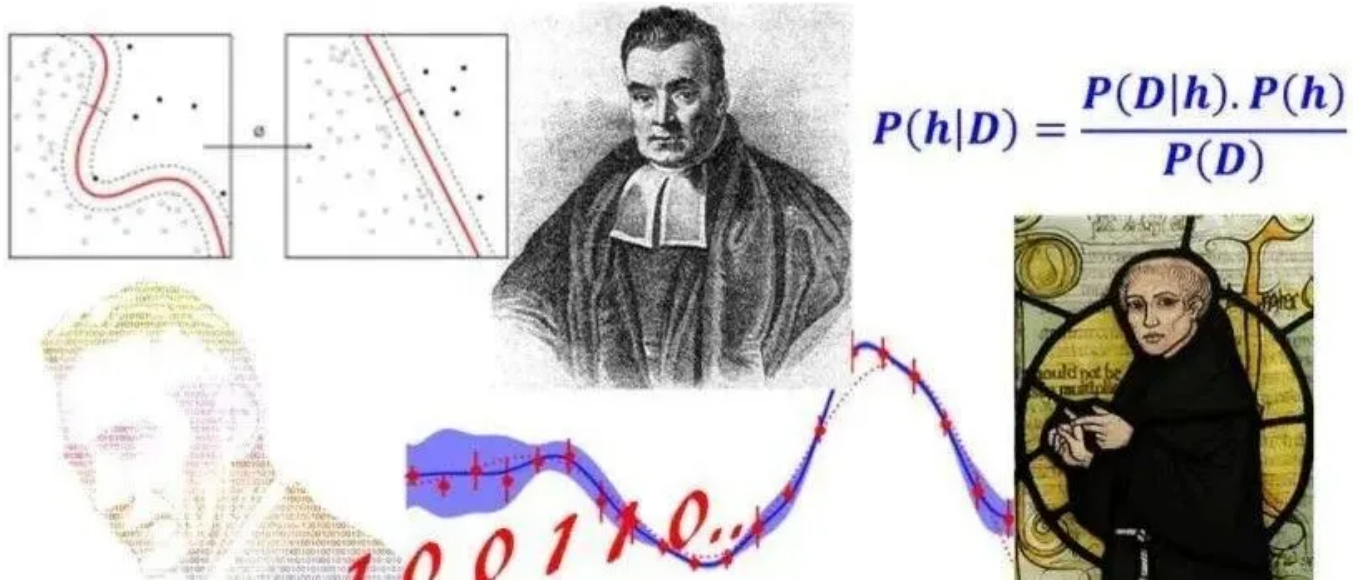
從貝葉斯定理到概率分佈：詳解概率論基本定義

機器學習算法與Python實戰 今天



轉自：機器之心

本文從最基礎的概率論到各種概率分佈全面梳理了基本的概率知識與概念，這些概念可能會幫助我們了解機器學習或開拓視野。這些概念是數據科學的核心，並經常出現在各種各樣的話題上。重溫基礎知識總是有益的，這樣我們就能發現以前並未理解的新知識。

簡介

在本系列文章中，我想探討一些統計學上的入門概念，這些概念可能會幫助我們了解機器學習或開拓視野。這些概念是數據科學的核心，並經常出現在各種各樣的話題上。重溫基礎知識總是有益的，這樣我們就能發現以前並未理解的新知識，所以我們開始吧。

第一部分將會介紹概率論基礎知識。

概率

我們已經擁有十分強大的數學工具了，為什麼我們還需要學習概率論？我們用微積分來處理變化無限小的函數，併計算它們的變化。我們使用代數來解方程，我們還有其他幾十個數學領域來幫助我們解決幾乎任何一種可以想到的難題。

難點在於我們都生活在一個混亂的世界中，多數情況下無法準確地測量事物。當我們研究真實世界的過程時，我們想了解許多影響實驗結果的隨機事件。不確定性無處不在，我們必須馴服它以滿足我們的需要。只有如此，概率論和統計學才會發揮作用。

