LSH for ℓ_2 distances

Lecture 15

October 18, 2022

LSH Approach for Approximate NNS

Use **locality-sensitive hashing** to solve simplified decision problem

Definition

A family of hash functions is (r, cr, p_1, p_2)-LSH with $p_1 > p_2$ and $c > 1$ if h drawn randomly from the family satisfies the following:

- $\Pr[h(x) = h(y)] \geq p_1$ when $\mathrm{dist}(x, y) \leq r$
- $\Pr[h(x) = h(y)] \leq p_2$ when $\mathrm{dist}(x, y) \geq cr$

Key parameter: the gap between p_1 and p_2 measured as $\rho = \frac{\log p_1}{\log p_2}$ usually small.

Two-level hashing scheme:

- Amplify basic locality sensitive hash family to create better family by repetition
- Use several copies of amplified hash functions

Layer binary search based on r on top of above scheme.

LSH Approach for Approximate NNS

Key parameter: the gap between p_1 and p_2 measured as $\rho = \frac{\log p_1}{\log p_2}$ usually small.

- $L \simeq n^\rho$ hash tables
- Storage: $n^{1+\rho}$ (ignoring log factors)
- Query time: kn^ρ (ignoring log factors) where $k = \log_{1/p_2} n$

LSH for Euclidean Distances

Now $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$ and $\text{dist}(x, y) = \|x - y\|_2$

First do dimensionality reduction (JL) to reduce d (if necessary) to $O(\log n)$ (since we are using c-approximation anyway)

LSH for Euclidean Distances

Now $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$ and $\text{dist}(x, y) = \|x - y\|_2$

First do dimensionality reduction (JL) to reduce d (if necessary) to $O(\log n)$ (since we are using c-approximation anyway)

What is a good basic locality-sensitive hashing scheme? That is, we want a hashing approach that makes nearby points more likely to collide than farther away points.

LSH for Euclidean Distances

Now $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$ and $\text{dist}(x, y) = \|x - y\|_2$

First do dimensionality reduction (JL) to reduce d (if necessary) to $O(\log n)$ (since we are using c-approximation anyway)

What is a good basic locality-sensitive hashing scheme? That is, we want a hashing approach that makes nearby points more likely to collide than farther away points.

Projections onto random lines plus bucketing

Random unit vector

Question: How do we generate a random unit vector in $\mathbb{R}^d$ (same as a uniform point on the sphere S^{n-1})?

Random unit vector

Question: How do we generate a random unit vector in $\mathbb{R}^d$ (same as a uniform point on the sphere S^{n-1})?

- Pick d independent rvs $Z_1, Z_2, \ldots, Z_d$ where each $Z_i \simeq \mathcal{N}(0, 1)$ and let $g = (Z_1, Z_2, \ldots, Z_d)$ (also called a random Guassian vector)
- g is symmetric and hence is a random direction
- to obtain random unit vector normalize $g' = g/\|g\|_2$
- When d is large $\|g\|_2^2 = \sum_i Z_i^2$ is concentrated around d and hence $\|g\|_2 = (1 \pm \epsilon)\sqrt{d}$ with high probability.
- Thus $g/\sqrt{d}$ is a proxy for random unit vector and is easier to work with in many cases

Projection onto a random guassian vector

Lemma

Suppose $\boldsymbol{x} \in \mathbb{R}^{\boldsymbol{d}}$ *and* $\boldsymbol{g}$ *is a random Guassian vector. Let* $\boldsymbol{Y} = \boldsymbol{x} \cdot \boldsymbol{g}$*. Then* $\boldsymbol{Y} \sim \mathcal{N}(0, \|\boldsymbol{x}\|_2)$ *and hence* $\boldsymbol{E}[\boldsymbol{Y}^2] = (\|\boldsymbol{x}\|_2)^2$*.*

Hashing scheme

- Pick a random unit Guassian vector $\boldsymbol{u}$
- Pick a random shift $\boldsymbol{a} \in (0, \boldsymbol{r}]$
- For vector $\boldsymbol{x}$ set $\boldsymbol{h}_{\boldsymbol{u},\boldsymbol{a}} = \lfloor \frac{\boldsymbol{x} \cdot \boldsymbol{u} + \boldsymbol{a}}{\boldsymbol{r}} \rfloor$

