KAIST AI
Graduate School of AI

Lecture 8: High-dimensional Space (Chapter 2 of Textbook B)

Jinwoo Shin

AI503: Mathematics for AI

Introduction

KAIST AI
Graduate School of AI

- High dimensional data are extensively used in machine learning.
  - e.g. Images, videos
- However, high dimensional space is very different from 2D and 3D.
  - Volume of the unit ball in d-dimensions goes to zero as dimension goes to infinity.
  - Volume of a high dimensional unit ball is concentrated near its surface and also its equator.

Roadmap

Roadmap

KAIST AI
Graduate School of AI

Markov's Inequality

Lemma. Markov's inequality

Let x be a nonnegative random variable. Then for $a > 0$,

$$Prob(x \geq a) \leq \frac{E(x)}{a}$$

Proof. For a continuous nonnegative random variable x with probability density p,

$$E(x) = \int_0^\infty xp(x)dx = \int_0^a xp(x)dx + \int_a^\infty xp(x)dx \geq \int_a^\infty xp(x)dx \geq a\int_a^\infty p(x)dx = a \cdot \text{Prob}(x \geq a)$$

Chebyshev's Inequality

Lemma. Chebyshev's inequality

Let x be a random variable. Then for $c > 0$,

$$Prob(|x - E(x)| \geq c) \leq \frac{Var(x)}{c^2}$$

Proof. Since $|x - E(x)|^2$ is nonnegative random variable and $E(|x - E(x)|^2) = Var(x)$, we can apply Markov's inequality:

$$Prob(|x - E(x)| \geq c) = Prob(|x - E(x)|^2 \geq c^2) \leq \frac{E(|x - E(x)|^2}{c^2} = \frac{Var(x)}{c^2}$$

The Law of Large Numbers

Theorem. The Law of Large Numbers

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

$$Prob\left(\left|\frac{x_1 + x_2 + \ldots + x_n}{n} - E(x)\right| \geq \varepsilon\right) \leq \frac{Var(x)}{n\varepsilon^2}$$

Proof. By Chebychev's inequality,

$$\begin{aligned} Prob\left(\left|\frac{x_1 + x_2 + \cdots + x_n}{n} - E(x)\right| \geq \varepsilon\right) &\leq \frac{Var(\frac{x_1+x_2+\cdots+x_n}{n})}{\varepsilon^2} \\ &= \frac{1}{n^2\varepsilon^2} Var(x_1 + x_2 + \cdots + x_n) \\ &= \frac{Var(x)}{n\varepsilon^2} \end{aligned}$$

Remark. Average of the samples will be close to the expectation of the random variable.

Roadmap

The Geometry of High Dimensions

KAIST AI
Graduate School of AI

- High-dimensional objects have most of their volume in near the surface.
  - Let A denote any object in d dimensions.
  - Shrink A by a small amount ε to produce a new object $(1-\varepsilon)A = \{(1-\varepsilon)x | x \in A\}$.

$$\frac{\text{volume}((1-\varepsilon)A)}{\text{volume}(A)} = (1-\varepsilon)^d \leq e^{-\varepsilon d}$$

  - The above ratio rapidly approaches zero as $d \to \infty$.
  - Nearly all of the volume of A must be in $A \setminus (1-\varepsilon)A$.

Roadmap

KAIST AI
Graduate School of AI

Volume Near the Surface

KAIST AI
Graduate School of AI

Theorem. Volume near the surface

Consider n points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ drawn at random from the unit ball. With probability $1 - O(\frac{1}{n})$,

$$|\mathbf{x}_i| \geq 1 - \frac{2 \ln n}{d}, \quad \text{for all } i$$

Proof. From the analysis of the previous section, the probability that $|\mathbf{x}_i| < 1 - \varepsilon$ is less than $e^{-\varepsilon d}$.

$$\text{Prob}\left(|\mathbf{x}_i| < 1 - \frac{2 \ln n}{d}\right) \leq e^{-(\frac{2 \ln n}{d})d} = \frac{1}{n^2}$$

By the union bound, the probability there exists an i such that $\left(|\mathbf{x}_i| < 1 - \frac{2 \ln n}{d}\right)$ is at most $1/n$.

Volume of the Unit Ball

KAIST AI
Graduate School of AI

Theorem. Surface and volume of unit ball in d-dimension

The surface area $A(d)$ and the volume $V(d)$ of a unit-radius ball in d-dimensions are given by

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

where $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\,dt$. Note that $\Gamma(n) = (n-1)!$ for integer n.

Proof. See Page 17-18 of Textbook B.