如今，這些學科處於人工智能，粒子物理學，社會科學，生物信息學以及日常生活中的中心。

如果我們要談論統計學，最好先確定什麼是概率。其實，這個問題沒有絕對的答案。我們接下來將闡述概率論的各種觀點。

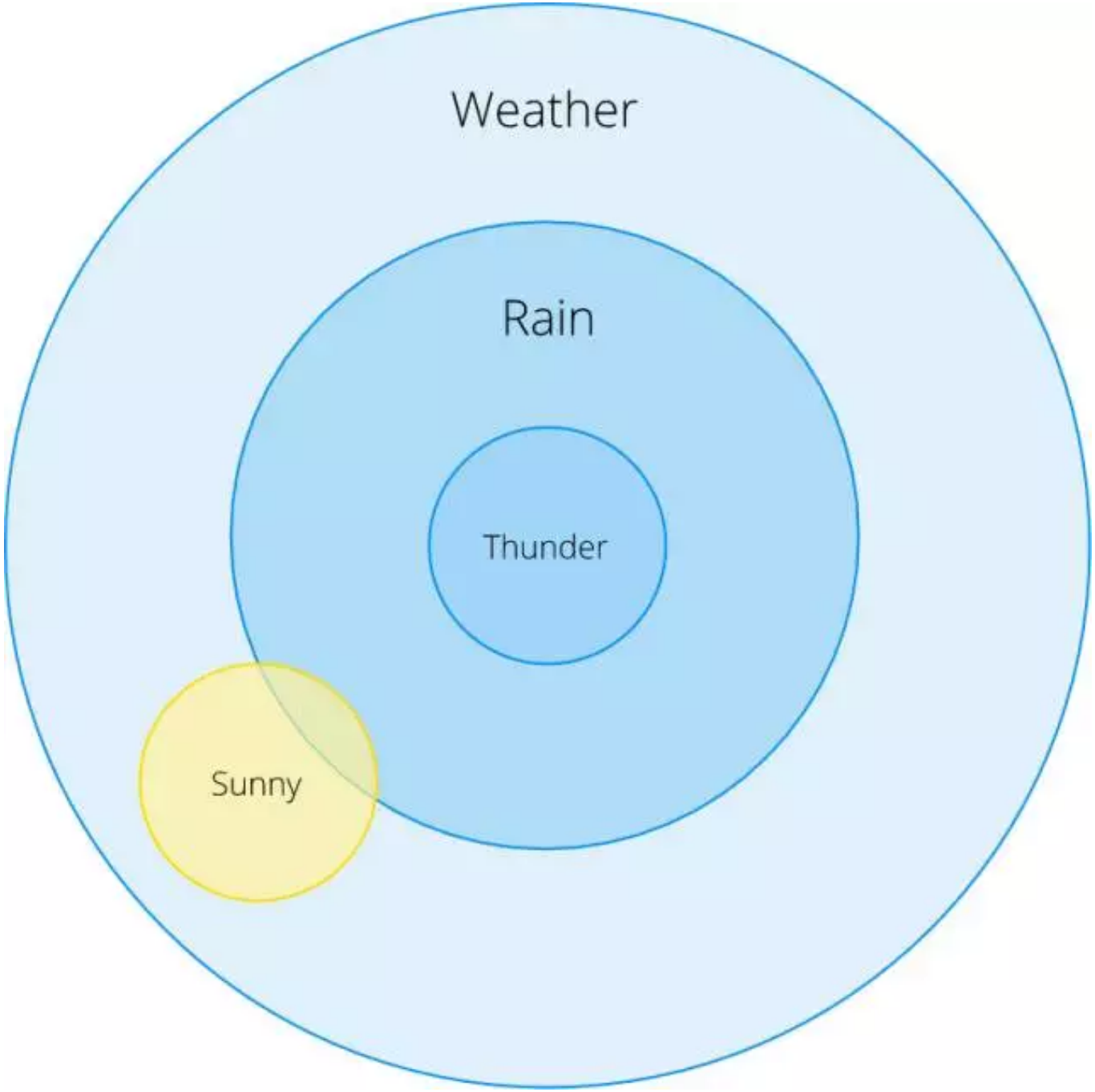
頻率

想像一下，我們有一枚硬幣，想驗証投擲後正反面朝上頻率是否相同。我們如何解決這一問題？我們試著進行一些實驗，如果硬幣正面向上記錄1，如果反面向上記錄0。重複投擲1000 次並記錄0 和1 的次數。在我們進行了一些繁瑣的時間實驗後，我們得到了這些結果：600 個正面（1）和400 反面（0）。如果我們計算過去正面和反面的頻率，我們將分別得到60%和40%。這些頻率可以被解釋為硬幣出現正面或者反面的概率。這被稱為頻率化的概率。

條件概率

通常，我們想知道某些事件發生時其它事件也發生的概率。我們將事件B 發生時事件A 也發生的條件概率寫為P（A | B）。以下雨為例：

- 打雷時下雨的概率有多大?
- 晴天時下雨的概率有多大?


Weather
Rain
Thunder
Sunny

從這個歐拉圖，我們可以看到P（Rain | Thunder）= 1：當我們看到雷聲時，總會下雨（當然，這不完全正確，但是我們在這個例子中保證它成立）。

P（Rain | Sunny）是多少呢？直覺上這個概率很小，但是我們怎樣才能在數學上做出這個準確的計算呢？條件概率定義為：

$$P(Rain|Sunny) = \frac{P(Rain, Sunny)}{P(Sunny)}$$

換句話說，我們用Rain 且Sunny 的概率除以Sunny 的概率。

相依事件與獨立事件

如果一個事件的概率不以任何方式影響另一個事件，則該事件被稱為獨立事件。以擲骰子且連續兩次擲得2 的概率為例。這些事件是獨立的。我們可以這樣表述

$$P(roll2) = P(roll2_{\text{1st time}})P(roll2_{\text{2nd time}})$$

但是為什麼這個公式可行？首先，我們將第一次投擲和第二次投擲的事件重命名為A 和B，以消除語義影響，然後將我們看到的兩次投擲的的聯合概率明確地重寫為兩次投擲的單獨概率乘積：

