Locality-Sensitive Hashing

Anil Maheshwari

anil@scs.carleton.ca
School of Computer Science
Carleton University
Canada

Outline

Introduction

Objectives

How to find efficiently

1. Similar documents among a collection of documents
2. Similar web-pages among web-pages
3. Similar fingerprints among a database of fingerprints
4. Similar sets among a collection of sets
5. Similar images from a database of images
6. Similar vectors in higher dimensions.

Similarity of Documents

Similarity of Documents

Problem Definition

Input: A collection of web-pages.

Output: Report near duplicate web-pages.

k-shingles

Any substring of k words that appears in the document.

Text Document $=$ “What is the likely date that the regular classes may resume in Ontario”

$2-$shingles: What is, is the, the likely, $\ldots$, in Ontario

$3-$shingles: What is the, is the likely, $\ldots$, resume in Ontario

In practice: $9-$shingles for English Text and $5-$shingles for e-mails

Similarity between sets

Text Document $D \rightarrow$ Set S

1. Form all the k-shingles of D
2. S is the collection of all k-shingles of D

Jaccard Similarity

For a pair of sets S and T, the Jaccard Similarity is defined as $\text{SIM}(S,T) = \frac{|S \cap T|}{|S \cup T|}$

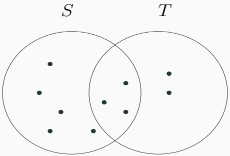


Figure 1: $|S| = 8, |T| = 5, |S \cup T| = 9, |S \cap T| = 3$, $\text{SIM}(S,T) = \frac{|S \cap T|}{|S \cup T|} = \frac{3}{9} = \frac{1}{3}$

Problem: Find Similar Sets

New Problem

Given a constant $0 \leq s \leq 1$ and a collection of sets $\mathcal{S}$, find the pairs of sets in $\mathcal{S}$ with Jaccard similarity $\geq s$

$U = \{$Cruise, Ski, Resorts, Safari, Stay@Home$\}$

$S_1 = \{$Cruise, Safari$\}$ $S_3 = \{$Ski, Safari, Stay@Home$\}$

$S_2 = \{$Resorts$\}$ $S_4 = \{$Cruise, Resorts, Safari$\}$

Problem: Given $\mathcal{S} = \{S_1, S_2, S_3, S_4\}$ and $s = \frac{1}{2}$, report all pairs that are s-similar.

$\text{SIM}(S_1, S_2) = \frac{0}{3} = 0$ $\text{SIM}(S_2, S_3) = \frac{0}{4} = 0$

$\text{SIM}(S_1, S_3) = \frac{1}{4}$ $\text{SIM}(S_2, S_4) = \frac{1}{3}$

$\text{SIM}(S_1, S_4) = \frac{2}{3}$ $\text{SIM}(S_3, S_4) = \frac{1}{5}$

Characteristic Matrix Representation of Sets

$U = \{$Cruise, Ski, Resorts, Safari, Stay@Home$\}$

$\mathcal{S} = \{S_1, S_2, S_3, S_4\}$, where each $S_i \subseteq U$
e.g. $S_1 = \{$Cruise, Safari$\}$ and $S_2 = \{$Resorts$\}$

Characteristic matrix for $\mathcal{S}$:

	S_1	S_2	S_3	S_4
Cruise	1	0	0	1
Ski	0	0	1	0
Resorts	0	1	0	1
Safari	1	0	1	1
Stay@Home	0	0	1	0

MinHash Signatures via Random Permutation

Permute Rows of characteristic matrix - $\pi : 01234 \rightarrow 40312$

		S_1	S_2	S_3	S_4
0	Cruise	1	0	0	1
1	Ski	0	0	1	0
2	Resorts	0	1	0	1
3	Safari	1	0	1	1
4	Stay@Home	0	0	1	0

		S_1	S_2	S_3	S_4
0(1)	Ski	0	0	1	0
1(3)	Safari	1	0	1	1
2(4)	Stay@Home	0	0	1	0
3(2)	Resorts	0	1	0	1
4(0)	Cruise	1	0	0	1

Minhash Signatures for a set S_i w.r.t. π is the **row-number** of first non-zero element in the column corresponding to S_i

