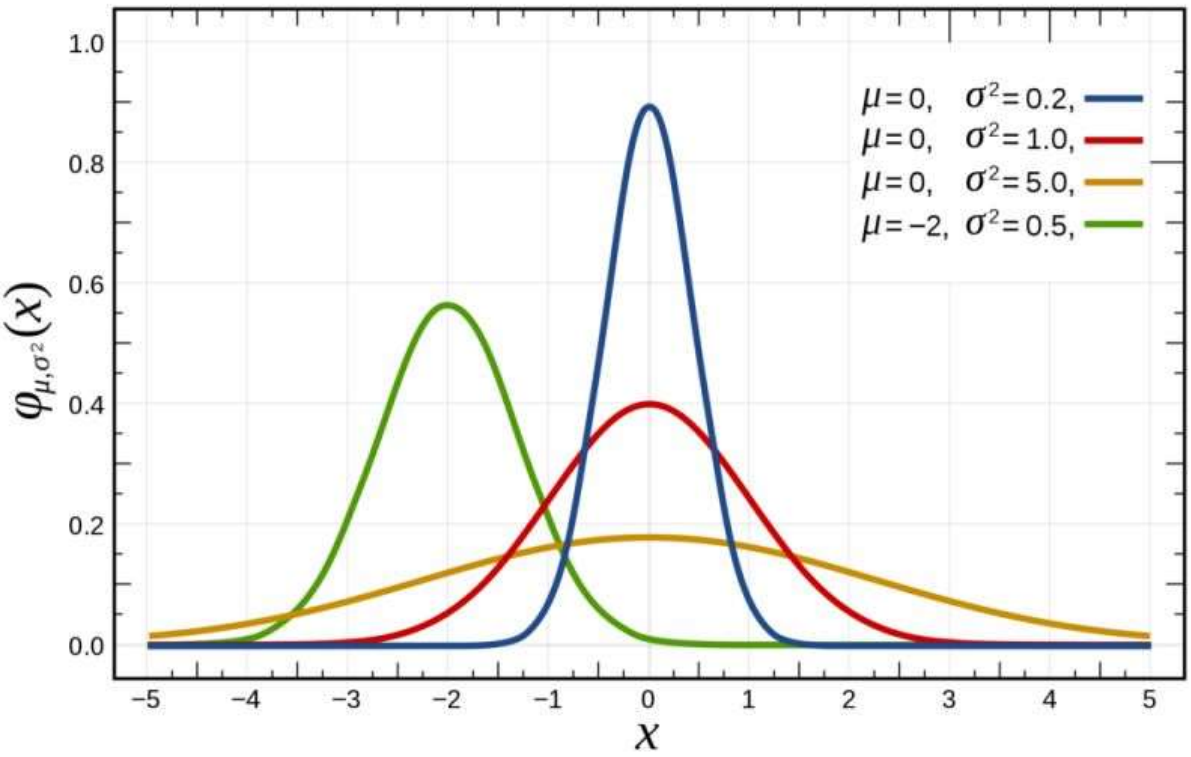
学界 | 为什么数据科学家都钟情于最常见的正态分布？



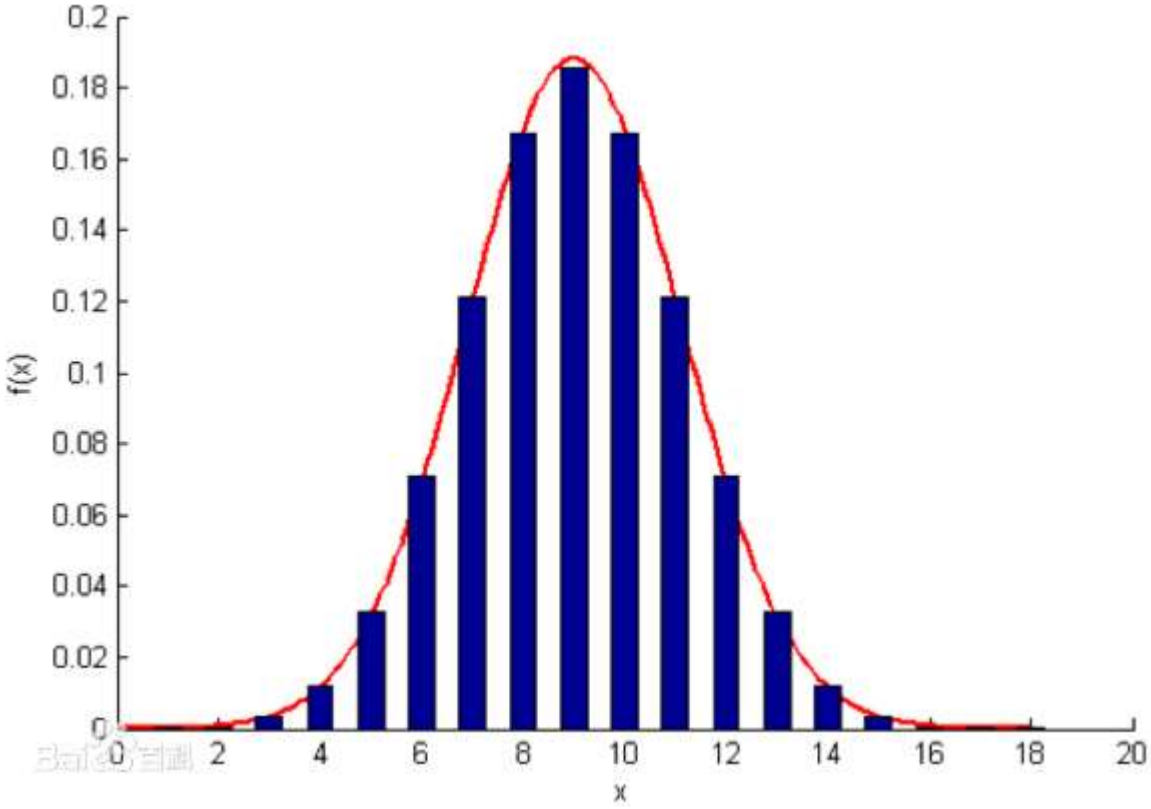
大数据文摘出品

编译：JonyKai、元元、云舟

对于深度学习和机器学习工程师们来说，正态分布是世界上所有概率模型中最重要的一个。即使你没有参与过任何人工智能项目，也一定遇到过高斯模型，今天就让我们来看看高斯过程为什么这么受欢迎。

高斯分布（Gaussian distribution），也称正态分布，最早由A.棣莫弗在求二项分布的渐近公式中得到。C.F.高斯在研究测量误差时从另一个角度导出了它。P.S.拉普拉斯和高斯研究了它的性质。