$$P(A, B) = P(A)P(B)$$

現在用P（A）乘以P（B）（沒有變化，可以取消）並重新回顧條件概率的定義：

$$P(A) = \frac{P(A)P(B)}{P(B)} = \frac{P(A,B)}{P(B)} = P(A \mid B)$$

如果我們從右到左閱讀上式，我們會發現P(A | B) = P(A)。這就意味著事件A 獨立於事件B！P（B）也是一樣，獨立事件的解釋就是這樣。

貝葉斯概率論

贝叶斯可以作为一种理解概率的替代方法。频率统计方法假设存在我们正在寻找的模型参数的一个最佳的具体组合。另一方面，贝叶斯以概率方式处理参数，并将其视为随机变量。在贝叶斯统计中，每个参数都有自己的概率分布，它告诉我们给已有数据的参数有多种可能。数学上可以写成

$$P(\Theta \mid D)$$

这一切都从一个允许我们基于先验知识来计算条件概率的简单的定理开始：

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

尽管贝叶斯定理很简单，但它具有巨大的价值，广泛的应用领域，甚至是贝叶斯统计学的特殊分支。有一个关于贝叶斯定理的非常棒的博客文章，如果你对贝叶斯的推导感兴趣---这并不难。

抽样与统计

假设我们正在研究人类的身高分布，并渴望发表一篇令人兴奋的科学论文。我们测量了街上一些陌生人的身高，因此我们的测量数据是独立的。我们从真实人群中随机选择数据子集的过程称为抽样。统计是用来总结采样值数据规律的函数。你可能见过的统计量是样本均值：

$$\bar{x} = \frac{1}{n}\sum_{i=0}^{n} x_i$$

另一个例子是样本方差：

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

这个公式可以得出所有数据点偏离平均值的程度。

分布

什么是概率分布？这是一个定律，它以数学函数的形式告诉我们在一些实验中不同可能结果的概率。对于每个函数，分布可能有一些参数来调整其行为。

当我们计算硬币投掷事件的相对频率时，我们实际上计算了一个所谓经验概率分布。事实证明，世界上许多不确定的过程可以用概率分布来表述。例如，我们的硬币结果是一个伯努利分布，如果我们想计算一个 n 次试验后硬币正面向上的概率，我们可以使用二项式分布。

引入一个类似于概率环境中的变量的概念会方便很多--随机变量。每个随机变量都具有一定的分布。随机变量默认用大写字母表示，我们可以使用 ~ 符号指定一个分布赋给一个变量。

$$X \sim Bernoulli(0.6)$$

上式表示随机变量 X 服从成功率（正面向上）为 0.6 的伯努利分布。

连续和离散概率分布

概率分布可分为两种：离散分布用于处理具有有限值的随机变量，如投掷硬币和伯努利分布的情形。离散分布是由所谓的概率质量函数（PMF）定义的，连续分布用于处理连续的（理论上）有无限数量的值的随机变量。想想用声音传感器测量的速度和加速度。连续分布是由概率密度函数（PDF）定义的。

这两种分布类型在数学处理上有所不同：通常连续分布使用积分 ∫ 而离散分布使用求和Σ。以期望值为例：

$$E[X] = \sum_x xP_{discrete}(x), \text{if } X \sim P_{discrete}$$

$$E[X] = \int_x xP_{continuous}(x)dx, \text{if } X \sim P_{continuous}(x)$$

下面我们将详细介绍各种常见的概率分布类型，正如上所说，概率分布可以分为离散型随机变量分布和连续性随机变量分布。离散型随机变量分布常见的有伯努利分布（Bernoulli Distribution）、二项分布（Binomial Distribution）、泊松分布（Poisson Distribution）等，而常见的连续型随机变量分布包括均匀分布（Uniform Distribution）、指数分布（Exponential Distribution）、正态分布等。

常见的数据类型

在解释各种分布之前，我们先看看常见的数据类型有哪些，数据类型可分为离散型和连续型。

离散型数据：数据只能取特定的值，比如，当你掷一个骰子的时候，可能的结果只有 1，2，3，4，5，6 而不会是 1.5 或者 2.45。

连续型数据：数据可以在给定的范围内取任何值，给定的范围可以是有限的或无限的，比如一个女孩的体重或者身高，或者道路的长度。一个女孩的体重可以是 54 kgs，54.5 kgs，或 54.5436kgs。

分布的类型

伯努利分布

最简单的离散型随机变量分布是伯努利分布，我们从这里开始讨论。

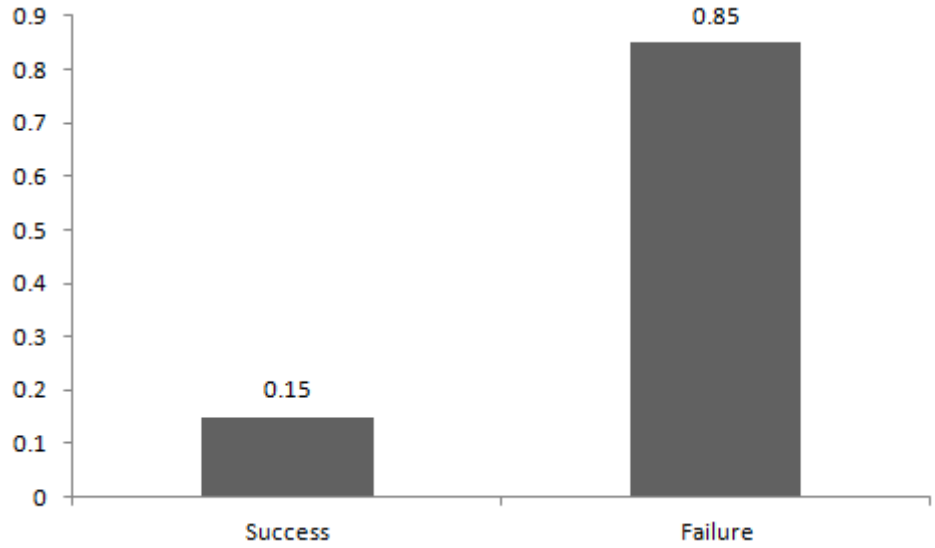
一个伯努利分布只有两个可能的结果，记作 1（成功）和 0（失败），只有单次伯努利试验。设定一个具有伯努利分布的随机变量 X，取值为 1 即成功的概率为 p，取值为 0 即失败的概率为 q 或者 1-p。

