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 $\Omega$: metric space with metric $\rho$.
- Dataset $X \subset \Omega$, always finite, with metric $\rho$.
- 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 $\mu$ on $\Omega$.
- Treat $X$ as i.i.d. sample $\sim \mu$ of size $n$
Similarity Workloads

- Universe $\Omega$: metric space with metric $\rho$.
- Dataset $X \subset \Omega$, always finite, with metric $\rho$.
- 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 $\mu$ on $\Omega$.
- Treat $X$ as i.i.d. sample $\sim \mu$ of size $n$
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 $\Sigma_d$ 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 \( \Sigma_d \) of dimension \( d \): The set of all binary sequences of length \( d \).
Examples of previous work:

Let $n$ the size of $X$ vary, but the space $(\Omega, \rho, \mu)$ be fixed.

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

Let the dimension and hence $(\Omega, \rho, \mu)$ 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 $\Omega$.
- Both $d$ and $n$ grow.
- Need to make assumptions about the sequence of $\Omega$’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 $\Omega$’s exhibits concentration of measure, the “intrinsic dimension” grows.
4. Statistical Learning Theory: linking properties of $\Omega$’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 \((\Omega_d, \rho_d)\) 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 \(d \geq D\), the probability that at least half the queries on dataset \(X_d\) take less than \((1 - \varepsilon)n_d\) time is less than \(\eta\).
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 \((\Omega_d, \rho_d)\) 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 \(d \geq D\), the probability that at least half the queries on dataset \(X_d\) take less than \((1 - \varepsilon)n_d\) time is less than \(\eta\).
Pivot indexing scheme

Build an index:

1. Pick \( \{p_1 \ldots p_k\} \) from \( X \)
2. Calculate \( n \times k \) array of distances

\[
\rho(x, p_i), \; 1 \leq i \leq k, \; x \in X
\]

Perform query given \( q \) and \( r \):

1. Compute \( \rho_k(q, x) := \sup_{1 \leq i \leq k} |\rho(q, p_i) - \rho(x, p_i)| \).
2. Since \( \rho(q, x) \geq \rho_k(q, x) \), 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 \( \{p_1 \ldots p_k\} \) from \( X \)
2. Calculate \( n \times k \) array of distances

\[ \rho(x, p_i), 1 \leq i \leq k, \; x \in X \]

Perform query given \( q \) and \( r \):
1. Compute \( \rho_k(q, x) := \sup_{1 \leq i \leq k} |\rho(q, p_i) - \rho(x, p_i)| \).
2. Since \( \rho(q, x) \geq \rho_k(q, x) \), 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 $C_{q,r,p_1,...,p_k}$ denote all the discarded points in $X$:
  \[ \{ x \in X | \rho_k(q, x) > r \} \]
  
- Let $n = |X|$.
- Total cost: $k + n - |C_{q,r,p_1,...,p_k}|$. 

Ilya Volnyansky, Vladimir Pestov

Curse of Dimensionality in Pivot-based Indexes
The cost model

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

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

- Let $n = |X|$.
- Total cost: $k + n - |C_{q,r,p_1,...,p_k}|$. 
A function $f : \Omega \rightarrow \mathbb{R}$ is 1-Lipschitz if

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

Examples:

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

Its median is a number $M$ such that

$$\mu\{\omega | f(\omega) \leq M\} \geq 1/2 \text{ and } \mu\{\omega | f(\omega) \geq M\} \geq 1/2$$
A function $f : \Omega \to \mathbb{R}$ is 1-Lipschitz if

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

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

Its median is a number $M$ such that

$$\mu\{\omega | f(\omega) \leq M\} \geq 1/2 \text{ and } \mu\{\omega | f(\omega) \geq M\} \geq 1/2$$
Concentration of Measure

A sequence of spaces $(\Omega_d, \rho_d, \mu_d)^\infty_{d=1}$ 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 $M$:

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

Examples:
- The Spheres $\mathbb{S}^d$ in $\mathbb{R}^{d+1}$
- The Balls $\mathbb{B}^d$.
- The Hamming Cubes $\Sigma^d$. 
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 \(M\):

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

Examples:
- The Spheres \(S^d\) in \(\mathbb{R}^{d+1}\)
- The Balls \(B^d\).
- The Hamming Cubes \(\Sigma^d\).
The concentration functions of various spheres

