统计学与概率论基础

机器学习（入门） DC学院

基本概念

概率定义

> 概率是一个在0到1之间的**实数**，是对随机事件发生的**可能性**的**度量**。
>
> **不可能事件**概率为0，符号为ϕ
>
> **必然事件**概率为1，符号为**Ω**

概率说明

> **概率**，通常是指一个具有**不确定性**的事件**发生**的**可能性**。

注意：
概率更偏理论上的定义，是理论值。

> **频率**是在**有限次**实验中事件**发生次数**占总实验次数的**百分比**，是**经验值**。
>
> 根据大数定理，可以认为在**无穷多次**实验中的频率值无限**接近**于概率值，此时可以用频率**代替**概率。

古典概率（事前概率）

> **满足以下条件**：
> 1)随机现象所能发生的事件是有限的、互不相容的
> 2)每个基本事件发生的可能性相等

例子：
抛硬币，掷骰子

离散概率

> **满足以下条件**：
> 随机现象所能发生的事件是有限的、互不相容的

例子：
抛硬币，掷骰子，预测明天是否下雨

连续概率

满足以下条件:
事件发生的可能性是无限个（不可数），且在一个区间内

例子:
某校同学的身高

样本空间

定义
随机事件**一切可能发生的结果组成的集合**称为此随机事件的样本空间，符号为**Ω**

随机事件发生的任何结果都**必然**存在于样本空间=>样本空间是**必然**发生

例子:
抛硬币的样本空间：{正面,反面}

事件域

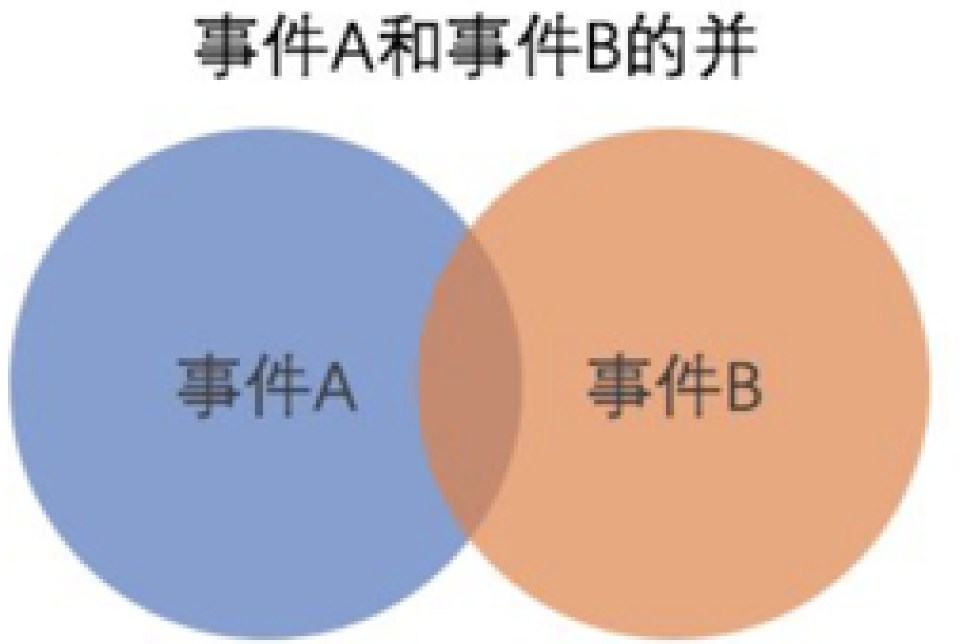
定义
样本空间中所有子集组成的集合类，符号**F**

例子:
抛硬币的事件域：{ϕ，正面,反面，Ω}

条件概率

条件概率在已知事件A发生的情况下，事件B发生的概率，记为：
P（B | A）

事件交集：事件A与事件B同时发生的概率，记为：
P（AB）



设（Ω， 𝒻，P）是一个概率空间， B∈ 𝒻，且 P(B)>0,则对任意 A∈𝒻记：
P（B | A）= P（AB）/ P(A)

例子：
先后掷两次硬币，假设1为正面，0为背面则样本空间为：
（1，1），（1，0），（0，1），（0，0）

事件域为16种情况：（ ∅，……， Ω ）

记事件A为至少有一次为正面其概率：P(A) = 3/4

记事件B为一次正面一次背面其概率：P（B）=1/2

在已抛出一次正面情况下抛出一次背面概率
P（B | A）= P（AB）/ P(A)=2/3

全概率公式

完备事件组：事件之间两两互斥，所有事件的并集是整个样本空间（必然事件）。

完备事件组满足：

$$BiBj = \varnothing (i \neq j)$$
$$B1 + B2 + \cdots = \Omega$$

全概率公式推导：
假设我们要研究事件A。我们希望能够求出P(A)，但是经过一番探索，却发现P(A)本身很难直接求出，不过却能够比较容易地求出各个P(Bi)，以及相应的条件概率P(A|Bi)。
Bi是两两互斥的。

$$A = A\Omega = AB1 + AB2 + AB3 + \cdots$$

显然，AB1，AB2，AB3，…也是两两互斥的。

一说到两两互斥，我们就想到了概率的加法定理：

$$P(A)=P(A\Omega)=P(AB1+AB2+AB3+\cdots)=P(AB1)+P(AB2)+P(AB3)+\cdots$$

再根据条件概率的定义，最后推导出全概率公式：

$$P(A)=P(B1)P(A|B1)+P(B2)P(A|B2)+P(B3)P(A|B3)+\cdots$$

随机变量

随机变量并不是变量，它实际上是将（样本空间中的）结果映射到真值的函数。

有连续随机变量和离散随机变量两种。

优点：

可以用数学分析的方法来研究随机现象。

例子

在掷骰子时，我们常常关心的是两颗骰子的点和数，而并不真正关心其实际结果，就是说，我们关心的也许是其点和数为7，而并不关心其实际结果是否是（1，6）或（2，5）或（3，4）或（4，3）或（5，2）或（6，1）。