是一个在数学、物理及工程等领域都非常重要的概率分布，在统计学的许多方面有着重大的影响力。

正态曲线呈钟型，两头低，中间高，左右对称因其曲线呈钟形，因此人们又经常称之为钟形曲线。



若随机变量X服从一个数学期望为μ、方差为σ^2的正态分布，记为N(μ，σ^2)。其概率密度函数为正态分布的期望值μ决定了其位置，其标准差σ决定了分布的幅度。当μ = 0,σ = 1时的正态分布是标准正态分布。

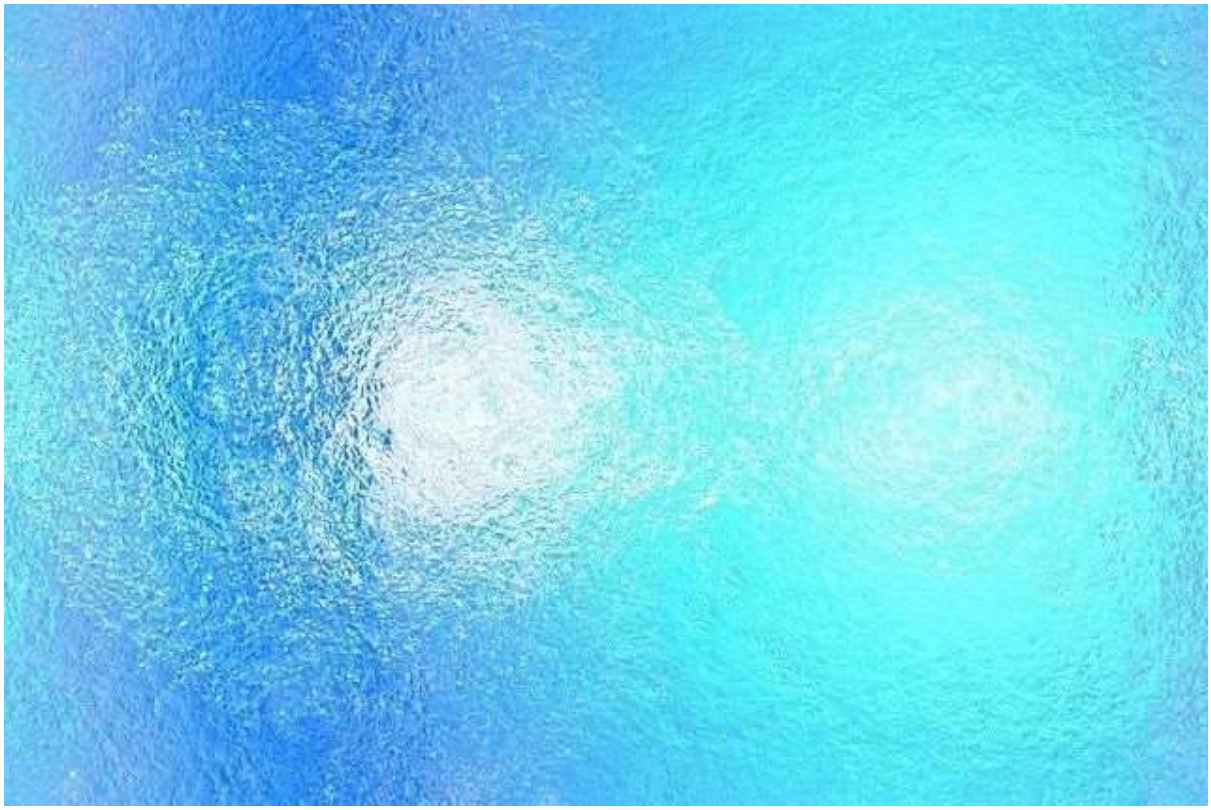
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\mu - x)^2}{2\sigma^2}}$$

