COL866: Foundations of Data Science

Ragesh Jaiswal, IITD

Theorem (Law of large numbers)

Let $x_1, x_2, ..., x_n$ be n independent samples of a random variable x. Then

$$\mathbf{Pr}\left[\left|\frac{x_1 + x_2 + ... + x_n}{n} - \mathbf{E}(x)\right| \geq \varepsilon\right] \leq \frac{\mathbf{Var}(x)}{n\varepsilon^2}.$$

- The above theorem gives a sense of how concentrated the sum of independent random variables is around the mean value.
- Such tail bounds are extremely useful in randomised analysis.
- Here is a general theorem for sum of independent random variables.

Theorem (Master tail bounds theorem)

Let $x = x_1 + ... + x_n$, where $x_1, ..., x_n$ are mutually *independent random variables with zero mean and variance at most σ^2. Let $0 \leq a \leq \sqrt{2}n\sigma^2$. Assume that $|\mathbf{E}(x_i^s)| \leq \sigma^2(s!)$ for $s = 3, 4, ..., \lfloor \frac{a^2}{4n\sigma^2} \rfloor$. Then*

$$\mathbf{Pr}(|x| \geq a) \leq 3e^{-\frac{a^2}{12n\sigma^2}}.$$

Theorem (Law of large numbers)

Let $x_1, x_2, ..., x_n$ be n independent samples of a random variable x. Then

$$\mathbf{Pr}\left[\left|\frac{x_1 + x_2 + ... + x_n}{n} - \mathbf{E}(x)\right| \geq \varepsilon\right] \leq \frac{\mathbf{Var}(x)}{n\varepsilon^2}.$$

- Let us try to use the above theorem to get answers to the initial questions the were raised w.r.t. high dimensional spaces.
  - The volume of a unit ball goes to zero as dimension goes to infinity.
  - The volume of a unit ball is concentrated near its *surface* and is also concentrated at its *equator*.
  - If one generates a random point in d-dimensional space using a Gaussian to generate coordinates independently, the distance between all pair of points will *mostly* be the same when d is large.

Claim

The volume of a unit ball goes to zero as dimension goes to infinity.

Argument

- Let x denote a gaussian random variable with zero mean and variance $1/2\pi$.
- Let $\mathbf{z}$ denote a d-dimensional random point sampled by taking d independent copies of x in each coordinate.
- Claim 1: The gaussian probability density is bounded below by some constant throughout the unit ball.

Claim

The volume of a unit ball goes to zero as dimension goes to infinity.

Argument

- Let x denote a gaussian random variable with zero mean and variance $1/2\pi$.
- Let $\mathbf{z}$ denote a d-dimensional random point sampled by taking d independent copies of x in each coordinate.
- Claim 1: The gaussian probability density is bounded below by some constant throughout the unit ball.
- Claim 2: With high probability $||\mathbf{z}||^2 = \Theta(d)$.

Claim

The volume of a unit ball goes to zero as dimension goes to infinity.

Argument

- Let x denote a gaussian random variable with zero mean and variance $1/2\pi$.
- Let $\mathbf{z}$ denote a d-dimensional random point sampled by taking d independent copies of x in each coordinate.
- Claim 1: The gaussian probability density is bounded below by some constant throughout the unit ball.
- Claim 2: With high probability $||\mathbf{z}||^2 = \Theta(d)$.
- So, as d goes to infinity, the probability that $\mathbf{z}$ is in the unit ball goes to 0 (from the Law of large numbers).
- This implies that the integral of the probability density function within the unit ball goes to 0 as d goes to infinity.

Claim

The volume of a unit ball goes to zero as dimension goes to infinity.

Argument

- Let x denote a gaussian random variable with zero mean and variance $1/2\pi$.
- Let $\mathbf{z}$ denote a d-dimensional random point sampled by taking d independent copies of x in each coordinate.
- Claim 1: The gaussian probability density is bounded below by some constant throughout the unit ball.
- Claim 2: With high probability $||\mathbf{z}||^2 = \Theta(d)$.
- So, as d goes to infinity, the probability that $\mathbf{z}$ is in the unit ball goes to 0 (from the Law of large numbers).
- This implies that the integral of the probability density function within the unit ball goes to 0 as d goes to infinity.
- From claim 1, this implies that the volume of the unit ball goes to 0 as d goes to infinity.

Claim

If one generates a random point in d-dimensional space using a Gaussian to generate coordinates independently, the distance between all pair of points will *mostly* be the same when d is large.

Argument

- Consider points $\mathbf{y} = (y_1, ..., y_d)$ and $\mathbf{z} = (z_1, ..., z_d)$ constructed by sampling y_i's and z_i's independently from a zero mean and unit variance gaussian.
- Claim 1: $\mathbf{E}[(y_i - z_i)^2] = 2$.
- Claim 2: $||\mathbf{y} - \mathbf{z}||^2 \approx 2d$ with high probability.

Claim

The volume of a unit ball is concentrated at its *equator*.

Argument

- Consider points $\mathbf{y} = (y_1, ..., y_d)$ and $\mathbf{z} = (z_1, ..., z_d)$ constructed by sampling y_i's and z_i's independently from a zero mean and unit variance gaussian.
- Claim 1: $\mathbf{E}[(y_i - z_i)^2] = 2$.
- Claim 2: $||\mathbf{y} - \mathbf{z}||^2 \approx 2d$ with high probability.
- Claim 3: $||\mathbf{y}||^2 \approx d$ and $||\mathbf{z}||^2 \approx d$ with high probability.
- So, $\mathbf{y}$ and $\mathbf{z}$ are approximately orthogonal.
- Scaling these points to be unit length and calling (scaled) $\mathbf{y}$ as the "north pole", we see that much of the surface area of the unit ball must lie near the equator.

High Dimensional Geometry

Claim

Most of the volume of any high dimensional object is near its surface.

Argument

- Consider any object $A \in \mathbb{R}^d$ and its "shrinked" version $\langle 1-\varepsilon \rangle A = \{(1-\varepsilon)x | x \in A\}$.
- Claim 1: $Volume(\langle 1-\varepsilon \rangle A) = (1-\varepsilon)^d \cdot Volume(A)$.

Claim

Most of the volume of any high dimensional object is near its surface.

Argument

- Consider any object $A \in \mathbb{R}^d$ and its "shrinked" version $\langle 1-\varepsilon \rangle A = \{(1-\varepsilon)x | x \in A\}$.
- Claim 1: $Volume(\langle 1-\varepsilon \rangle A) = (1-\varepsilon)^d \cdot Volume(A)$.
  - Partition A into infinitesimal cubes, then $\langle 1-\varepsilon \rangle A$ is the union of the cubes shrinked by a factor of $(1-\varepsilon)$.

Corollary

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

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

Theorem (Volume and surface area of unit ball)

The surface area $A(d)$ and the volume $V(d)$ of a unit ball in $\mathbb{R}^d$ is given by:

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

The Γ function (analogous to factorial) is defined recursively as $\Gamma(x) = (x-1) \cdot \Gamma(x-1), \Gamma(1) = \Gamma(2) = 1$, and $\Gamma(1/2) = \sqrt{\pi}$.

Claim

Most of the volume of a unit ball in $\mathbb{R}^d$ is concentrated near its "equator".

Claim

Most of the volume of a unit ball in $\mathbb{R}^d$ is concentrated near its "equator".

Claim rephrased

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

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 $|x_1| \leq \frac{c}{\sqrt{d-1}}$.

End