$h(S_1) = 1$
$h(S_2) = 3$
$h(S_3) = 0$
$h(S_4) = 1$

Key Lemma

Lemma

For any two sets S_i and S_j in a collection of sets $\mathcal{S}$ where the elements are drawn from the universe U, the probability that the minhash value $h(S_i)$ equals $h(S_j)$ is equal to the Jaccard similarity of S_i and S_j, i.e., $Pr[h(S_i) = h(S_j)] = \mathsf{SIM}(S_i, S_j) = \frac{|S_i \cap S_j|}{|S_i \cup S_j|}$.

		S_1	S_2	S_3	S_4
0	Ski	0	0	1	0
1	Safari	1	0	1	1
2	Stay@Home	0	0	1	0
3	Resorts	0	1	0	1
4	Cruise	1	0	0	1

$$Pr[h(S_1) = h(S_4)] = \mathsf{SIM}(S_1, S_4) = \frac{|S_1 \cap S_4|}{|S_1 \cup S_4|} = \frac{2}{3}$$

Proof of Key Observation

Consider the rows corresponding to the columns of S_i and S_j.

Let $x =$ Number of rows where both the columns have a 1.

Let $y =$ Number of rows where exactly one of the columns has a 1.

S_1	S_4		
0	0		
1	1	$\rightarrow$	x
0	0		
0	1	$\rightarrow$	y
1	1	$\rightarrow$	x

Observe that $|S_i \cap S_j| = x$ and $|S_i \cup S_j| = x + y$.

Note that the rows where both the columns have 0's can't be the minHash signature of S_i or S_j.

Probability that $h(S_i) = h(S_j)$ is same as that the row corresponding to x is the 'first one' as compared to the rows corresponding to y.

Thus, $Pr[h(S_i) = h(S_j)] = \frac{x}{x+y} = \frac{|S_i \cap S_j|}{|S_i \cup S_j|} = \text{SIM}(S_i, S_j)$ □

MinHash Signature matrix for $|\mathcal{S}| = 11$ sets with 12 hash functions

S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}
2	2	1	0	0	1	3	2	5	0	3
1	3	2	0	2	2	1	4	2	1	2
3	0	3	0	4	3	2	0	0	4	2
0	4	3	1	5	3	3	2	3	5	4
2	1	1	0	4	1	2	1	4	2	5
4	2	1	0	5	2	3	2	3	5	4
2	4	3	0	5	3	3	4	4	5	3
0	2	4	1	3	4	3	2	2	2	4
0	2	1	0	5	1	1	1	1	5	1
0	5	1	0	2	1	3	2	1	5	4
1	3	1	0	5	2	3	3	6	3	2
0	5	2	1	5	1	2	2	6	5	4

LSH

LSH for MinHash

Partitioning of a signature matrix into $b = 4$ bands of $r = 3$ rows each.

Band	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}
	2	2	1	0	0	1	3	2	5	0	3
I	1	3	2	0	2	2	1	4	2	1	2
	3	0	3	0	4	3	2	0	0	4	2
	0	4	3	1	5	3	3	2	3	5	4
II	2	1	1	0	4	1	2	1	4	2	5
	4	2	1	0	5	2	3	2	3	5	4
	2	4	3	0	5	3	3	4	4	5	3
III	0	2	4	1	3	4	3	2	2	2	4
	0	2	1	0	5	1	1	1	1	5	1
	0	5	1	0	2	1	3	2	1	5	4
IV	1	3	1	0	5	2	3	3	6	3	2
	0	5	2	1	5	1	2	2	6	5	4

Band 3: $\{S_3, S_6, S_{11}\}$ are hashed into the same bucket, and so are $\{S_8, S_9\}$

Probability of finding similar sets

Lemma

Let $s > 0$ be the Jaccard similarity of two sets. The probability that the minHash signature matrix agrees in all the rows of at least one of the bands for these two sets is $f(s) = 1 - (1 - s^r)^b$.

