COL866: Foundations of Data Science

Ragesh Jaiswal, IITD

Claim

For any unit length vector $\mathbf{v} \in \mathbb{R}^d$ defining "north", most of the volume of the unit ball lies in the thin slab containing points whose dot product with $\mathbf{v}$ is $O(1/\sqrt{d})$ (that is, the dot product is close to 0).

Argument

- Let $\mathbf{v}$ be the first coordinate vector. That is, $\mathbf{v} = (1, 0, 0, ..., 0)$.
- We will argue that most of the volume of the unit ball has $|x_1| = O(1/\sqrt{d})$.
- Theorem: For any $c \geq 1$ and $d \geq 3$, at least a $(1 - \frac{2}{c}e^{-c^2/2})$ fraction of the volume of the d-dimensional unit ball has $|x_1| \leq \frac{c}{\sqrt{d-1}}$.

Claim

Most of the volume of a unit ball in $\mathbb{R}^d$ is contained in an annulus of width $O(1/d)$ near the boundary.

Claim

For any unit length vector $\mathbf{v} \in \mathbb{R}^d$ defining "north", most of the volume of the unit ball lies in the thin slab containing points whose dot product with $\mathbf{v}$ is $O(1/\sqrt{d})$ (that is, the dot product is close to 0).

Claim

If we draw two random points from the unit ball, then with high probability their vectors will be nearly orthogonal to each other.

Claim

Most of the volume of a unit ball in $\mathbb{R}^d$ is contained in an annulus of width $O(1/d)$ near the boundary.

Claim

For any unit length vector $\mathbf{v} \in \mathbb{R}^d$ defining "north", most of the volume of the unit ball lies in the thin slab containing points whose dot product with $\mathbf{v}$ is $O(1/\sqrt{d})$ (that is, the dot product is close to 0).

Claim

If we draw two random points from the unit ball, then with high probability their vectors will be nearly orthogonal to each other.

Argument

- Both have length $1 - O(1/d)$ (whp).
- The dot product of these vectors are $\pm O(1/\sqrt{d})$ (whp).
- So, the angle between them is close to $\pi/2$ (whp).

Claim

If we draw two random points from the unit ball, then with high probability their vectors will be nearly orthogonal to each other.

Argument

- Both have length $1 - O(1/d)$ (whp).
- The dot product of these vectors are $\pm O(1/\sqrt{d})$ (whp).
- So, the angle between them is close to $\pi/2$ (whp).

Theorem

Consider drawing n points $\mathbf{x_1}, ..., \mathbf{x_n}$ at random from the unit ball. The following holds with probability $1 - O(1/n)$.

1. $||\mathbf{x_i}|| \geq 1 - \frac{2 \ln n}{d}$ *for all i, and*
2. $|\langle \mathbf{x_i}, \mathbf{x_j} \rangle| \leq \frac{\sqrt{6 \ln n}}{\sqrt{d-1}}$ *for all $i \neq j$.*

Claim

The volume of a unit ball in $\mathbb{R}^d$ goes to 0 as d goes to infinity.

Argument

- Consider a box of side $\frac{2c}{\sqrt{d-1}}$ for $c = 2\sqrt{\ln d}$ centered around the origin.

Claim

The volume of a unit ball in $\mathbb{R}^d$ goes to 0 as d goes to infinity.

Argument

- Consider a box of side $\frac{2c}{\sqrt{d-1}}$ for $c = 2\sqrt{\ln d}$ centered around the origin.
- The fraction of volume of the unit ball with $|x_1| \geq \frac{c}{\sqrt{d-1}}$ is at most $\frac{2}{c}e^{-c^2/2} = \frac{1}{d^2\sqrt{\ln d}} < \frac{1}{d^2}$.
- So, the ratio of volume of box to the volume of a unit ball is at least $1/2$.
- The volume of the box goes to 0 as d goes to infinity since the volume is $\left(4\sqrt{\frac{\ln d}{d-1}}\right)^d$.
- So, volume of the unit cube goes to 0 as $d \to \infty$.

Generating a random point from a unit ball

Question

How do we generate a random point from a unit ball in $\mathbb{R}^d$?

- Idea 1: Pick $x_1, ..., x_d$ randomly from the interval $[-1, +1]$. If $\mathbf{x} = (x_1, ..., x_d)$ is inside the unit ball, then output $\mathbf{x}$, else repeat.
  - When d is small (say $d = 2, 3$), then this idea indeed works. Does it work for large values of d?

Question

How do we generate a random point from a unit ball in $\mathbb{R}^d$?

- Idea 1: Pick $x_1, ..., x_d$ randomly from the interval $[-1, +1]$. If $\mathbf{x} = (x_1, ..., x_d)$ is inside the unit ball, then output $\mathbf{x}$, else repeat.
  - When d is small (say $d = 2, 3$), then this idea indeed works. Does it work for large values of d?
- Idea 2: Randomly sample $x_1, ..., x_d$ independently from a zero mean and unit variance Gaussian (i.e., with pdf $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$). Normalize the vector $\mathbf{x} = (x_1, ..., x_d)$ to a unit vector (i.e., output $\frac{\mathbf{x}}{||\mathbf{x}||}$).
  - From spherical symmetry, the output point is a random point on the surface of the unit ball.
  - The pdf of $\mathbf{x} = (x_1, ..., x_d)$ is given by $\frac{1}{(2\pi)^{d/2}} \cdot e^{-\frac{x_1^2 + ... + x_d^2}{2}}$.

