Curse of Dimensionality in Pivot-based Indexes

Ilya Volnyansky, Vladimir Pestov

Department of Mathematics and Statistics
University of Ottawa
Ottawa, Ontario, Canada

SISAP 2009, Prague, 29/09/2009

Outline

1. Overview
   - The Setting for Similarity Search
   - Previous Work
2. Our Work
   - Framework
   - Concentration of Measure
   - Statistical Learning Theory
   - Asymptotic Bounds

Similarity Workloads

- Universe Ω: metric space with metric ρ.
- Dataset $X \subset \Omega$, always finite, with metric ρ.
- A *range query*: given $q \in \Omega$ and $r > 0$ find $\{x \in X | \rho(x, q) < r\}$

For analysis purposes, we add:

- A measure μ on Ω.
- Treat X as i.i.d. sample $\sim \mu$ of size n

Similarity Workloads

- Universe Ω: metric space with metric ρ.
- Dataset $X \subset \Omega$, always finite, with metric ρ.
- A *range query*: given $q \in \Omega$ and $r > 0$ find $\{x \in X | \rho(x, q) < r\}$

For analysis purposes, we add:

- A measure μ on Ω.
- Treat X as i.i.d. sample $\sim \mu$ of size n

Curse of dimensionality conjecture

All indexing schemes suffer from the curse of dimensionality: (conjecture)

> If $d = \omega(\log n)$ and $d = n^{o(1)}$, any sequence of indexes built on a sequence of datasets $X_d \subset \Sigma_d$ allowing similarity search in time polynomial in d must use $n^{\omega(1)}$ space.

Handbook of Discrete and Computational Geometry

The Hamming cube Σ_d of dimension d: The set of all binary sequences of length d.

Curse of dimensionality conjecture

All indexing schemes suffer from the curse of dimensionality: (conjecture)

> If $d = \omega(\log n)$ and $d = n^{o(1)}$, any sequence of indexes built on a sequence of datasets $X_d \subset \Sigma_d$ allowing similarity search in time polynomial in d must use $n^{\omega(1)}$ space.

Handbook of Discrete and Computational Geometry

The Hamming cube Σ_d of dimension d: The set of all binary sequences of length d.

Fixed dimension

Examples of previous work:

Let n the size of X vary, but the space (Ω, ρ, μ) be fixed.

- The usual “asymptotic” analysis in the CS sense.
- Does not investigate the curse of dimensionality.

Fixed *n*

Let the dimension and hence (Ω, ρ, μ) vary but the size *n* of *X* stay the same.

- e.g. [Weber 98], [Chávez 01]
- Too small sample size *n* makes it easier to index spaces of high dimension *d*.
- When both *d* and *n* vary, the math is more challenging.

Points to keep in mind

- Distinction between X and Ω.
- Both d and n grow.
- Need to make assumptions about the sequence of Ω's
- (?) Need to make assumption about the indexes.

Gameplan

1. Pick an index type to analyze.
2. Pick a cost model.
3. The sequence of Ω's exhibits concentration of measure, the "intrinsic dimension" grows.
4. Statistical Learning Theory: linking properties of Ω's and properties of X's.
5. Conclusion: if all conditions are met, the Curse of Dimensionality will take place.

Main Result

From a sequence of metric spaces with measure $(\Omega_d, \rho_d, \mu_d)$, where $d = 1, 2, 3, \ldots$ take i.i.d. samples (datasets) $X_d \sim \mu_d$. Assume

- $(\Omega_d, \rho_d, \mu_d)$ display the concentration of measure.
- The VC dimension of closed balls in (Ω_d, ρ_d) is $O(d)$.
- We build a pivot-index using k pivots, where $k = o(n_d/d)$.
- Sample size n_d satisfies $d = \omega(\log n_d)$ and $d = n_d^{o(1)}$.

Suppose we perform queries of radius=NN. Then:
If we fix arbitrarily small $\varepsilon, \eta > 0$, $\exists D$ **such that for all** $d \geqslant D$, **the probability that** ***at least half*** **the queries on dataset** X_d **take less than** $(1 - \varepsilon)n_d$ **time is less than** η.