Volume of the Unit Ball

Theorem. Surface and volume of unit ball in d-dimension

The surface area $A(d)$ and the volume $V(d)$ of a unit-radius ball in d-dimensions are given by

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

where $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\,dt$. Note that $\Gamma(n) = (n-1)!$ for integer n.

- e.g. $A(2) = 2\pi$, $A(3) = 4\pi$ and $V(2) = \pi$, $V(3) = \frac{4}{3}\pi$
- Remark. As d goes to infinity, the volume of the ball goes to zero.
  - $\Gamma(\frac{d}{2})$ grows as the factorial of $\frac{d}{2}$ while $\pi^{\frac{d}{2}}$ increases exponentially.

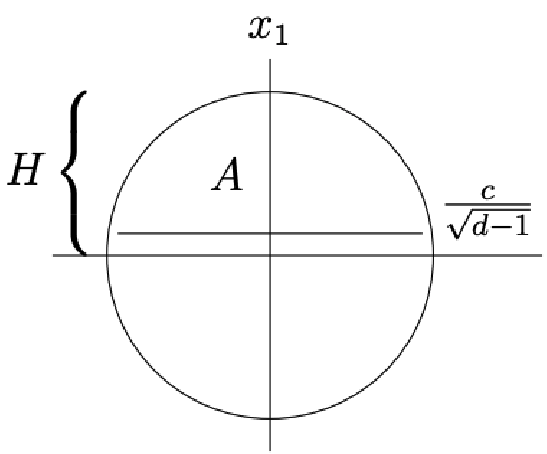
Volume Near the Equator

Most of volume of the unit ball is concentrated near its equator.

Theorem. Volume of unit ball near the equator

For $c \geq 1$ and $d \geq 3$, at least a $1 - \frac{2}{c} e^{-\frac{c^2}{2}}$ fraction of the volume of the d-dimensional unit ball has $|x_1| \leq \frac{c}{\sqrt{d-1}}$.

Proof. See Page 19-20 of Textbook B.



Near Orthogonality

If we choose two random point from unit ball, their vectors are nearly orthogonal with high probability.

- From previous analysis that most of volume is in near the surface, both vectors will be close to the surface and have length $1 - O(\frac{1}{d})$.
- Without loss of generality, let the first vector points 'north'.
- Then, the second vector have a projection of only $\pm O(\frac{1}{\sqrt{d}})$ with high probability.
- Thus, their dot product will be $\pm O(\frac{1}{\sqrt{d}})$.

Near Orthogonality

Theorem. Near Orthogonality

Consider n points $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{\boldsymbol{n}}$ at random from the unit ball. With probability $1 - O(\frac{1}{n})$,

$$|x_i \cdot x_j| \leq \frac{\sqrt{6 \ln n}}{\sqrt{d-1}} \quad \text{for all } i \neq j$$

Proof. See Page 21 of Textbook B.

Roadmap

KAIST AI
Graduate School of AI

(1) Preliminary: The Law of Large Numbers

(2) The Geometry of High Dimensions

(3) Properties of the Unit Ball

(4) Gaussians in High Dimension

(5) Application: Random Projection and Johnson-Lindenstrauss Lemma

(6) Application: Separating Gaussians

Gaussian in High Dimension

Consider d-dimensional spherical Gaussian with zero mean and variance σ^2 in each dimension.

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}\sigma^d} exp\left(-\frac{|\boldsymbol{x}|^2}{2\sigma^2}\right)$$

Gaussian Annulus Theorem

KAIST AI
Graduate School of AI

Theorem. Gaussian Annulus Theorem

For a d-dimensional spherical Gaussian with unit variance in each direction, for any $\beta \leq \sqrt{d}$, more than $1 - 3e^{-c\beta^2}$ of the probability mass lies within the annulus $\sqrt{d} - \beta \leq |x| \leq \sqrt{d} + \beta$, where c is a fixed positive constant.

Proof. See Page 24-25 of Textbook B.

Gaussian Annulus Theorem

KAIST AI
Graduate School of AI

Theorem. Gaussian Annulus Theorem

For a d-dimensional spherical Gaussian with unit variance in each direction, for any $\beta \leq \sqrt{d}$, more than $1 - 3e^{-c\beta^2}$ of the probability mass lies within the annulus $\sqrt{d} - \beta \leq |x| \leq \sqrt{d} + \beta$, where c is a fixed positive constant.

- Nearly all the probability of spherical Gaussian is concentrated in a annulus of width $O(1)$ at radius $\sqrt{d}$.
- High-level intuition:
  - $E(|x|^2) = \sum_{i=1}^{d} E(x_i^2) = dE(x_1^2) = d$.
  - Distance of a point from the center is $\sqrt{d}$ with high probability by the law of large number.