Band	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}
	2	2	1	0	0	1	3	2	5	0	3
I	1	3	2	0	2	2	1	4	2	1	2
	3	0	3	0	4	3	2	0	0	4	2
	0	4	3	1	5	3	3	2	3	5	4
II	2	1	1	0	4	1	2	1	4	2	5
	4	2	1	0	5	2	3	2	3	5	4
	2	4	3	0	5	3	3	4	4	5	3
III	0	2	4	1	3	4	3	2	2	2	4
	0	2	1	0	5	1	1	1	1	5	1
	0	5	1	0	2	1	3	2	1	5	4
IV	1	3	1	0	5	2	3	3	6	3	2
	0	5	2	1	5	1	2	2	6	5	4

Proof

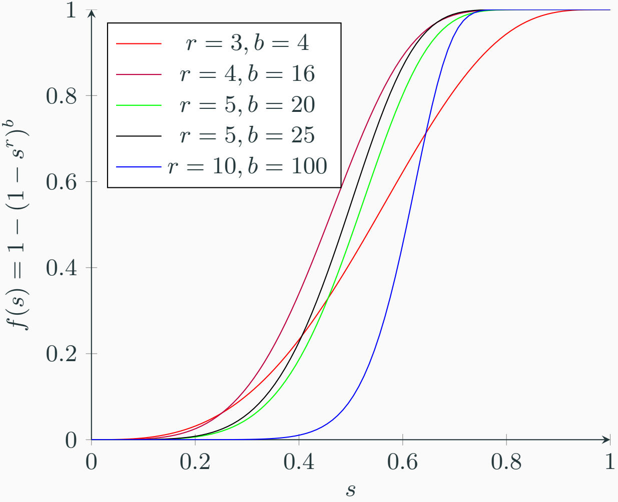
Claim: Pr(signatures agree in all rows of ≥ 1 bands for S_i and S_j with Jaccard Similarity s)= $f(s) = 1 - (1 - s^r)^b$. Answer the following:

1. Probability that the signature agrees in a row
2. Probability that the signature agrees in all rows of a band
3. Probability that the signature doesn't agree in at least one of the rows of a band
4. Probability that the signature doesn't agree in any of the bands
5. Probability that the signature agrees in at least one of the bands

Understanding $f(s)$

$f(s) = 1 - (1 - s^r)^b$ for different values of s, b, and r:

(b, r) $f(s) = 1 - (1 - s^r)^b \searrow$	$(4, 3)$	$(16, 4)$	$(20, 5)$	$(25, 5)$	$(100, 10)$
$s = 0.2$	0.0316	0.0252	0.0063	0.0079	0.0000
$s = 0.4$	0.2324	0.3396	0.1860	0.2268	0.0104
$s = 0.5$	0.4138	0.6439	0.4700	0.5478	0.0930
$s = 0.6$	0.6221	0.8914	0.8019	0.8678	0.4547
$s = 0.8$	0.9432	0.9997	0.9996	0.9999	0.9999
$s = 1.0$	1.0	1.0	1.0	1.0	1.0
Threshold $t = (\frac{1}{b})^{(\frac{1}{r})}$	0.6299	0.5	0.5492	0.5253	0.6309



$r = 3, b = 4$
$r = 4, b = 16$
$r = 5, b = 20$
$r = 5, b = 25$
$r = 10, b = 100$
$f(s) = 1 - (1 - s^r)^b$
s
0
0.2
0.4
0.6
0.8
1

Comments on S-Curve

1. For what values of s, $f''(s) = 0$?
   $s = (\frac{r-1}{br-1})^{\frac{1}{r}}$
2. For values of $br >> 1$, $s \approx (\frac{1}{b})^{\frac{1}{r}}$
3. Steepest slope occurs at $s \approx (1/b)^{(1/r)}$
4. If the Jaccard similarity s of the two sets is above the threshold $t = (\frac{1}{b})^{\frac{1}{r}}$, the probability that they will be found potentially similar is very high.
5. Consider the entries in the row corresponding to $s = 0.8$ in the table and observe that most of the values for $f(s = 0.8) \to 1$ as $s > t$.