Main Result

From a sequence of metric spaces with measure $(\Omega_d, \rho_d, \mu_d)$, where $d = 1, 2, 3, \dots$ take i.i.d. samples (datasets) $X_d \sim \mu_d$. Assume

- $(\Omega_d, \rho_d, \mu_d)$ display the concentration of measure.
- The VC dimension of closed balls in (Ω_d, ρ_d) is $O(d)$.
- We build a pivot-index using k pivots, where $k = o(n_d/d)$.
- Sample size n_d satisfies $d = \omega(\log n_d)$ and $d = n_d^{o(1)}$.

Suppose we perform queries of radius=NN. Then:
If we fix arbitrarily small $\varepsilon, \eta > 0$, $\exists D$ **such that for all** $d \geqslant D$**, the probability that** ***at least half*** **the queries on dataset** X_d **take less than** $(1-\varepsilon)n_d$ **time is less than** η**.**

Pivot indexing scheme

Build an index:

1. Pick $\{p_1 \dots p_k\}$ from X
2. Calculate $n \times k$ array of distances

$$\rho(x, p_i),\ 1 \leqslant i \leqslant k,\ x \in X$$

Perform query given q and r :

1. Compute $\rho_K(q, x) := \sup_{1 \leqslant i \leqslant k} |\rho(q, p_i) - \rho(x, p_i)|$.
2. Since $\rho(q, x) \geqslant \rho_K(q, x)$, no need to compute $\rho(q, x)$ if $\rho_K(q, x) > r$
3. Compute $\rho(q, x)$ otherwise.

Pivot indexing scheme

Build an index:

1. Pick $\{p_1 \dots p_k\}$ from X
2. Calculate $n \times k$ array of distances

$$\rho(x, p_i),\ 1 \leqslant i \leqslant k,\ x \in X$$

Perform query given q and r :

1. Compute $\rho_k(q, x) := \sup_{1 \leqslant i \leqslant k} |\rho(q, p_i) - \rho(x, p_i)|$.
2. Since $\rho(q, x) \geqslant \rho_k(q, x)$, no need to compute $\rho(q, x)$ if $\rho_k(q, x) > r$
3. Compute $\rho(q, x)$ otherwise.

The cost model

- Only one cost: $\rho(q, x)$
- Computing $\rho_k(q, x)$ costs k.
- Let $C_{q,r,p_1,\dots,p_k}$ denote all the discarded points in X:

$$\{x \in X | \rho_k(q, x) > r\}$$

- Let $n = |X|$.
- Total cost: $k + n - |C_{q,r,p_1,\dots,p_k}|$.

The cost model

- Only one cost: $\rho(q, x)$
- Computing $\rho_k(q, x)$ costs k.
- Let $C_{q,r,p_1,\ldots,p_k}$ denote all the discarded points in X:

$$\{x \in X | \rho_k(q, x) > r\}$$

- Let $n = |X|$.
- Total cost: $k + n - |C_{q,r,p_1,\ldots,p_k}|$.

Concentration of Measure

A function $f : \Omega \to \mathbb{R}$ is 1-Lipschitz if

$$|f(\omega_1) - f(\omega_2)| \leqslant \rho(\omega_1, \omega_2)\ \forall \omega_1, \omega_2 \in \Omega$$

Examples:

- $f(x) = x$
- $f(x) = \frac{1}{2}x$
- $f(x) = \sqrt{(x^2 + 1)}$

Its median is a number M such that

$$\mu\{\omega | f(\omega) \leqslant M\} \geqslant 1/2 \text{ and } \mu\{\omega | f(\omega) \geqslant M\} \geqslant 1/2$$

Concentration of Measure

A function $f : \Omega \to \mathbb{R}$ is 1-Lipschitz if

$$|f(\omega_1) - f(\omega_2)| \leqslant \rho(\omega_1, \omega_2) \ \forall \omega_1, \omega_2 \in \Omega$$

Examples:

- $f(x) = x$
- $f(x) = \frac{1}{2}x$
- $f(x) = \sqrt{(x^2 + 1)}$

Its median is a number M such that

$$\mu\{\omega | f(\omega) \leqslant M\} \geqslant 1/2 \text{ and } \mu\{\omega | f(\omega) \geqslant M\} \geqslant 1/2$$

Concentration of Measure

A sequence of spaces $(\Omega_d, \rho_d, \mu_d)_{d=1}^{\infty}$ exhibits (normal) *concentration of measure* if there are $C, c > 0$ such that for every 1-Lipschitz function $f : \Omega \to \mathbb{R}$ with median M:

$$\forall \epsilon > 0, \quad \mu\{\omega | \ |f(\omega) - M| > \epsilon\} < Ce^{-c\epsilon^2 d}$$

Examples:

- The Spheres $\mathbb{S}^d$ in $\mathbb{R}^{d+1}$
- The Balls $\mathbb{B}^d$.
- The Hamming Cubes Σ^d.