高斯概率分布的数学表达式

在自然现象中随处可见

所有模型都是错的，但有些是有用的

—George Box

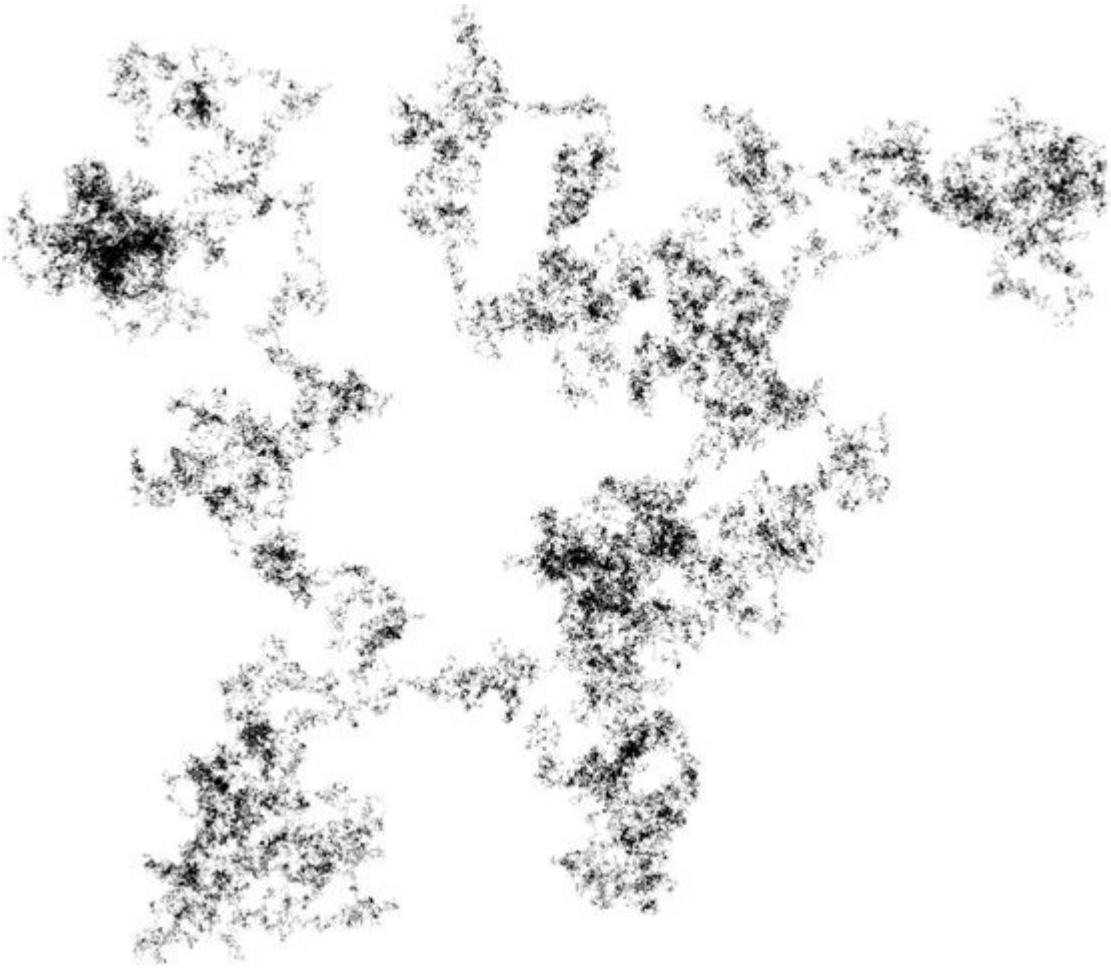


正在扩散的粒子的位置可以用正态分布来描述

正态分布有极其广泛的实际背景，生产与科学实验中很多随机变量的概率分布都可以近似地用正态分布来描述。例如，在生产条件不变的情况下，产品的强力、抗压强度、口径、长度等指标；同一种生物体的身长、体重等指标；同一种种子的重量；测量同一物体的误差；弹着点沿某一方向的偏差；某个地区的年降水量；以及理想气体分子的速度分量，等等。

一般来说，如果一个量是由许多微小的独立随机因素影响的结果，那么就可以认为这个量具有正态分布。从理论上看，正态分布具有很多良好的性质，许多概率分布可以用它来近似；还有一些常用的概率分布是由它直接导出的，例如对数正态分布、t分布、F分布等。

数学原因：中心极限定理



二维空间上进行200万步的随机游走之后得到的图案

中心极限定理的内容为：大量独立随机变量的和经过适当标准化之后趋近于正态分布，与这些变量原本的分布无关。比如，随机游走的总距离就趋近于正态分布。下面我们介绍三种形式的中心极限定理：

独立同分布的中心极限定理

设随机变量X1，X2，.....Xn，.....独立同分布，并且具有有限的数学期望和方差：E(Xi)=μ，D(Xi)=σ^2 (i=1,2....)，则对任意x，分布函数为

$$F_n(x) = P\left\{\frac{\sum_{i=1}^{n} - n\mu}{\sqrt{n}\sigma} \le x\right\}$$

满足

$$\lim_{n\to\infty} F_n(x) = \lim_{n\to\infty}\left\{\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \le x\right\} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{t^2}{2}}dt = \emptyset(x)$$

该定理说明，当n很大时，随机变量

$$Y_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma}$$

