The Volume of High-dimensional Unit Ball

http://freemind.pluskid.org/misc/the-volume-of-high-dimensional-unit-ball

posted on Free Mind on February 9, 2016
generated with pandoc on February 29, 2016
category: MISC

tags: Fun, Intuition

高维度数据是现代机器学习的一个重要特点，而人脑无法直观地对超过三维的东西进行 visualize （至少我自己做不到）一直是我觉得作为一个人类最痛苦的事情之一，如果什么时候有机会做一回超人或者宇宙人什么的，能够选择技能树的话，我大概会优先选择这一项吧。正因为如此在高维空间中的许多现象看起来非常“反直觉”。比如今天我们要讲的结论是：高维空间的单位球（半径为 1 的球）的体积随着维度的增大趋向于 0。我跟一个朋友提起这件事情的时候她说这是她在听说了“奇数和整数一样多”这件事以来让她最为震惊的事情。

首先，本文的主要内容来自于 Avrim Blum、John Hopcroft 和 Ravindran Kannan 的一本叫做《Foundations of Data Science》的书里的第二章。这本书还没有出版，在作者主页上可以下载到 draft。

我们都知道一维的单位球（是一个线段）的“体积”（也就是长度）为 2，二维情况下的“体积”也就是（面积）为 $\pi \approx 3.14$，三维的体积是 $4\pi/3 \approx 4.19$。看起来似乎是在单调递增。但是只看三个 case 要总结出规律其实比较勉强，这也是为什么只能 visualize 到三维为止是一个局限性非常大的技能，因为基本上总结不出在高维里应该“长什么样”的直观规律来。

不过要讲 intuition 的话，其实还是有的。虽然几何课上没有学过高维球体的“体积”的公式，但是我们知道立方体的体积永远都是所有的边相乘。所以 n 维的 unit cube 的体积永远都是 1，然后我们可以将 unit ball 和 unit cube 相比较，就能得到一些信息。

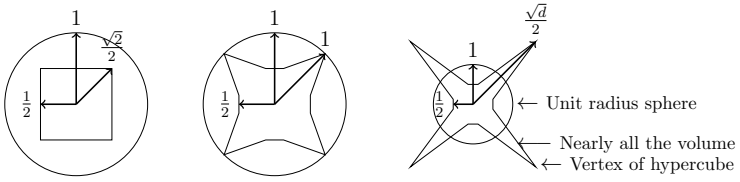

Figure 1: .....

这里我直接盗取了书中的插图 2.3：如上图最左边所示，我们可以将一个单位正方形和一个单位圆同时置于原点。从原点到圆上所有点的距离都是 1，而从圆到正方形的最短距离是到边的中点的距离，为 $1/2$，而最远的距离为到两边交点的距离，为 $\sqrt{2}/2$。中间一个图是 4 维的情况，虽然在网上见过各种各样的 4 维立方体的 visualization，但是我还是对它长什么样并没有一个什么具体的概念。好在这里我们的目的只是比较两个东西的体积大小，而有几样东西我们是知道的：

- 原点到球面上任意一点的距离仍然是 1，和维度无关。

- 原点到 cube 的 “面” 的距离仍然是 $1/2$，因为 cube 的边长被固定为 1，而这是边长的一半。
- 原点到 cube 的 “顶点” 的距离可以根据勾股定理算出来，比如 4 维的情况就是 $\sqrt{4(1/2)^2} = 1$。

可以看到虽然 2 维的时候 unit cube 还完全包含在 unit ball 内部（因此 unit ball 的体积大于 1），但是到 4 维的时候 unit cube 的顶点已经接触到 unit ball 的外壳。推广到 d 维的时候，顶点到原点的距离变成 $\sqrt{d(1/2)^2} = \sqrt{d}/2$，当 d 变大的时候这个长度可以变得很大，最终效果就如上面图中最右边所示。此外，虽然图中只画了 4 个顶点，但是实际上 d 维 cube 的顶点数目是 2^d 个，最终结果就是 unit cube 的大部分 volume 集中在顶点的地方，而内部的部分的比例越来越小。相比较起来单位球的体积也就越来越小。如果你觉得这个 intuition 还不够直观，可以通过积分来具体地计算 d 维球体的体积。在极坐标下单位球的体积可以写成

$$V(d) = \int_{S^d} \int_{r=0}^{1} r^{d-1}\, dr\, d\Omega = \int_{s^d} d\Omega \int_{r=0}^{1} r^{d-1}\, dr = \frac{1}{d} \int_{s^d} d\Omega = \frac{A(d)}{d}$$

其中 $A(d)$ 是 d 维单位球面的表面积，而 $A(d)$ 的计算可以通过一个迂回的办法来实现。考虑如下的积分

$$I(d) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left(-(x_1^2 + x_2^2 + \cdots + x_d^2)\right)\, dx_d \cdots dx_2\, dx_1$$

这个积分很好计算，将 exp 中的变量分开，我们可以分别对每个 $x_1, \ldots, x_d$ 求积分，其中每个独立的积分可以直接通过一元正态分布的概率密度公式得到：

$$I(d) = \left(\int_{\infty}^{\infty} \exp(-x^2)\, dx\right)^d = (\sqrt{\pi})^d = \pi^{d/2}$$

而另一方面，我们也可以通过极坐标来计算 $I(d)$，如下：

$$I(d) = \int_{S^d} d\Omega \int_0^{\infty} \exp(-r^2) r^{d-1}\, dr = A(d) \int_0^{\infty} \exp(-r^2) r^{d-1}\, dr$$

而剩下部分的积分通过一步变量代换，可以发现和 Gamma 函数的定义匹配起来：

$$\int_0^{\infty} \exp(-r^2) r^{d-1}\, dr = \frac{1}{2} \int_0^{\infty} e^{-t} t^{d/2-1}\, dt = \frac{1}{2}\Gamma\left(d/2\right)$$

结合两种不同的计算 $I(d)$ 的方法，可以得到 d 维单位球的表面积公式，以及体积公式为：

$$A(d) = \frac{2\pi^{d/2}}{\Gamma(d/2)}, \quad V(d) = \frac{2}{d}\frac{\pi^{d/2}}{\Gamma(d/2)}$$

现在我们可以很清楚地看到，球体的体积公式中，分子是指数级增长，而分母的 Γ 函数是更加可怕的阶层级别的增长，随着 $d \to \infty$, $V(d) \to 0$。

到了这里，即便是超级 counter intuitive，也还是只能接受了。然而“问题” 到底出在哪里呢？仔细想一想，好像 “体积” 的概念最初在提出来的时候并没有考虑到非常高维度的情况，也没有考虑会将不同维度下的 “体积” 用来相互比较。一个 cube 的体积是所有边长相乘，似乎总是会出现指数增长或者指数 decay 这样的不太好的情况，如果我们 normalize 一下会怎样呢？比如我们定义一个新的体积，在 d 维的情况下计算原来体积的基础上再开 d 次方根：

$$\tilde{V}(d) = (V(d))^{1/d} = \left(\frac{2}{d}\right)^{1/d} \frac{\sqrt{\pi}}{(\Gamma(d/2))^{1/d}}$$

例如一个 Gamma 函数的 lower bound：$\Gamma(d/2) \geq (d/(2e))^{d/2}$，我们得到：

$$\tilde{V}(d) \leq \left(\frac{2}{d}\right)^{1/d} \sqrt{\frac{2e\pi}{d}}$$

其中随着 d 趋向于无穷大，第一项趋向于 1，然而第二项还是会趋向于零。所以看来我们的 normalization 的补救并不能阻止高维 unit ball 的体积趋向于零。