Concentration of Measure

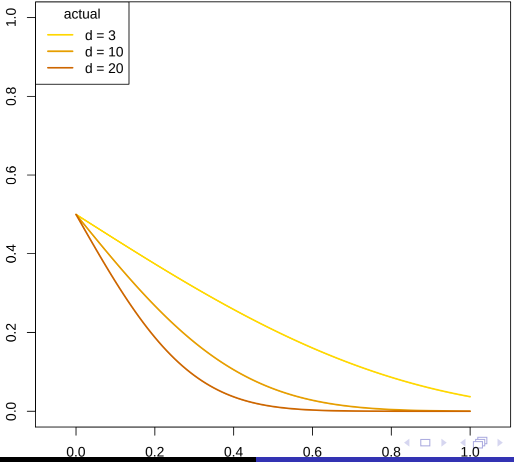
A sequence of spaces $(\Omega_d, \rho_d, \mu_d)_{d=1}^{\infty}$ exhibits (normal) *concentration of measure* if there are C, $c > 0$ such that for every 1-Lipschitz function $f : \Omega \to \mathbb{R}$ with median M:

$$\forall \epsilon > 0, \quad \mu\{\omega|\ |f(\omega) - M| > \epsilon\} < Ce^{-c\epsilon^2 d}$$

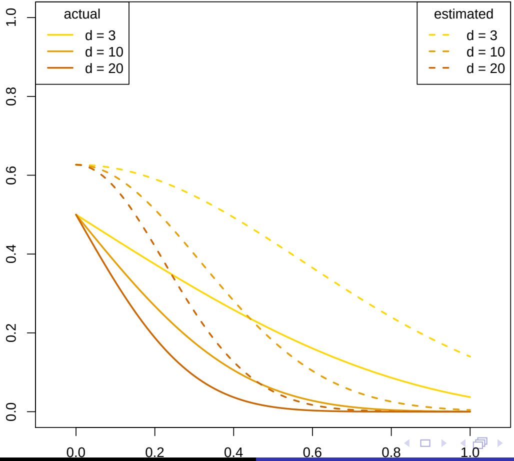
Examples:

- The Spheres $\mathbb{S}^d$ in $\mathbb{R}^{d+1}$
- The Balls $\mathbb{B}^d$.
- The Hamming Cubes Σ^d.

The concentration functions of various spheres



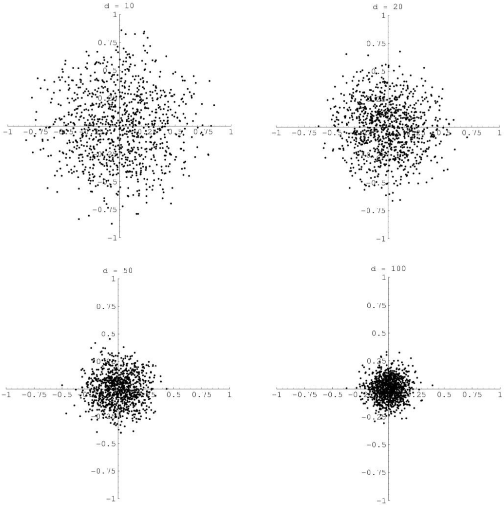
The concentration functions of various spheres



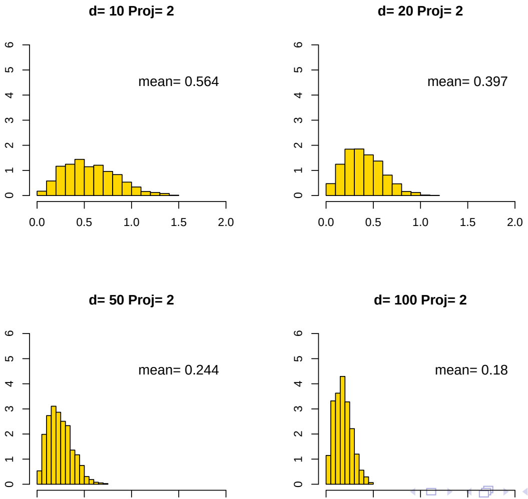
The concentration of measure in spheres

- We can replace $f : \Omega \to \mathbb{R}$ by $f : \Omega \to \mathbb{R}^N$.
- Suppose $f : \mathbb{S}^d \to \mathbb{R}^2$.
- $d = 10, 20, 50, 100$.