联合分布

两个及以上随机变量组成的随机变量的概率分布叫做**联合分布**。

用P(X=a,Y=b)或PX,Y(a,b)来表示X取值为a且Y取值为b时的概率。

用P(X,Y)来表示它们的联合分布。

注意：

$$\sum_x \sum_y P(X=x,Y=y)=1$$

例子：

1. 假设在投掷一个骰子的样本空间Ω上定义一个随机变量X，如果骰子是均匀的则X的分布为：

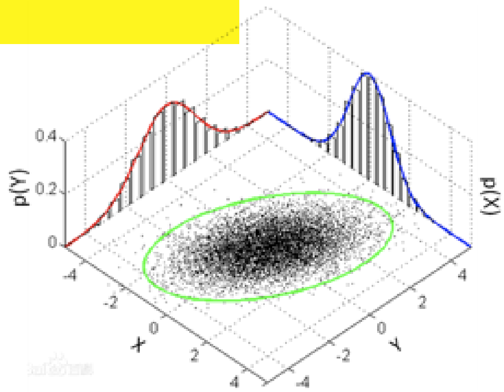
   PX(1)=PX(2)=…=PX(6)=1/6

2. 在投掷一个骰子的样本空间上定义一个随机变量X。定义一个指示变量Y，当抛硬币结果为正面朝上时取1，反面朝上时取0。假设骰子和硬币都是均匀的，则X和Y的联合分布如下：

P	X=1	X=2	X=3	X=4	X=5	X=6
Y=0	1/12	1/12	1/12	1/12	1/12	1/12
Y=1	1/12	1/12	1/12	1/12	1/12	1/12

边缘分布

边缘分布是指一个随机变量对于其自身的概率分布。



红色曲线为Y的边缘分布曲线
蓝色曲线为X的边缘分布曲线

为了得到一个随机变量的边缘分布，我们将该分布中的所有其它变量相加，准确来说，就是：

$$P(x) = \sum_y P(x,y) = \sum_y P(x|y)P(y)$$

例子：

P	**X=1**	**X=2**	**X=3**	**X=4**	**X=5**	**X=6**
Y=0	1/12	1/12	1/12	1/12	1/12	1/12
Y=1	1/12	1/12	1/12	1/12	1/12	1/12

X=1的边缘概率为1/12+1/12=1/6

条件分布

对于二维随机变量(X，Y)，可以考虑在其中一个随机变量取得(可能的)固定值的条件下，另一随机变量的概率分布，这样得到的X或Y的概率分布叫做条件概率分布，简称**条件分布**

条件分布为概率论中用于探讨不确定性的关键工具之一。它明确了在另一随机变量已知的情况下（或者更通俗来说，当已知某事件为真时）的某一随机变量的分布

公式:
当给定Y=b时，X=a的条件概率定义为：

$$P(X=a|Y=b)=\frac{P(X=a,Y=b)}{P(Y=b)}$$

注意：当Y=b的概率为0时，上式不成立

独立性:

在概率论中，独立性是指随机变量的分布不因知道其它随机变量的值而改变

数学解释:

从数学角度来说，随机变量X独立于Y:

$$\boldsymbol{P}(\boldsymbol{X}) = \boldsymbol{P}(\boldsymbol{X}|\boldsymbol{Y})$$

链式法则:

复合函数的导数将是构成复合这有限个函数在相应点的导数的乘积，就像锁链一样一环套一环，故称链式法则，如:

$$P(X1, X2, \ldots, Xn) = P(X1)P(X2|X1)\ldots P(Xn|X1, X2, \ldots, Xn-1)$$

链式法则用于贝叶斯定理:

$$\mathrm{P}(\mathrm{X}|\mathrm{Y}) = \frac{\mathrm{P}(\mathrm{Y}|\mathrm{X})\mathrm{P}(\mathrm{X})}{\mathrm{P}(\mathrm{Y})}$$

$$\mathrm{P}(\mathrm{Y}) = \sum_{\mathrm{a} \in \mathrm{Val}(\mathrm{X})} \mathrm{P}(\mathrm{X}=\mathrm{a}, \mathrm{Y}) = \sum_{\mathrm{a} \in \mathrm{Val}(\mathrm{X})} \mathrm{P}(\mathrm{Y}|\mathrm{X}=\mathrm{a})\mathrm{P}(\mathrm{X}=\mathrm{a})$$

期望:

数学期望(mean)（或均值，亦简称期望）是试验中每次可能结果的概率乘以其结果的总和，它反映随机变量**平均取值的大小**，随机变量的期望记为E(x)。

公式:

$$E(x) = \sum_{\mathrm{a} \in \mathrm{Val}(\mathrm{X})} \mathrm{a}P(X=\mathrm{a})$$

方差:

一个随机变量的方差描述的是它的**离散程度**，也就是该变量离其期望值的距离。方差的算术平方根称为该随机变量的标准差。

离散型随机变量方差计算公式：

$$Var(X) = E(X^2) - \left[(E(X))\right]^2$$