若随机变量 X 服从伯努利分布，则概率函数为：

$$P(x) = \begin{cases} 1-p, & x = 0 \\ p, & x = 1 \end{cases}$$

成功和失败的概率不一定要相等。比如当我和一个运动员打架的时候，他的胜算应该更大，在这时候，我的成功概率是 0.15，而失败概率是 0.85。

下图展示了我们的战斗的伯努利分布。



如上图所示，我的成功概率=0.15，失败概率=0.85。期望值是指一个概率分布的平均值，对于随机变量 X，对应的期望值为：E(X) = 1*p + 0*(1-p) = p，而方差为 V(X) = E(X^2) – [E(X)]^2 = p – p^2 = p(1-p)

实际上还有很多关于伯努利分布的例子，比如明天是晴天还是雨天，这场比赛中某一队输还是赢，等等。

二项分布

现在回到掷硬币的案例中，当掷完第一次，我们可以再掷一次，也就是存在多个伯努利试验。第一次为正不代表以后也会为正。那么设一个随机变量 X，它表示我们投掷为正面的次数。X 可能会取什么值呢？在投掷硬币的总次数范围内可以是任何非负整数。

如果存在一组相同的随机事件，即一组伯努利试验，在上例中为连续掷硬币多次。那么某随机事件出现的次数即概率服从于二项分布，也称为多重伯努利分布。

任何一次试验都是互相独立的，前一次试验不会影响当前试验的结果。两个结果概率相同的试验重复 n 次的试验称为多次伯努利试验。二项分布的参数为 n 和 p，n 是试验的总次数，p 是每一次试验的成功概率。