Roadmap

(1) Preliminary: The Law of Large Numbers

(2) The Geometry of High Dimensions

(3) Properties of the Unit Ball

(4) Gaussians in High Dimension

(5) Application: Random Projection and Johnson-Lindenstrauss Lemma

(6) Application: Separating Gaussians

Random Projection and Johnson-Lindenstrauss Lemma

- Nearest neighbor search is a broadly used subroutine in tasks involving high-dimensional data.
  - We are presented to "query" points in R^d to find the nearest point to the query point.
  - Since the number of queries is often large, the time for each query should be very small.
- For this kind of problem, dimension reduction which (approximately) preserves relative distances can be very useful.
- We will show simple (approximately) distance-preserving projection exists using Gaussian Annulus Theorem.

Random Projection and Johnson-Lindenstrauss Lemma

The projection $f : R^d \rightarrow R^k$ is as follows.

- Pick k Gaussian vectors $u_1, u_2, ..., u_k$ in R^d with unit variance.
- For any vector v, the projection $f(v)$ is
$$f(v) = (u_1 \cdot v, u_2 \cdot v, ..., u_k \cdot v)$$
- Remark. $f(v_1 - v_2) = f(v_1) - f(v_2)$ for every v_1 and v_2.

We will show that $|f(v)| \approx \sqrt{k}|v|$.

- Then, $|v_1 - v_2| \approx \frac{1}{\sqrt{k}}|f(v_1 - v_2)| = \frac{1}{\sqrt{k}}|f(v_1) - f(v_2)|$.

The projection f preserves the relative distance with factor $\sqrt{k}$

The Random Projection Theorem

Theorem. The Random Projection Theorem

Let v be a fixed vector in R^d and let f be defined as above. There exists constant $c > 0$ such that for $\varepsilon \in (0, 1)$,

$$Prob\left(\left||f(v)| - \sqrt{k}|v|\right| \geq \varepsilon\sqrt{k}|v|\right) \leq 3e^{-ck\varepsilon^2}$$

where the probability is taken over the random draws of vectors u_i used to construct f.

Proof. We may assume that $|v| = 1$ by scaling the inner equality by $|v|$.
Since $u_i \cdot v = \sum_{j=1}^{d} u_{ij} v_j$, the random variable $u_i \cdot v$ has Gaussian density with zero mean and unit variance.

$$Var(u_i \cdot v) = Var\left(\sum_{j=1}^{d} u_{ij} v_j\right) = \sum_{j=1}^{d} v_j^2 Var(u_{ij}) = \sum_{j=1}^{d} v_j^2 = 1$$

Since $u_1 \cdot v, u_2 \cdot v, ..., u_k \cdot v$ are independent Gaussian random variables, $f(v)$ is a random vector from a k-dimensional spherical Gaussian with unit variance.
Therefore, the theorem follows from the Gaussian Annulus Theorem.

Theorem. The Random Projection Theorem

Let v be a fixed vector in R^d and let f be defined as above. There exists constant $c > 0$ such that for $\varepsilon \in (0, 1)$,

$$Prob\left(\left||f(v)| - \sqrt{k}|v|\right| \geq \varepsilon\sqrt{k}|v|\right) \leq 3e^{-ck\varepsilon^2}$$

where the probability is taken over the random draws of vectors u_i used to construct f.

The random projection theorem establishes that the probability of the length of the projection of a single vector differing significantly from its expected value is exponentially small in k, the dimension of the target subspace.

Theorem. Johnson-Lindenstrauss Lemma

For any $0 < \varepsilon < 1$ and any integer n, let $k \geq \frac{3}{c\varepsilon^2} \ln n$ with c as in Gaussian Annulus Theorem. For any set of n points in R^d, the random projection $f : R^d \to R^k$ defined above has the property that for all pairs of points v_i and v_j, with probability at least $1 - \frac{3}{2n}$,

$$(1-\varepsilon)\sqrt{k}|v_i - v_j| \leq |f(v_i) - f(v_j)| \leq (1+\varepsilon)\sqrt{k}|v_i - v_j|$$

where the probability is taken over the random draws of vectors u_i used to construct f.

Proof. Direct consequence of the random projection theorem.

Roadmap

KAIST AI
Graduate School of AI

Separating Gaussians

Mixtures of Gaussians are often used to model heterogeneous data coming from multiple sources.

- e.g. Heights record from men and women group.

We will show the algorithm which separates samples of Gaussian mixture to each Gaussian samples with high probability.

- Specifically, we will focus on a mixture of two spherical unit-variance Gaussians whose means are separated by a distance $\Omega(d^{1/4})$.

