切片值表示出当$y=y^*$时x取不同值的相对概率，但其本身没有形成有效的概率分布，因为它们仅仅构成联合分布的一小部分，其总和不会是1，而联合概率分布自身总和为1.为计算条件概率分布，因此需要规范化切片中的总概率：

$$Pr(x|y=y^*)=\frac{Pr(x,y=y^*)}{\int Pr(x,y=y^*)dx}=\frac{Pr(x,y=y^*)}{Pr(y=y^*)}$$

其中，使用边缘概率关系式去简化分母。通常情况下不会显式定义$y=y^*$，所以条件概率关系式可简化缩写为：

$$Pr(x|y)=\frac{Pr(x,y)}{Pr(y)}$$

重新整理得到：

$$Pr(x,y)=Pr(x|y)\cdot Pr(y)$$
$$Pr(x,y)=Pr(y|x)\cdot Pr(x)$$

当有两个以上的变量时，可以不断用条件概率分布将联合概率分布分解为乘积形式：

$$\begin{aligned}Pr(w,x,y,z)&=\Pr(w,x,y|z)\cdot Pr(z)\\&=Pr(w,x|y,z)\cdot Pr(y|z)\cdot Pr(z)\\&=Pr(w|x,y,z)\cdot Pr(x|y,z)\cdot Pr(y|z)\cdot Pr(z)\end{aligned}$$

贝叶斯公式

结合上面的公式，我们可以得到$Pr(x|y)$和$Pr(y|x)$之间的关系：

$$Pr(x|y)\cdot Pr(y)=Pr(y|x)\cdot Pr(x)$$

重新整理后得到：

$$\begin{aligned}Pr(y|x)&=\frac{Pr(x|y)\cdot Pr(y)}{Pr(x)}\\&=\frac{Pr(x|y)\cdot Pr(y)}{\int Pr(x,y)dy}\\&=\frac{Pr(x|y)\cdot Pr(y)}{\int Pr(x|y)\cdot Pr(y)dy}\end{aligned}$$

其中第二行和第三行分别利用边缘概率和条件概率的定义对分母进行展开。这三个式子通常统称为贝叶斯公式

贝叶斯公式中每项都有一个名称。等号左边的$Pr(y|x)$叫做后验概率，代表给定x下y的概率。相反，$Pr(y)$叫做先验概率，表示在考虑x之前y的概率。$Pr(x|y)$叫做似然性，分母$Pr(x)$是证据。

在计算机视觉中，常常用条件概率$Pr(x|y)$来表示变量x和y的关系。然而，我们主要感兴趣的可能是变量y，在这种情况下，概率$Pr(y|x)$就用贝叶斯公式来计算。

独立性

如果从变量x不能获得变量y的任何信息（反之亦然），就称x,y是独立的，可以表示为：

$$Pr(x|y) = Pr(x)$$

$$Pr(y|x) = Pr(y)$$

代入前式可得，独立变量的联合概率$Pr(x,y)$是边缘概率$Pr(x)$和$Pr(y)$的乘积，即：

$$Pr(x,y) = Pr(x|y) \cdot Pr(y) = Pr(x) \cdot Pr(y)$$

期望

给定一个函数$f(\cdot)$和每个x所对应的概率$Pr(x = x^*)$，函数对变量x的每个值x^*都返回一个值，有时希望求函数的期望输出。如果从概率分布中抽取大量样本，计算每个样本的函数，并求这些值的平均值，其结果就是期望。更确切地说，在离散及连续的情况下，一个随机变量x的函数$f(\cdot)$的期望值分别定义为：

$$E[f(x)] = \sum_x f(x)Pr(x)$$

$$E[f(x)] = \int f(x)Pr(x)dx$$

将这种思路推广到二元随机变量的函数$f[\cdot]$，则有：

$$E[f(x,y)] = \iint f(x,y) \cdot Pr(x,y)dxdy$$

对于某些特殊的函数$f(\cdot)$，期望被赋予特殊的名称。这些特殊函数常用来概括复杂概率分布的性质。

函数$f(\cdot)$	期望
x	均值μ_x
x^k	关于零的第k阶矩阵
$(x-\mu_x)^k$	关于均值的第k阶矩阵
$(x-\mu_x)^2$	方差
$(x-\mu_x)^3$	偏度
$(x-\mu_x)^4$	峰度
$(x-\mu_x)(y-\mu_y)$	x和y的协方差

期望有四条性质，这些性质能够通过期望的原始定义简单证得：

1. 若随机变量x是常数k，则其期望是常数本身：

$$E[k] = k$$

2. 常数k与函数$f(x)$的乘积所得函数的期望是$f(x)$期望的k倍：

$$E[k \cdot f(x)] = k \cdot E[f(x)]$$

3. 随机变量都是x时：函数$f(x)$和$g(x)$相加所得函数的期望是两个函数期望的总和：

$$E[f(x) + g(x)] = E[f(x)] + E[g(x)]$$

4. 若随机变量x和y相互独立，则函数$f(x)$和$g(y)$相乘所得函数的期望是两个函数期望的乘积：

$$E[f(x) \cdot g(y)] = E[f(x)] \cdot E[g(y)]$$