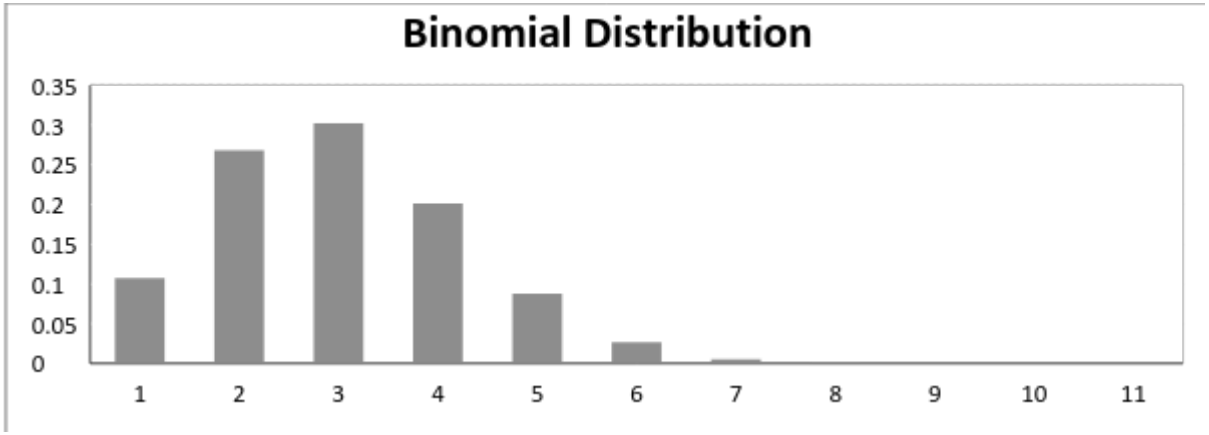
根据以上所述，一个二项分布的性质为：

1. 每一次试验都是独立的；
2. 只有两个可能的结果；
3. 进行 n 次相同的试验；
4. 所有试验中成功率都是相同的，失败的概率也是相同的。

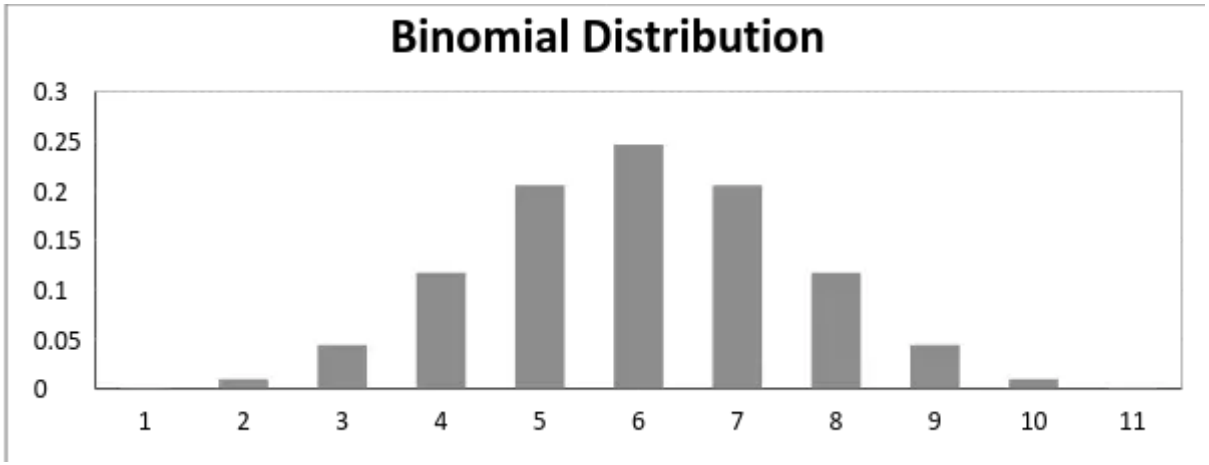
二项分布的数学表达式为：

$$P(x) = \frac{n!}{(n-x)!\,x!} p^x q^{n-x}$$

成功概率和失败概率不相等的二项分布看起来如下图所示：



而成功概率和失败概率相等的二项分布看起来如下图所示：



二项分布的平均值表示为 μ = n*p，而方差可以表示为 Var(X) = n*p*q。

泊松分布

如果你在一个呼叫中心工作，一天内会接到多少次呼叫呢？多少次都可能！在呼叫中心一天能接到多少次呼叫可以用泊松分布建模。这里有几个例子：

1. 一天内医院接到的紧急呼叫次数；
2. 一天内地方接到的偷窃事件报告次数；
3. 一小时内光顾沙龙的人数；
4. 一个特定城市里报告的自杀人数；
5. 书的每一页的印刷错误次数。

现在你可以按相同的方式构造很多其它的例子。泊松分布适用于事件发生的时间和地点随机分布的情况，其中我们只对事件的发生次数感兴趣。泊松分布的主要特点为如下：

1. 任何一个成功事件不能影响其它的成功事件；
2. 经过短时间间隔的成功概率必须等于经过长时间间隔的成功概率；
3. 时间间隔趋向于无穷小的时候，一个时间间隔内的成功概率趋近零。

在泊松分布中定义的符号有：

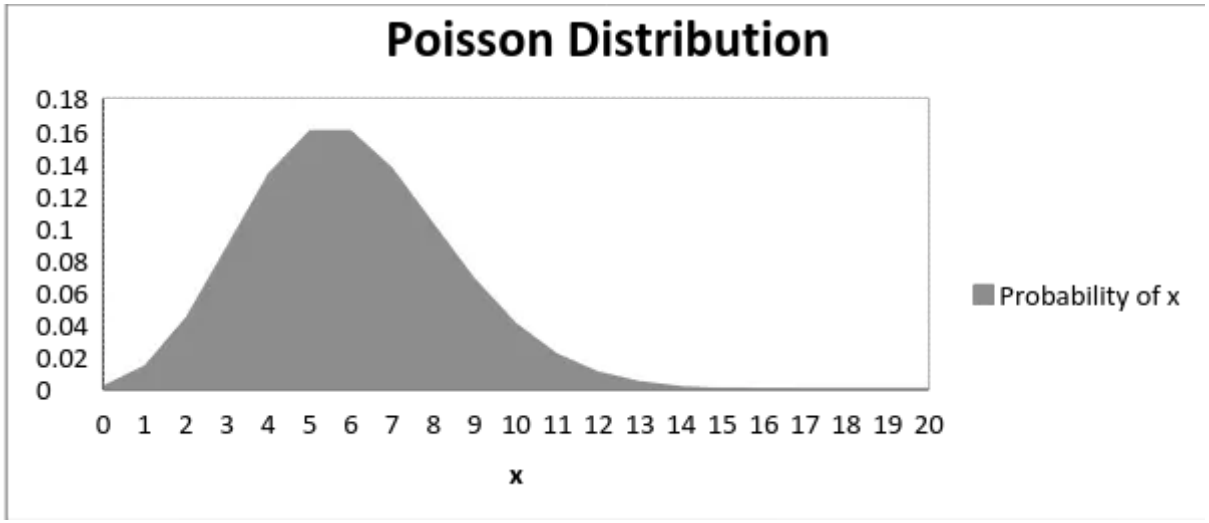
- λ是事件的发生率；
- t 是事件间隔的长度；
- X 是在一个时间间隔内的事件发生次数。