Ilya Volnyansky, Vladimir Pestov
Curse of Dimensionality in Pivot-based Indexes
The concentration functions of various spheres
The concentration of measure in spheres

- We can replace \( f : \Omega \rightarrow \mathbb{R} \) by \( f : \Omega \rightarrow \mathbb{R}^N \).
- Suppose \( f : S^d \rightarrow \mathbb{R}^2 \).
- \( d = 10, 20, 50, 100 \).
The concentration of measure in spheres

Ilya Volnyansky, Vladimir Pestov

Curse of Dimensionality in Pivot-based Indexes
Distribution of distances of projected spheres

- **d = 10 Proj = 2**
  - mean = 0.564

- **d = 20 Proj = 2**
  - mean = 0.397

- **d = 50 Proj = 2**
  - mean = 0.244

- **d = 100 Proj = 2**
  - mean = 0.18

Ilya Volnyansky, Vladimir Pestov

Curse of Dimensionality in Pivot-based Indexes
Ilya Volnyansky, Vladimir Pestov

Curse of Dimensionality in Pivot-based Indexes
Observe that

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

is a 1-Lipschitz function, as the $\Delta$-inequality:

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

Leads to:

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

and hence

$$
|\rho(\omega_1, p) - \rho(\omega_2, \omega_p)| \leq \rho(\omega_1, \omega_2).
$$
\( \rho(\cdot, p) \) is a 1-Lipschitz function.

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

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

- Recall \( C_{q,r,p_1,\ldots,p_k} = \{ \omega \in \Omega | \rho_k(q,\omega) > r \} \).
- 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
  \[ \mu_d(C_{q,r,p_1,\ldots,p_k}) < Ce^{-cr^2d}. \]
- We want to know about \( |C_{q,r,p_1,\ldots,p_k}| \).
\(\rho(\cdot, p)\) is a 1-Lipschitz function.

- Recall \(C_{q,r,p_1,\ldots,p_k} = \{\omega \in \Omega | \rho_k(q, \omega) > r\}\).
- 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 \(\mu_d(C_{q,r,p_1,\ldots,p_k}) < Ce^{-cr^2d}\).
- 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 \((\Omega, \mu)\). If we let \( A \) be a collection of subsets with the “finite Vapnik-Chervonenkis (VC) dimension \( \Delta \)” property then

\[
\sup_{A \in \mathcal{A}} \left| n \left( \mathbb{P}_{\mu} \right) - P \right| \xrightarrow{P} 0
\]

Furthermore:

We know the rate of convergence: \( \exp(-\Delta \varepsilon^2 n) \).
Let $X$ be an i.i.d. sample of size $n$ from $(\Omega, \mu)$. If we let $\mathcal{A}$ be a collection of subsets with the “finite Vapnik-Chervonenkis (VC) dimension $\Delta$” property then

$$\sup_{A \in \mathcal{A}} \left| \frac{1}{n} \mu_n(A) - \mu(A) \right| \xrightarrow{P} 0$$

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 $\mathbb{R}^d$
  \[ \{ x \in \mathbb{R}^d \mid \| x - v \| < r \} \]
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 $2d$
Below $\Delta$ denotes the VC dimension of $\mathcal{C}$:

- For $(\mathbb{R}^d, L^2)$, $\Delta \leq k(8d + 12) \ln(6k)$.
- For $(\mathbb{R}^d, L^\infty)$, $\Delta \leq k(16d + 4) \ln(6k)$.
- For $(\Sigma^d, \rho)$, $\Delta \leq k(8d + 8 \log_2 d + 4) \ln(6k)$. 
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 \((\Omega_d, \rho_d)\) 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 \(d \geq D\), the probability that at least half the queries on dataset \(X_d\) take less than \((1 - \varepsilon)n_d\) time is less than \(\eta\).
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 $(\Omega_d, \rho_d)$ 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 $d \geq D$, the probability that at least half the queries on dataset $X_d$ take less than $(1 - \varepsilon)n_d$ time is less than $\eta$. 

Ilya Volnyansky, Vladimir Pestov
Curse of Dimensionality in Pivot-based Indexes
Discussion

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