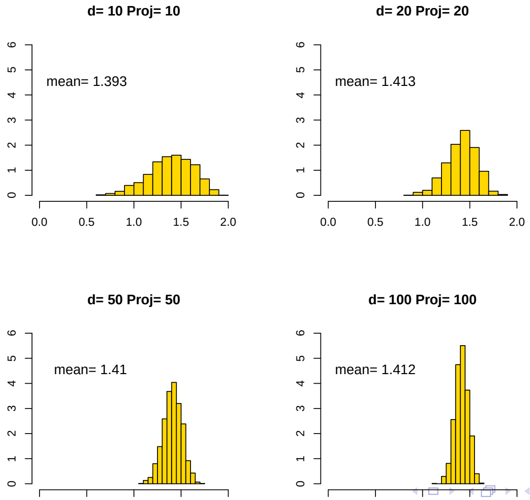
The concentration of measure in spheres



Distribution of distances of projected spheres



Distribution of distances of spheres



Connection to indexing

Observe that

$$\rho(\cdot, p) : \Omega \to \mathbb{R} : \omega \mapsto \rho(\omega, p)$$

is a 1-Lipschitz function, as the Δ-inequality:

$$\rho(\omega_1, p) \leqslant \rho(\omega_1, \omega_2) + \rho(\omega_2, \omega_p)$$
$$\rho(\omega_2, p) \leqslant \rho(\omega_2, \omega_1) + \rho(\omega_1, \omega_p)$$

Leads to:

$$\rho(\omega_1, p) - \rho(\omega_2, \omega_p) \leqslant \rho(\omega_1, \omega_2)$$
$$\rho(\omega_2, p) - \rho(\omega_1, \omega_p) \leqslant \rho(\omega_2, \omega_1)$$

and hence $|\rho(\omega_1, p) - \rho(\omega_2, \omega_p)| \leqslant \rho(\omega_1, \omega_2)$.

Connection to indexing

$\rho(\cdot, p)$ is a 1-Lipschitz function.

- Recall $\mathcal{C}_{q,r,p_1,\ldots,p_k} = \{\omega \in \Omega | \rho_k(q, \omega) > r\}$.
- Compare to $C_{q,r,p_1,\ldots,p_k} = \{x \in X | \rho_k(q, x) > r\}$.
- If concentration of measure is present, it follows that $\mu_d(\mathcal{C}_{q,r,p_1,\ldots,p_k}) < Ce^{-cr^2 d}$.
- We want to know about $|C_{q,r,p_1,\ldots,p_k}|$.

Connection to indexing

$\rho(\cdot, p)$ is a 1-Lipschitz function.

- Recall $\mathcal{C}_{q,r,p_1,\ldots,p_k} = \{\omega \in \Omega | \rho_k(q,\omega) > r\}$.
- Compare to $C_{q,r,p_1,\ldots,p_k} = \{x \in X | \rho_k(q,x) > r\}$.
- If concentration of measure is present, it follows that $\mu_d(\mathcal{C}_{q,r,p_1,\ldots,p_k}) < Ce^{-cr^2 d}$.
- We want to know about $|C_{q,r,p_1,\ldots,p_k}|$.

Connection to indexing

$\rho(\cdot, p)$ is a 1-Lipschitz function.

- Recall $\mathcal{C}_{q,r,p_1,\ldots,p_k} = \{\omega \in \Omega | \rho_k(q, \omega) > r\}$.
- Compare to $C_{q,r,p_1,\ldots,p_k} = \{x \in X | \rho_k(q, x) > r\}$.
- If concentration of measure is present, it follows that $\mu_d(\mathcal{C}_{q,r,p_1,\ldots,p_k}) < Ce^{-cr^2 d}$.
- We want to know about $|C_{q,r,p_1,\ldots,p_k}|$.