设 X 是一个泊松随机变量，那么 X 的概率分布称为泊松分布。以μ表示一个时间间隔 t 内平均事件发生的次数，则 μ=λ*t；

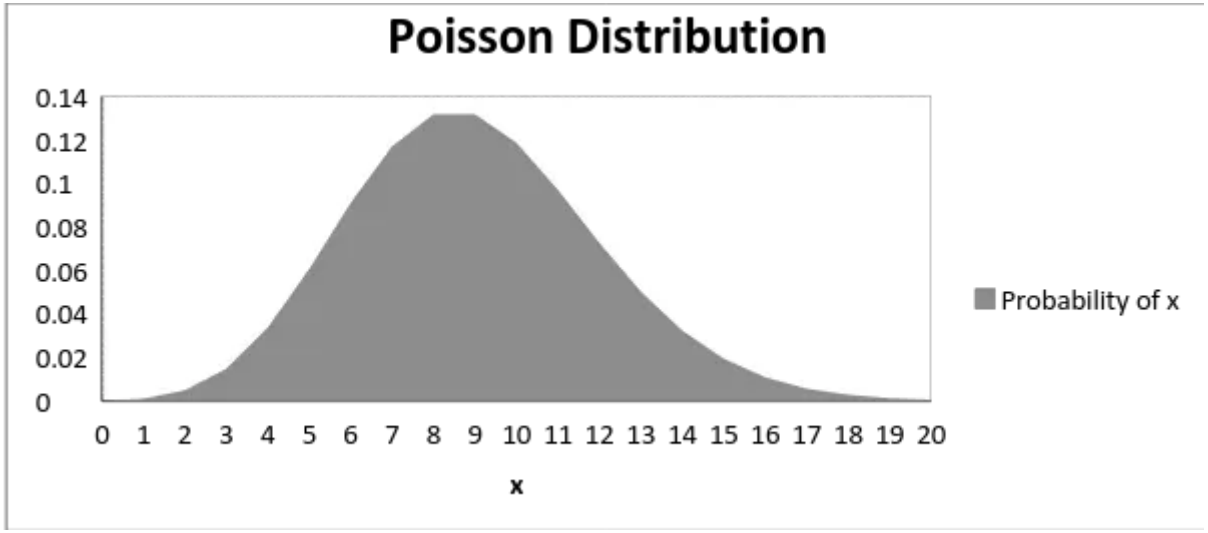
X 的概率分布函数为：

$$P(X = x) = e^{-\mu}\frac{\mu^x}{x!} \qquad for\ x\ =\ 0, 1, 2, \ldots..$$

泊松分布的概率分布图示如下，其中μ为泊松分布的参数：



下图展示了均值增加时的分布曲线的变化情况：



如上所示，当均值增加时，曲线向右移动。泊松分布的均值和方差为：

均值：E(X) = μ

方差：Var(X) = μ

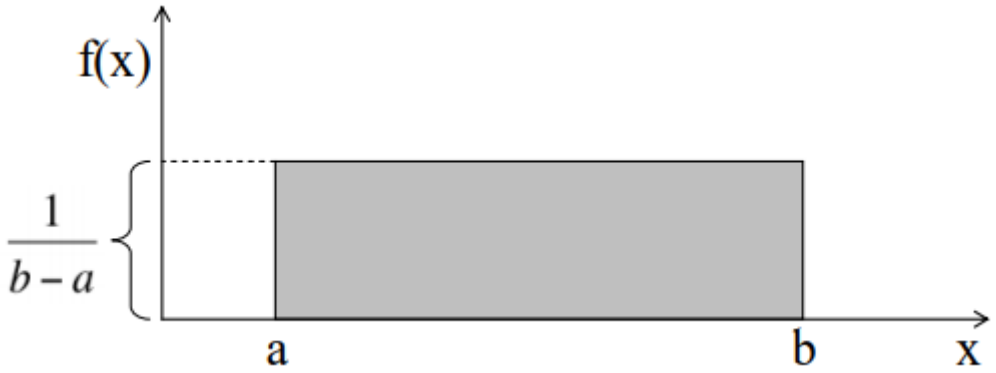
均匀分布

假设我们在从 a 到 b 的一段线段上等距地选择一个区间的概率是相等的，那么概率在整个区间 [a,b] 上是均匀分布的，概率密度函数也不会随着变量的更改而更改。均匀分布和伯努利分布不同，随机变量的取值都是等概率的，因此概率密度就可以表达为区间长度分之一，如果我们取随机变量一半的可能值，那么其出现的概率就为 1/2。