Analysis

Suppose x, y are such that $\|x - y\|_2 \leq r$. What is $p_1 = \Pr[h_{u,a}(x) = h_{u,a}(y)]$

Suppose x, y are such that $\|x - y\|_2 \geq cr$. What is $p_2 = \Pr[h_{u,a}(x) = h_{u,a}(y)]$

Analysis

Suppose x, y are such that $\|x - y\|_2 \leq r$. What is $p_1 = \Pr[h_{u,a}(x) = h_{u,a}(y)]$

Suppose x, y are such that $\|x - y\|_2 \geq cr$. What is $p_2 = \Pr[h_{u,a}(x) = h_{u,a}(y)]$

Let $q = x - y$. Let $s = \|q\|_2$ be length of q.
From Lemma $q \cdot g$ is distributed as $\mathcal{N}(0, s^2)$.

Analysis

Suppose x, y are such that $\|x - y\|_2 \leq r$. What is $p_1 = \Pr[h_{u,a}(x) = h_{u,a}(y)]$

Suppose x, y are such that $\|x - y\|_2 \geq cr$. What is $p_2 = \Pr[h_{u,a}(x) = h_{u,a}(y)]$

Let $q = x - y$. Let $s = \|q\|_2$ be length of q.
From Lemma $q \cdot g$ is distributed as $\mathcal{N}(0, s^2)$.

Observations:

- $h(x) \neq h(y)$ if $|q \cdot g| \geq r$
- If $|q \cdot g| < r$ then $h(x) = h(y)$ with probability $1 - |q \cdot g|/r$

Analysis

Suppose x, y are such that $\|x - y\|_2 \leq r$. What is $p_1 = \Pr[h_{u,a}(x) = h_{u,a}(y)]$

Suppose x, y are such that $\|x - y\|_2 \geq cr$. What is $p_2 = \Pr[h_{u,a}(x) = h_{u,a}(y)]$

Let $q = x - y$. Let $s = \|q\|_2$ be length of q.
From Lemma $q \cdot g$ is distributed as $\mathcal{N}(0, s^2)$.

Observations:

- $h(x) \neq h(y)$ if $|q \cdot g| \geq r$
- If $|q \cdot g| < r$ then $h(x) = h(y)$ with probability $1 - |q \cdot g|/r$

Thus collision probability depends only on s

Analysis

Let $q = x - y$. Let $s = \|q\|_2$ be length of q.
From Lemma $q \cdot g$ is distributed as $\mathcal{N}(0, s^2)$.

- $h(x) \neq h(y)$ if $|q \cdot g| \geq r$
- If $|q \cdot g| < r$ then $h(x) = h(y)$ with probability $1 - |q \cdot g|/r$

For a fixed s collision probability is

$$p(s) = \int_0^r f(t)(1 - t/r)dt$$

where f is the density function of $|\mathcal{N}(0, s^2)|$.
Rewriting

$$p(s) = \int_0^r \frac{1}{s} f(\frac{t}{s})(1 - t/r)dt$$

where f is the density function of the $|\mathcal{N}(0, 1)|$.

Analysis

$$p(s) = \int_0^r \frac{1}{s} f(\frac{t}{s})(1 - t/r)dt$$

where f is the density function of the $|\mathcal{N}(0,1)|$.

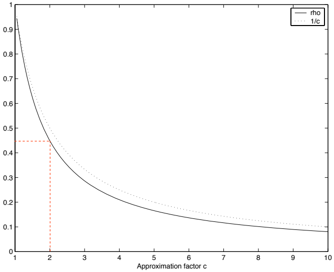
Recall $p_1 = p(r)$ and $p_2 = p(cr)$ and we are interested in $\rho = \frac{\log p_1}{\log p_2}$.

Analysis

$$p(s) = \int_0^r \frac{1}{s} f(\frac{t}{s})(1 - t/r)dt$$

where f is the density function of the $|\mathcal{N}(0,1)|$.