Glivenko-Cantelli and the generalization

Let X be an i.i.d. sample of size n from $(\mathbb{R}, \mu)$ (any* prob. measure). If we let $\mu_n(A) := |X \cap A|$ then

$$\sup_{A \in \mathcal{A}} | \mu_n(A) - \mu(A) | \xrightarrow{P} 0$$

where

$$\mathcal{A} = \{(a, b] | a, b \in \mathbb{R}\}.$$

This is known as the Glivenko-Cantelli theorem.

Generalization of Glivenko-Cantelli

Let X be an i.i.d. sample of size n from (Ω, μ). If we let $\mathcal{A}$ be a collection of subsets with the “finite Vapnik-Chervonenkis (VC) dimension Δ” property then

$$\sup_{A \in \mathcal{A}} |\ \mu_n(A) - \mu(A)\ | \xrightarrow{P} 0$$

Furthermore:
We know the rate of convergence: $\exp(-\Delta\varepsilon^2 n)$.

Generalization of Glivenko-Cantelli

Let X be an i.i.d. sample of size n from (Ω, μ). If we let $\mathcal{A}$ be a collection of subsets with the "finite Vapnik-Chervonenkis (VC) dimension Δ" property then

$$\sup_{A \in \mathcal{A}} | \, \mu_n(A) - \mu(A) \, | \xrightarrow{P} 0$$

Furthermore:
We know the rate of convergence: $\exp(-\Delta \varepsilon^2 n)$.

Examples of Spaces with bounds on VC

- The VC dimension of half-spaces in $\mathbb{R}^d$ is $d + 1$.
- The VC-dimension of all open (or closed) balls in $\mathbb{R}^d$

$$\{x \in \mathbb{R}^d | \, \|x - v\| < r\}$$

  is also $d + 1$.
- axis-aligned rectangular parallelepipeds in $\mathbb{R}^d$,

$$[a_1, b_1] \times [a_2, b_2] \times \ldots \times [a_d, b_d]$$

  have a VC dimension of $2d$

Bounds on k-fold Intersections of Spherical Shells

Below Δ denotes the VC dimension of $\mathcal{C}$:

- For $(\mathbb{R}^d, L^2)$, $\Delta \leqslant k(8d + 12)\ln(6k)$.
- For $(\mathbb{R}^d, L^\infty)$, $\Delta \leqslant k(16d + 4)\ln(6k)$.
- For (Σ^d, ρ), $\Delta \leqslant k(8d + 8\log_2 d + 4)\ln(6k)$.

Main Result

From a sequence of metric spaces with measure $(\Omega_d, \rho_d, \mu_d)$, where $d = 1, 2, 3, \ldots$ take i.i.d. samples (datasets) $X_d \sim \mu_d$. Assume

- $(\Omega_d, \rho_d, \mu_d)$ display the concentration of measure.
- The VC dimension of closed balls in (Ω_d, ρ_d) is $O(d)$.
- We build a pivot-index using k pivots, where $k = o(n_d/d)$.
- Sample size n_d satisfies $d = \omega(\log n_d)$ and $d = n_d^{o(1)}$.

Suppose we perform queries of radius=NN. Then:
If we fix arbitrarily small $\varepsilon, \eta > 0$, $\exists D$ **such that for all** $d \geqslant D$, **the probability that** ***at least half*** **the queries on dataset** X_d **take less than** $(1-\varepsilon)n_d$ **time is less than** η.

Main Result

From a sequence of metric spaces with measure $(\Omega_d, \rho_d, \mu_d)$, where $d = 1, 2, 3, \ldots$ take i.i.d. samples (datasets) $X_d \sim \mu_d$. Assume

- $(\Omega_d, \rho_d, \mu_d)$ display the concentration of measure.
- The VC dimension of closed balls in (Ω_d, ρ_d) is $O(d)$.
- We build a pivot-index using k pivots, where $k = o(n_d/d)$.
- Sample size n_d satisfies $d = \omega(\log n_d)$ and $d = n_d^{o(1)}$.

Suppose we perform queries of radius=NN. Then:
If we fix arbitrarily small $\varepsilon, \eta > 0$, $\exists D$ **such that for all** $d \geqslant D$**, the probability that *at least half* the queries on dataset** X_d **take less than** $(1-\varepsilon)n_d$ **time is less than** η**.**

Discussion

1. Rigorous, linear bounds.
2. Independent of choice of pivots.
3. Somewhat artificial situation of growth in d and n.