近似地服从标准正态分布N(0，1)。因此，当n很大时，

$$\sum_{i=1}^{n} X_i = \sqrt{n}\sigma Y_n + n\mu$$

近似地服从正态分布N(nμ，nσ^2)．该定理是中心极限定理最简单又最常用的一种形式，在实际工作中，只要n足够大，便可以把独立同分布的随机变量之和当作正态变量。这种方法在数理统计中用得很普遍，当处理大样本时，它是重要工具。

棣莫佛 - 拉普拉斯定理

设随机变量X(n=1,2,...,)服从参数为n，p(0<p<1)的二项分布，则对于任意有限区间(a，b)有

$$\lim_{n\to\infty} P\left\{a < \frac{x_n - np}{\sqrt{np(1-p)}} \le b\right\} = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}dt$$

该定理表明，正态分布是二项分布的极限分布，当数充分大时，我们可以利用上式来计算二项分布的概率。

不同分布的中心极限定理

设随机变量X1，X2，.....Xn，.....独立同分布，它们的概率密度分别为$f_{xk}(x)$，并有$E(X_k)=\mu k$，$D(X_k)=\sigma_k^2$，（k=1,2.....）

$$B_n^2 = \sum_{i=1}^{n}\sigma_k^2$$

$$Y_n = \frac{\sum_{k=1}^{n} X_k - \sum_{k=1}^{n}\mu k}{B_n}$$

若对任意正数τ，有：

$$\lim_{n\to\infty}\frac{1}{B_n^2}\sum_{k=1}^{n}\int_{|x-\mu_k>\tau B_n|}\left(x-\mu_k\right)^2 f_{x_k}(x)\,dx=0$$

对任意x，随机变量Yn的分布函数Fn(x)，满足：

$$\lim_{n\to\infty}F_n(x)=\lim_{n\to\infty}P\left\{\frac{\sum_{k=1}^{n}X_k-\sum_{k=1}^{n}\mu k}{B_n}\le x\right\}=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{t^2}{2}}dt$$

