CS 6966 : High-Dimensional Data Analysis (Fall 2024) : Jeff M. Phillips

L1: Curse of Dimensionality: Basic Geometry

Vectors $x = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d$

For $a, b \in \mathbb{R}^d$

- ℓ_2 or Euclidean distance: $\|a - b\| = \|a - b\|_2 = \sqrt{\sum j = 1^d (a_j - b_j)^2}$
- ℓ_∞ or max distance: $\|a - b\|_\infty = \max_{j=1}^d |a_j - b_j|$

1. Volume of cube vs. inscribed sphere with dimension d.

$[-1, 1]^d$ square has volume 2^d for all d
$2 \times 2 \times \ldots \times 2 = 2^d$

volume of ball inscribed
radius = 1

$$\frac{r^d \pi^{d/2}}{\Gamma(d/2+1)} \approx \frac{1 \cdot \pi^{d/2}}{(d/2)!}$$

d	ball-vol	box-vol
2	3.14	4
4	4.93	16
6	5.16	64
8	4.05	256
10	2.55	1024
12	1.33	4096
14	0.60	16384
16	0.24	65536
18	0.08	262144

what happens for 1x1x1 box? Is it inscribed?

2. Approx orthogonality of high-d Gaussians

$a, b \sim G_d(0, 1)$ a d-dimensional Gaussian

$E[(a_i - b_i)^2] = E[a_i^2] + E[b_i^2] - 2E[a_i b_i] = Var[a_i] + Var[b_i] - 2E[a_i]E[b_i] = 2$
so $\|a - b\|^2 = 2d$

Need also: $E[||a||^2] = d$

Pythagorean if a orthogonal to b (w.r.t 0) then $\|a\| = \sqrt{d}$, $\|b\| = \sqrt{d}$ and hence $\|a - b\|^2 = \|a\|^2 + \|b\|^2 = 2d$

3. Big Annulus

For any object $A \subset \mathbb{R}^d$ let
$$(1 - \varepsilon)A = \{(1 - \varepsilon)x | x \in A\}$$

(imagine shrinking into origin .. but works more generally)

Thm: $Vol((1 - \varepsilon)A = (1 - \varepsilon)^d Vol(A)$

proof: decompose A into a set of d-dim cubes $C_1, C_2, ...$
so $Vol(A) = \sum_j Vol(C_j)$
Each C_j has side length l_j, and $Vol(C_j) = l_j^d$
We can replace $(1-\varepsilon)A$ by same series $(1-\varepsilon)C_1, (1-\varepsilon)C_2, ...$

$$Vol((1-\varepsilon)C_j) = ((1-\varepsilon)l_j)^d = (1-\varepsilon)^d Vol(C_j)$$

QED

Now notice that $(1-\varepsilon)^d \leq e^{-\varepsilon d}$
thus for fixed ε, as d grows larger than $1/\varepsilon$ and then
$e^{-\varepsilon d}$ exponentially decreases after that.

Consequence:

- For unit ball B subset $\mathbb{R}^d$:
  $1 - e^{-\varepsilon d}$ fraction of volume in $B \quad (1-\varepsilon)B$.

- For eps = 1/10, and d=100 then
  $1 - e^{-\varepsilon d} = 1 - e^{-10} =$ 0.99995
  within the last 10% of radius
- For eps = 1/20, and d=100 then
  $1 - e^{-\varepsilon d} = 1 - e^{-5} =$ 0.993
  within the last 5% of radius
- For eps = 1/25, and d=100 then
  $1 - e^{-\varepsilon d} = 1 - e^{-4} =$ 0.98
  within the last 4% of radius
- For eps = 1/50, and d=100 then
  $1 - e^{-\varepsilon d} = 1 - e^{-2} =$ 0.86
  within the last 2% of radius

4. Volume near Equator

Consider unit ball B subset $\mathbb{R}^d$
Let v = (1,0,0, ..., 0) – think of this as pointing "up" or "North"

Consider the "tropical zone" as being near the equator if the first coordinate has magnitude at most $c/\sqrt{d}$ for some $c \geq 1$ (think of c=10) and d=100. We show that

$$Vol_d(B_r) = \frac{r^d \pi^{d/2}}{\Gamma(d/2)} = r^d V_d$$

The "disk" at $1/\sqrt{d}$ above equator is a $(d-1)$-dimensional Ball with radius
$x = \sqrt{1 - 1/d}$ since $1^2 = x^2 + 1/d$
So $Vol_{d-1}(B_{\sqrt{1-1/d}}) = (1 - 1/d)^{d/2} V_{d-1}$

Also "disk" at $2/\sqrt{d}$ above equator is a $(d-1)$-dimensional Ball with radius
$\sqrt{1 - 4/d}$
So $Vol_{d-1)}(B_{\sqrt{1-4/d}}) = (1 - 4/d)^{d/2} V_{d-1}$

$$\frac{Vol_{d-1}(B_{\sqrt{1-1/d}})}{Vol_{d-1}(B_{\sqrt{1-4/d}})} = \frac{(1-1/d)^{d/2} V_{d-1}}{(1-4/d)^{d/2} V_{d-1}} \approx \frac{e^{-1/2}}{e^{-2}}$$

Each are "layers of a cake" with same height $1/\sqrt{d}$.
Volume decreases, geometrically so layers $j = 3 ... \sqrt{d} <$ layer 2

$$\sum_{j=2}^{\sqrt{d}} Vol_{d-1}(B_{\sqrt{1-j^2/d}}) < 2Vol_{d-1}(B_{\sqrt{1-4/d}})$$

So $e^{-1/2} > 2/e^2$ ... the lowest layer is larger than all above layers combined (for large d)

If we make the first layer $\kappa/\sqrt{d}$ for $\kappa > 1$, then the gap is even larger.

5. Spiky Boxes

Now consider a ball B with radius 1 in $\mathbb{R}^d$.
And a box $C = [-1/2, 1/2]^d$ with volume $Vol(C) = 1$

Note $(1 - 1/2)B \subset C$, where $(1 - 1/2)B$ is ball or radius $1/2$.

For $d = 2$, we have $C \subset B$.

For $d = 4$, still $C \subset B$
but $\|0 - (1/2, 1/2, 1/2, 1/2)\| = \sqrt{4(1/2)^2} = 1$
so the corner of C touches now touches boundary of B.

How about $d = 5$?
Corner of C outside of B.

How about $d = 8$?
Corner $c \in C$ has distance $\sqrt{8(1/2)^2} = 2$, far outside of B
center of face $(1/2, 0, ..., 0)$, well inside of B.

In general, corner a distance $\sqrt{d}/2$ from 0

Most of volume of the C is outside of B, since $Vol(B) \to 0$ as d grows

Question:

How do you sample a random point in B in $\mathbb{R}^d$ for large d?