假定随机变量 X 服从均匀分布，那么概率密度函数为：

$$f(x) = \frac{1}{b-a} \qquad for\ -\infty < a \le x \le b < \infty$$

均匀分布曲线图如下所示，其中概率密度曲线下面积为随机变量发生的概率：



我们可以看到均匀分布的概率分布图呈现为一个矩形，这也就是均匀分布又称为矩形分布的原因。在均匀分布中，a 和 b 都为参数，也即随机变量的取值范围。

服从均匀分布的随机变量 X 也有均值和方差，它的均值为 E(X) = (a+b)/2，方差为 V(X) = (b-a)^2/12

标准均匀分布的密度函数参数 a 取值为 0，b 取值为 1，因此标准均匀分布的概率密度可以表示为：

$$f(x) = \begin{cases} 1, & 0 \le x \le 1 \\ 0, & otherwise \end{cases}$$

指数分布

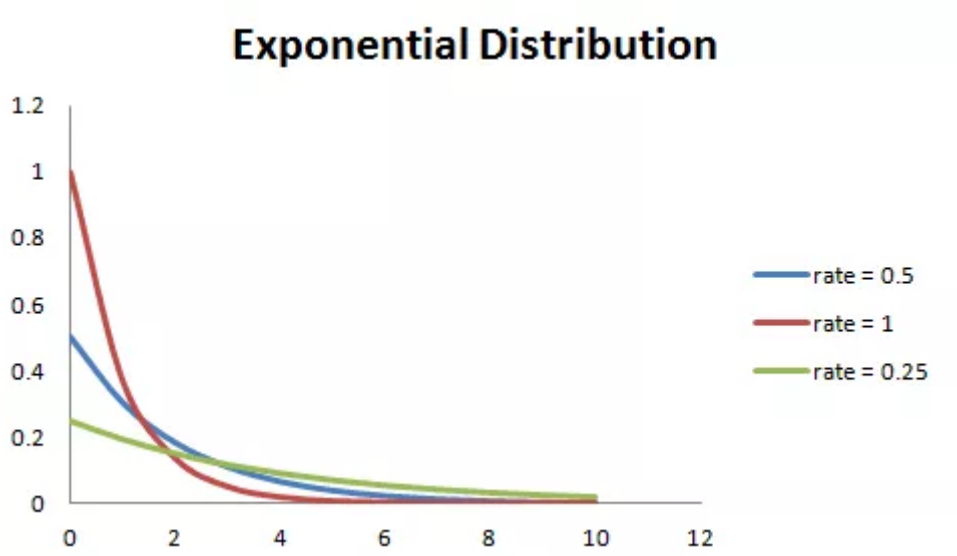
现在我们再次考虑电话中心案例，那么电话间隔的分布是怎么样的呢？这个分布可能就是指数分布，因为指数分布可以对电话的时间间隔进行建模。其它案例可能还有地铁到达时间的建模和空调设备周期等。

在深度学习中，我们经常会需要一个在 x=0 处取得边界点 (sharp point) 的分布。为了实现这一目的，我们可以使用指数分布（exponential distribution）：

$$p(x;\lambda) = \lambda \mathbf{1}_{x\geq 0} \exp(-\lambda x).$$

指数分布使用指示函数 (indicator function)1x≥0，以使当 x 取负值时的概率为零。

其中 λ >0 为概率密度函数的参数。随机变量 X 服从于指数分布，则该变量的均值可表示为 E(X) = 1/λ、方差可以表示为 Var(X) = (1/λ)^2。如下图所示，若λ较大，则指数分布的曲线下降地更大，若λ较小，则曲线越平坦。如下图所示：



以下是由指数分布函数推导而出的简单表达式：

P{X≤x} = 1 − exp(-λx)，对应小于 x 的密度函数曲线下面积。

P{X>x} = exp(-λx)，代表大于 x 的概率密度函数曲线下面积。

P{x1<X≤ x2} =exp(-λx1)-exp(-λx2)，代表 x1 点和 x2 点之间的概率密度函数曲线下面积。

正态分布（高斯分布）

实数上最常用的分布就是正态分布（normal distribution），也称为高斯分布（Gaussian distribution）。因为该分布的普遍性，尤其是中心极限定理的推广，一般叠加很多较小的随机变量都可以拟合为正态分布。正态分布主要有以下几个特点：

1. 所有的变量服从同一均值、方差和分布模式。

2. 分布曲线为钟型，并且沿 x=μ对称。

3. 曲线下面积的和为 1。

4. 该分布左半边的精确值等于右半边。

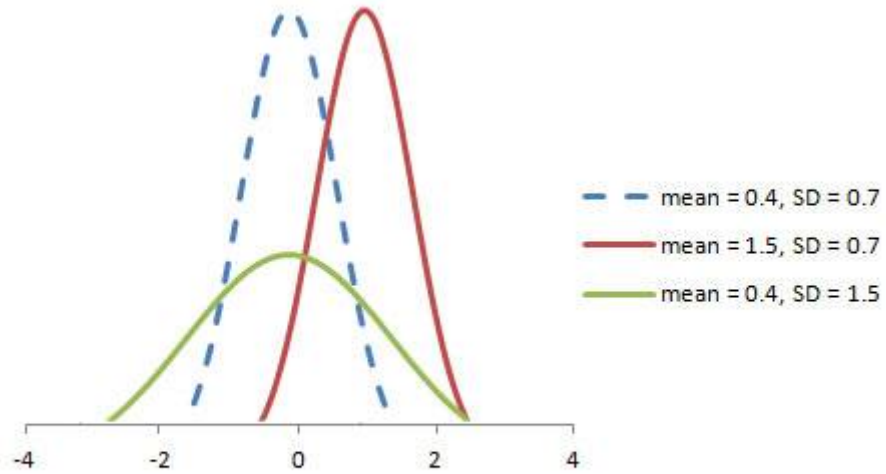
正态分布和伯努利分布有很大的不同，然而当伯努利试验的次数接近于无穷大时，他们的分布函数基本上是相等的。

若随机变量 X 服从于正态分布，那么 X 的概率密度可以表示为：

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\{-\frac{1}{2}(\frac{x-u}{\sigma})^2\}} \qquad for -\infty < x < \infty.$$