Recall $p_1 = p(r)$ and $p_2 = p(cr)$ and we are interested in $\rho = \frac{\log p_1}{\log p_2}$.
Show $\rho < 1/c$ by plot



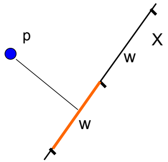
NNS for Euclidean distances

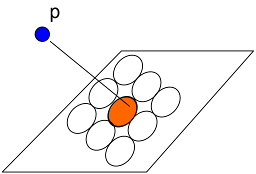
- For any fixed $c > 1$ use above scheme to obtain
  - Storage: $O(n^{1+1/c}\text{polylog}(n))$
  - Query time: $O(dn^{1/c}\text{polylog}(n))$
- Can use JL to reduce d to $O(\log n)$.

Improved LSH Scheme

[Andoni-Indyk'06]

- Basic LSH scheme projects points into lines
- Better scheme: pick some small constant t and project points into R^t
- Use lattice based space partitioning scheme to "bucket" instead of intervals





Figures from Piotr Indyk's slides

Improved LSH Scheme

[Andoni-Indyk'06]

- Basic LSH scheme projects points into lines
- Better scheme: pick some small constant t and project points into R^t
- Use lattice based space partitioning scheme to "bucket" instead of intervals
- Leads to $\rho \simeq 1/c^2 + O(\log t/\sqrt{t})$ and hence tends to $1/c^2$ for large t and fixed c
- Lower bound for LSH in ℓ_2 says $\rho \geq 1/c^2$

Data dependent LSH Scheme

LSH is data oblivious. That is, the hash families are chosen before seeing the data. Can one do better by choosing hash functions based on the given set of points?

Data dependent LSH Scheme

LSH is data oblivious. That is, the hash families are chosen before seeing the data. Can one do better by choosing hash functions based on the given set of points?

Yes.
[Andoni-Indyk-Ngyuyen-Razenshteyn'14, Andoni-Razensteyn'15]

- $\rho = 1/(2c^2 - 1)$ for ℓ_2 improving upon $1/c^2$ for data oblivious LSH (which is tight in worst case)
- $\rho = 1/(c^2 - 1)$ for ℓ_1/Hamming cube improving upon $1/c$ for data oblivious LSH

LSH Summary

- A modular hashing based scheme for similarity estimation
  - Non-trivial tradeoff between storage and query time
  - Simple to implement
  - Reduces dependence on dist to designing a class of hash functions. Rest of machinery common.
  - Provides speedups but uses more memory
- Main competitors are space partitioning data structures such as (randomized) variants of k-d trees: work well in low-dimensions
- Does not appear to be a clear winner
- Nearest neighbor search and related ideas not used in isolation. Need to take other factors into account.

Overall a new perspective on nearest neighbor search that provides practical flexibility and interesting tradeoffs and questions.

Digression: p-stable distributions

For F_2 estimation and JL and LSH we used important "stability" property of the Normal distribution.

Lemma

Let $Y_1, Y_2, \ldots, Y_d$ be independent random variables with distribution $\mathcal{N}(0, 1)$. $Z = \sum_i x_i Y_i$ has distribution $\|x\|_2 \mathcal{N}(0, 1)$

Standard Gaussian is 2-stable.

Digression: p-stable distributions

For F_2 estimation and JL and LSH we used important "stability" property of the Normal distribution.

Lemma

Let $Y_1, Y_2, \ldots, Y_d$ be independent random variables with distribution $\mathcal{N}(0,1)$. $Z = \sum_i x_i Y_i$ has distribution $\|x\|_2 \mathcal{N}(0,1)$

Standard Gaussian is 2-stable.

Definition

A distribution $\mathcal{D}$ is p-stable if $Z = \sum_i x_i Y_i$ has distribution $\|x\|_p \mathcal{D}$ when the Y_i are independent and each of them is distributed as $\mathcal{D}$.

Digression: p-stable distributions

For F_2 estimation and JL and LSH we used important "stability" property of the Normal distribution.

Lemma

Let $Y_1, Y_2, \ldots, Y_d$ be independent random variables with distribution $\mathcal{N}(0,1)$. $Z = \sum_i x_i Y_i$ has distribution $\|x\|_2 \mathcal{N}(0,1)$

Standard Gaussian is 2-stable.