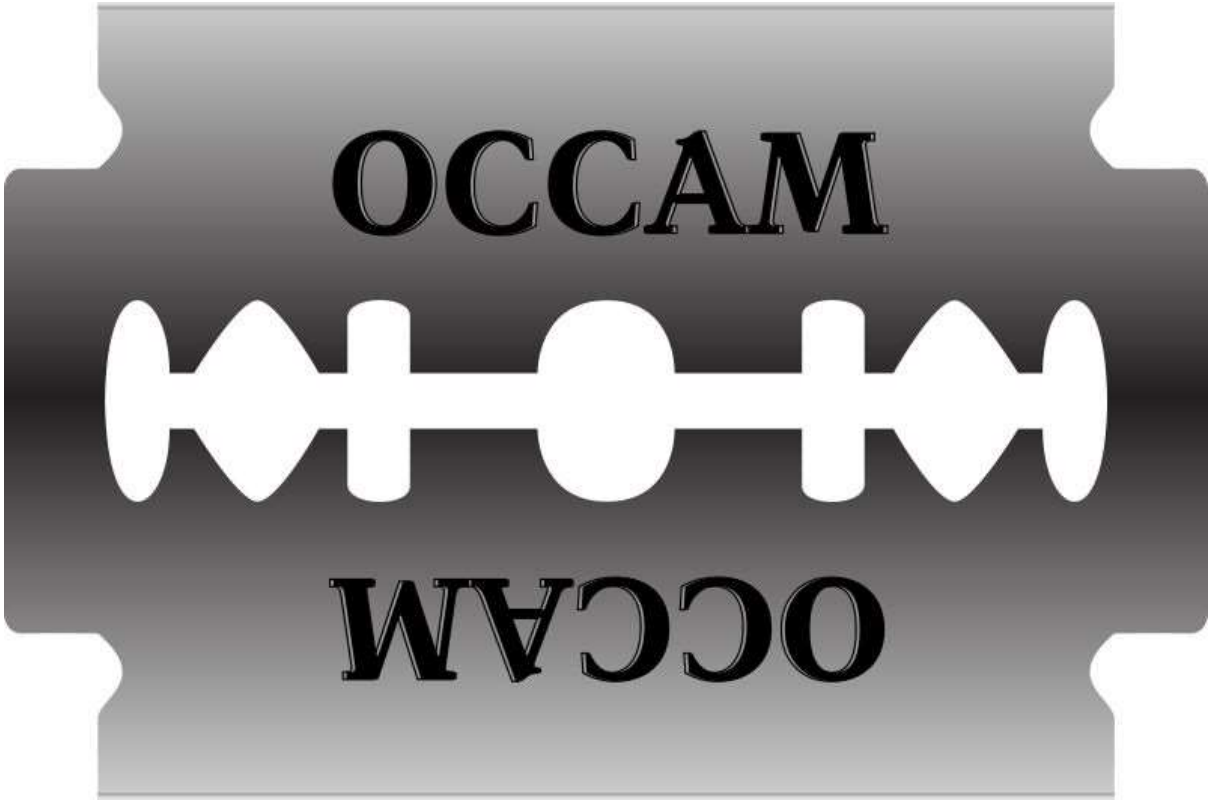
该定理说明：所研究的随机变量如果是有大量独立的而且均匀的随机变量相加而成，那么它的分布将近似于正态分布。

万变不离其宗

与其他很多分布不同，正态分布进行适当的变换之后，仍是正态分布。

- 两个正态分布之积仍是正态分布
- 两个独立的服从正态分布的随机变量之和服从正态分布
- 对一个正态分布进行高斯卷积还是正态分布
- 正态分布经过傅立叶变换之后仍是正态分布

简洁



奥卡姆剃刀强调一个哲学原则：在其他条件都相同下，最简单的解就是最好的解。

对于任何一个用正态分布拟合的随机分布，都可能存在一个多参数，更复杂，更准确的解法。但是我们仍然会倾向于选用正态分布，因为它在数学上很简洁。

- 它的均值（mean）、中值（median）和众数（mode）都相同
- 只需要用两个参数就可以确定整个分布

图形特性：

- 集中性：正态曲线的高峰位于正中央，即均数所在的位置。
- 对称性：正态曲线以均数为中心，左右对称，曲线两端永远不与横轴相交。
- 均匀变动性：正态曲线由均数所在处开始，分别向左右两侧逐渐均匀下降。
- 曲线与横轴间的面积总等于1，相当于概率密度函数的函数从正无穷到负无穷积分的概率为1。即频率的总和为100%。

相关报道：

https://towardsdatascience.com/why-data-scientists-love-gaussian-6e7a7b726859