Distance between Samples of One Spherical Gaussian

Consider one spherical unit variance Gaussian centered at the origin.

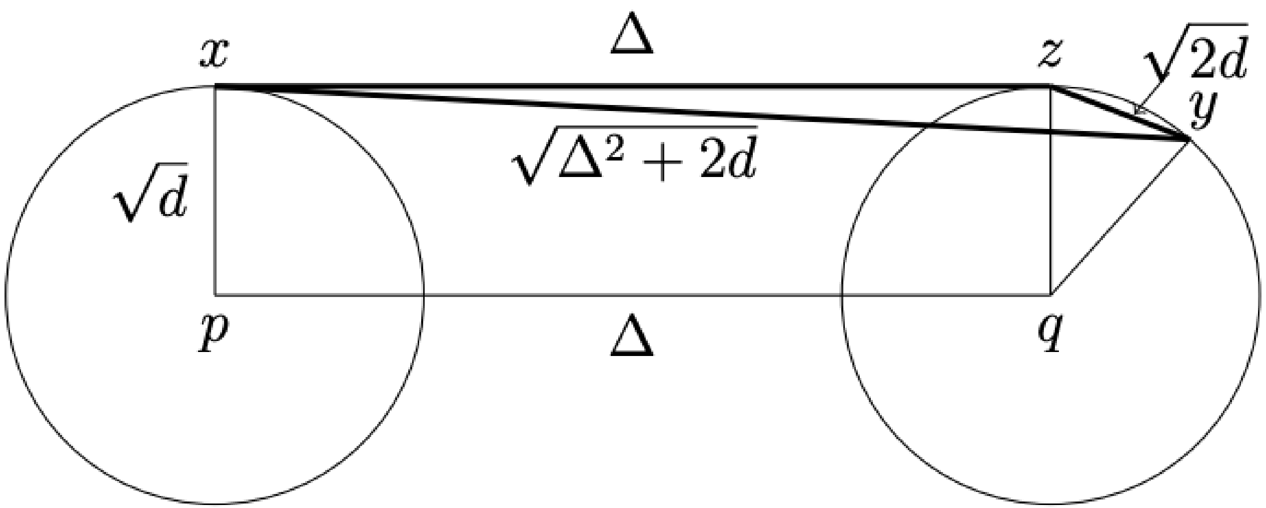
- Pick a point $\boldsymbol{x}$ from Gaussian. Without loss of generality, let the first axis align with $\boldsymbol{x}$.
- $\boldsymbol{x}$ lies on an annulus of width $O(1)$ at radius $\sqrt{d}$ with high probability, by Gaussian Annulus Theorem; $\boldsymbol{x} = (\sqrt{d} \pm O(1), 0, \ldots, 0)$.
- Independently choose $\boldsymbol{y}$ from the same Gaussian. Since each coordinate of $\boldsymbol{y}$ is from unit variance Gaussian, $\boldsymbol{y}$ is in $\{\boldsymbol{x} | -c \leq x_1 \leq c\}$ with high probability.
- Rotate the coordinate so that the component of $\boldsymbol{y}$ which is perpendicular to the first axis points the second coordinate; $\boldsymbol{y} = (O(1), \sqrt{d} \pm O(1), 0, \ldots, 0)$.
- Thus,
$$(\boldsymbol{x} - \boldsymbol{y})^2 = d \pm O(\sqrt{d}) + d \pm O(\sqrt{d}) = 2d + O(\sqrt{d})$$
- $|\boldsymbol{x} - \boldsymbol{y}| = \sqrt{2d} \pm O(1)$ with high probability.

Distance between Samples of Two Spherical Gaussian

Consider two spherical unit variance Gaussian with centers p and q separated by a distance Δ.

The distance between x and y, which are randomly chosen from each Gaussian, respectively, is close to $\sqrt{\Delta^2 + 2d}$.

$$|x - y|^2 \approx \Delta^2 + |z - q|^2 + |q - y|^2 = \Delta^2 + 2d \pm O(\sqrt{d})$$



Separating Gaussians

To separate samples from two Gaussians, the upper bound of the distance between a pair of points from the same Gaussian should be at most the lower bound of distance between points from different Gaussians.

This requires $\sqrt{2d} + O(1) \leq \sqrt{\Delta^2 + 2d} - O(1)$ which holds when $\Delta \in \omega(d^{1/4})$.

Thus when center of Gaussians are separated by $\omega(d^{1/4})$, samples are separable.

Algorithm for separating points from two Gaussians

1. Calculate all pairwise distances between points.
2. The cluster of smallest pairwise distances must come from a single Gaussian. Remove these points.
3. The remaining points come from the second Gaussian.

Questions?