Definition

A distribution $\mathcal{D}$ is p-stable if $Z = \sum_i x_i Y_i$ has distribution $\|x\|_p \mathcal{D}$ when the Y_i are independent and each of them is distributed as $\mathcal{D}$.

Question: Do p-stable distributions exist for $p \neq 2$?

p-stable distributions

Fact: p-stable distributions exist for all $p \in (0, 2]$ and do not exist for $p > 2$.

$p = 1$ is the Cauchy distribution which is the distribution of the ratio of two independent Guassian random variables. Has a closed form density function $\frac{1}{\pi(1+x^2)}$. Mean and variance are not finite.

p-stable distributions

Fact: p-stable distributions exist for all $p \in (0, 2]$ and do not exist for $p > 2$.

$p = 1$ is the Cauchy distribution which is the distribution of the ratio of two independent Guassian random variables. Has a closed form density function $\frac{1}{\pi(1+x^2)}$. Mean and variance are not finite.

For general p no closed form formula for density but can sample from the distribution.

p-stable distributions

Fact: p-stable distributions exist for all $p \in (0, 2]$ and do not exist for $p > 2$.

$p = 1$ is the Cauchy distribution which is the distribution of the ratio of two independent Guassian random variables. Has a closed form density function $\frac{1}{\pi(1+x^2)}$. Mean and variance are not finite.

For general p no closed form formula for density but can sample from the distribution.

Streaming, sketching, LSH ideas for ℓ_2 generalize to ℓ_p for $p \in (0, 2]$ via p-stable distributions and additional technical work.

Digression: Doubling dimension

Thesis/assumption: Real world data is high-dimensional in explicit representation but low-dimensional in "content".

Several interpretations of what it means for data to be low-dimensional

- Data lies in a low-dimensional manifold
- Data can be projected into low dimensions while preserving certain properties (JL for instance)
- Data has a latent low-dimensional description (SVD, PCA, tensor decomposition, etc)
- Data has low doubling dimension
- $\cdots$

Intrinsic dimension

Let (V, dist) be a finite metric space.

- $\text{dist}(x, y) = \text{dist}(y, x)$ for all $x, y \in V$ (symmetry)
- $\text{dist}(x, x) = 0$ for all $x \in V$ (reflexivity)
- $\text{dist}(x, y) + \text{dist}(y, z) \geq \text{dist}(x, z)$ for all $x, y, z \in V$ (triangle inequality)

Question: Can we quantify whether (V, dist) behaves like a low-dimensional Euclidean space? Does this have any benefits?

Doubling dimension

Property of $\mathbb{R}^{d}$: A ball of radius r can be covered by c^{d} balls of radius $r/2$ for some constant $c \leq 4$.

Doubling dimension

Property of $\mathbb{R}^d$: A ball of radius r can be covered by c^d balls of radius $r/2$ for some constant $c \leq 4$.

Given (V, d) let $B(p, r)$ be the ball or radius r around p and view it as a set of points:

$$B(p, r) = \{q \mid \text{dist}(p, q) \leq r\}$$

Doubling dimension

Property of $\mathbb{R}^d$: A ball of radius r can be covered by c^d balls of radius $r/2$ for some constant $c \leq 4$.

Given (V, d) let $B(p, r)$ be the ball or radius r around p and view it as a set of points:

$$B(p, r) = \{q \mid \text{dist}(p, q) \leq r\}$$

Definition

A finite metric space (V, dist) has doubling dimension d if for all $p \in V$ and all $r > 0$, $B(p, r)$ can be covered by 2^d balls of radius at most $r/2$.

Doubling dimensions

Definition

A finite metric space (V, dist) has doubling dimension d if for all $p \in V$ and all $r > 0$, $B(p, r)$ can be covered by 2^d balls of radius at most $r/2$.

Many algorithms/data structures for $\mathbb{R}^d$ can be extended to metric spaces with doubling dimension d with comparable running times.

Including approximate NNS.
See [Clarkson, Krauthgamer-Lee, HarPeled-Mendel]