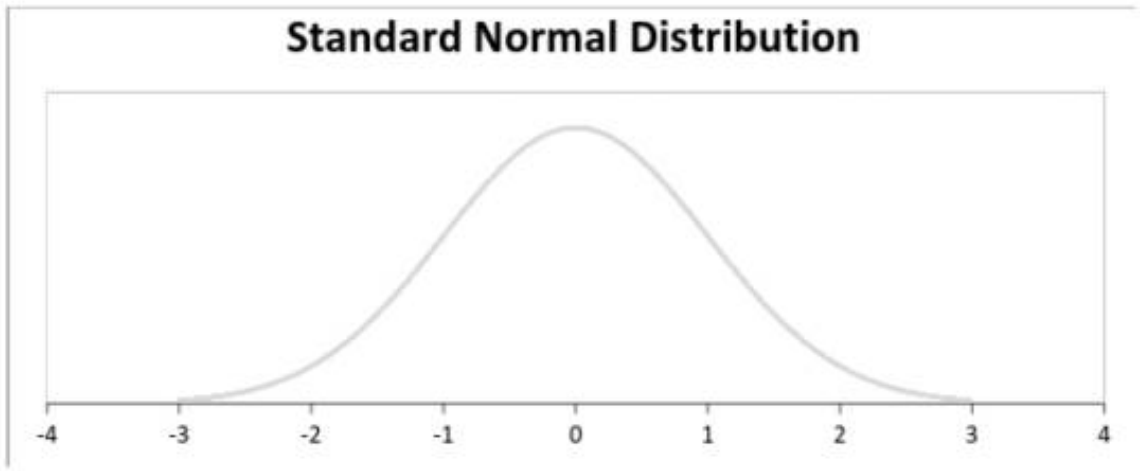
随机变量 X 的均值可表示为 E(X) = μ、方差可以表示为 Var(X) = σ^2。其中均值μ和标准差σ为高斯分布的参数。

随机变量 X 服从于正态分布 N (μ, σ)，可以表示为：



标准正态分布可以定义为均值为 0、方差为 1 的分布函数，以下展示了标准正态分布的概率密度函数和分布图：

$$f(x)=\frac{1}{\sqrt{2\pi}}e^{-x^2/2} \qquad for -\infty < x < \infty$$



分布之间的关系

伯努利分布和二项分布的关系

1. 二项分布是伯努利分布的单次试验的特例，即单词伯努利试验；
2. 二项分布和伯努利分布的每次试验都只有两个可能的结果；
3. 二项分布每次试验都是互相独立的，每一次试验都可以看作一个伯努利分布。

泊松分布和二项分布的关系

以下条件下，泊松分布是二项分布的极限形式：

1. 试验次数非常大或者趋近无穷，即 $n \to \infty$；
2. 每次试验的成功概率相同且趋近零，即 $p \to 0$；
3. $np = \lambda$ 是有限值。

正态分布和二项分布的关系 & 正态分布和泊松分布的关系

以下条件下，正态分布是二项分布的一种极限形式：

1. 试验次数非常大或者趋近无穷，即 $n \to \infty$；
2. p 和 q 都不是无穷小。

参数 $\lambda \to \infty$的时候，正态分布是泊松分布的极限形式。

指数分布和泊松分布的关系

如果随机事件的时间间隔服从参数为 λ的指数分布，那么在时间周期 t 内事件发生的总次数服从泊松分布，相应的参数为 λt。

测试

读者可以完成以下简单的测试，检查自己对上述概率分布的理解程度：

1. 服从标准正态分布的随机变量计算公式为：

a. (x+μ) / σ

b. (x-μ) / σ

c. (x-σ) / μ

2. 在伯努利分布中，计算标准差的公式为：

a. p (1 − p)

b. SQRT(p(p − 1))

c. SQRT(p(1 − p))

3. 对于正态分布，均值增大意味着：

a. 曲线向左移

b. 曲线向右移

c. 曲线变平坦

4. 假定电池的生命周期服从 λ = 0.05 指数分布，那么电池的最终使用寿命在 10 小时到 15 小时之间的概率为：

a.0.1341

b.0.1540

c.0.0079

结语

在本文中，我们从最基本的随机事件及其概念出发讨论对概率的理解。随后我们讨论了最基本的概率计算方法与概念，比如条件概率和贝叶斯概率等等。文中还讨论了随机变量的独立性和条件独立性。此外，本文更是详细介绍了概率分布，包括离散型随机变量分布和连续型随机变量分布。本文主要讨论了基本的概率定理与概念，其实这些内容在我们大学的概率论与数理统计课程中基本上都有详细的解释。而对于机器学习来说，理解概率和统计学知识对理解机器学习模型十分重要，以它为基础我们也能进一步理解结构化概率等新概念。

原文链接：

https://medium.com/towards-data-science/probabiliy-theory-basics-4ef523ae0820

https://www.analyticsvidhya.com/blog/2017/09/6-probability-distributions-data-science/

也可以加一下老胡的微信
围观朋友圈~~~

推荐阅读 （点击标题可跳转阅读）

深度学习的四个学习阶段!

2021年，机器学习研究风向要变了?

【机器学习】随机森林是我最喜欢的模型

Python之父：Python 4.0可能不会来了

【2021版】机器学习、深度学习调参手册

亚马逊首席科学家李沐博士：工作五年反思

【小抄】机器学习常见知识点总结（2021）

引用次数在15000次以上的都是什么神仙论文?

【下载】80页笔记看遍机器学习基本概念、算法、模型

老铁，三连支持一下，好吗? ↓↓↓