Question

How do we generate a random point from a unit ball in $\mathbb{R}^d$?

- Idea 2: Randomly sample $x_1, ..., x_d$ independently from a zero mean and unit variance Gaussian (i.e., with pdf $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$). Normalize the vector $\mathbf{x} = (x_1, ..., x_d)$ to a unit vector (i.e., output $\frac{\mathbf{x}}{||\mathbf{x}||}$).
  - From spherical symmetry, the output point is a random point on the surface of the unit ball.
  - The pdf of $\mathbf{x} = (x_1, ..., x_d)$ is given by $\frac{1}{(2\pi)^{d/2}} \cdot e^{-\frac{x_1^2+...+x_d^2}{2}}$.

Question

How do we sample a random point x from a zero mean and unit variance Gaussian?

Question

How do we sample a random point x from a zero mean and unit variance Gaussian?

- More general question: How do we sample a point x given its cumulative distribution function (cdf) $C(x)$? We assume that we can sample from a uniform distribution in the interval $[0, 1]$.
- Answer: Sample a uniform random number $u \in [0, 1]$ and output $x = C^{-1}(u)$.
- Since we do not have a closed form expression for the cdf of a Gaussian distribution, the above idea does not help in our case in a straightforward manner. However, we can use numerical approximations.

Question

How do we sample a random point x from a zero mean and unit variance Gaussian?

- More general question: How do we sample a point x given its cumulative distribution function (cdf) $C(x)$? We assume that we can sample from a uniform distribution in the interval $[0, 1]$.
- Answer: Sample a uniform random number $u \in [0, 1]$ and output $x = C^{-1}(u)$.
- Since we do not have a closed form expression for the cdf of a Gaussian distribution, the above idea does not help in our case in a straightforward manner. However, we can use numerical approximations.
- Another method is called the Box-Muller transform: Let U_1, U_2 denote uniform random numbers in $[0, 1]$. Then

$$X_1 = \sqrt{-2 \ln U_1} \cdot \cos(2\pi U_2) \quad \text{and} \quad X_2 = \sqrt{-2 \ln U_1} \cdot \sin(2\pi U_2)$$

are independent samples from zero mean and unit variance Gaussian.

Question

How do we generate a random point from a unit ball (surface and interior) in $\mathbb{R}^d$?

- Idea: Randomly sample $x_1, ..., x_d$ from zero mean and unit variance Gaussian and scale the vector $\frac{\mathbf{x}}{||\mathbf{x}||}$ on the surface of the unit ball by a scalar $\rho \in [0, 1]$. Here $\mathbf{x} = (x_1, ..., x_d)$.
- Question: Do we pick ρ from a uniform distribution over $[0, 1]$?

Question

How do we generate a random point from a unit ball (surface and interior) in $\mathbb{R}^d$?

- Idea: Randomly sample $x_1, ..., x_d$ from zero mean and unit variance Gaussian and scale the vector $\frac{\mathbf{x}}{||\mathbf{x}||}$ on the surface of the unit ball by a scalar $\rho \in [0, 1]$. Here $\mathbf{x} = (x_1, ..., x_d)$.
- Question: Do we pick ρ from a uniform distribution over $[0, 1]$? No
- The density of points at radius r is proportional to r^{d-1}.
- So, we should pick $\rho(r)$ with density dr^{d-1}.

Gaussians in High Dimension

- A one dimensional Gaussian has much of its probability mass close to the origin.
- Does this generalise to higher dimensions?
- A d-dimensional spherical Gaussian with 0 means and σ^2 variance in each coordinate has density:

$$p(\mathbf{x}) = \frac{1}{\sigma^d (2\pi)^{d/2}} e^{-\frac{||\mathbf{x}||^2}{2\sigma^2}}$$

- Let $\sigma^2 = 1$. Even though the probability density is high within the unit ball, the volume of of the unit ball is negligible and hence the probability mass within the unit ball is negligible.
- When the radius is $\sqrt{d}$, the volume becomes large enough to make the probability mass around the $\sqrt{d}$ radius significant.
- Even though the volume keeps increasing beyond the $\sqrt{d}$ radius, the probability density keeps diminishing. So, the probability mass much beyond the $\sqrt{d}$ radius is again negligible.

- Even though the probability density is high within the unit ball, the volume of of the unit ball is negligible and hence the probability mass within the unit ball is negligible.
- When the radius is $\sqrt{d}$, the volume becomes large enough to make the probability mass around the $\sqrt{d}$ radius significant.
- Even though the volume keeps increasing beyond the $\sqrt{d}$ radius, the probability density keeps diminishing. So, the probability mass much beyond the $\sqrt{d}$ radius is again negligible.
- This intuition is formalised in the next theorem.

Theorem (Gaussian Annulus Theorem)

For a d-dimensional spherical Gaussian with unit variance in each direction, for any $\beta \leq \sqrt{d}$, all but at most $3e^{-c\beta^2}$ of the probability mass lies within the annulus $\sqrt{d} - \beta \leq ||\mathbf{x}|| \leq \sqrt{d} + \beta$, where c is a fixed positive constant.

End