Computational Summary

- **Input:** Collection of m text documents of size $\mathcal{D}$
- k-shingles: Size $= k\mathcal{D}$
- Characteristic matrix of size $|U| \times m$, where U is the universe of all possible k-shingles
- Signature matrix of size $n \times m$ using n-permutations
- $\lceil \frac{n}{r} \rceil$ bands each consisting of r rows
- Hash maps from bands to buckets
- Output: All pairs of documents that are in the same bucket corresponding to a band
- Check whether the pairs correspond to similar documents!
- With the right choice of threshold
  Pr(the pair is similar)$\to 1$

What makes LSH works?

How can we apply for other ‘similarity’ problems?

How can we apply for ‘nearest neighbor’ problems?

Metric Spaces

Metric Spaces

Consider a finite set X. A *metric* or *distance measure* d on X is a function $d : X \times X \to [0, \infty)$ satisfying the following properties. For all elements $u, v, w \in X$:

1. Non-negativity: $d(u, v) \geq 0$.
2. Symmetric: $d(u, v) = d(v, u)$.
3. Identity: $d(u, v) = 0$ if and only if $u = v$.
4. Triangle Inequality: $d(u, v) + d(v, w) \geq d(u, w)$.

Examples: Euclidean distance among set of n-points in plane.

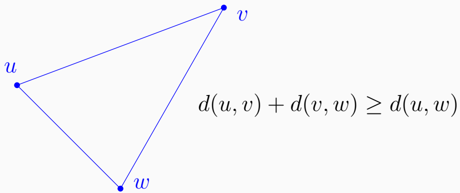
Euclidean Distance

Let X = Set of n-points in plane.
Euclidean distance between any two points $p_i = (x_i, y_i)$ and $p_j = (x_j, y_j)$ is $d(p_i, p_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$.

Euclidean Distance Metric

X with the Euclidean distance measure satisfies the metric properties.

1. Non-negativity: $d(u, v) \geq 0$.
2. Symmetric: $d(u, v) = d(v, u)$.
3. Identity: $d(u, v) = 0$ if and only if $u = v$.
4. Triangle Inequality: $d(u, v) + d(v, w) \geq d(u, w)$.



Hamming Distance Metric

X = Set of d-dimensional Boolean vectors.
Hamming distance HAM(u, v)= Number of coordinates in which two vectors $u, v \in X$ differ.

An Example:

$u =$	1	0	0	1	1	0	1	1
$v =$	1	1	0	0	1	1	1	1

HAM$(u, v) = 3$

Hamming Distance Metric

Hamming distance is a metric over the d-dimensional vectors.

1. Non-negativity: HAM$(u, v) \geq 0$.
2. Symmetric: HAM$(u, v) =$ HAM(v, u).
3. Identity: HAM$(u, v) = 0$ if and only if $u = v$.
4. Triangle Inequality: HAM$(u, v) +$ HAM$(v, w) \geq$ HAM(u, w).

Jaccard Distance Metric

$\mathcal{S}$ = A collection of sets.
Jaccard Similarity doesn't satisfy metric properties, e.g. $\text{SIM}(S, S) = 1$.

Define Jaccard Distance between two sets $S_i, S_j \in \mathcal{S}$ as
$\text{JD}(S_i, S_j) = 1 - \text{SIM}(S_i, S_j)$.

Jaccard Distance Metric

Set $\mathcal{S}$ with the Jaccard distance measure satisfies the metric properties.

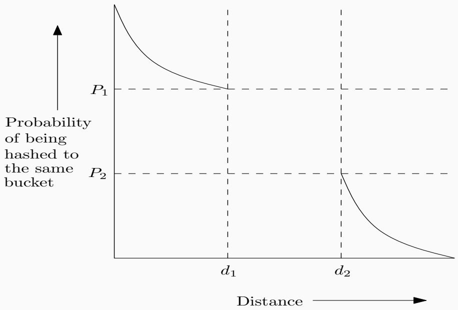
1. Non-negativity: $\text{JD}(S_i, S_j) \geq 0$.
2. Symmetric: $\text{JD}(S_i, S_j) = \text{JD}(S_j, S_i)$.
3. Identity: $\text{JD}(S_i, S_j) = 0$ if and only if $S_i = S_j$.
4. Triangle Inequality: $\text{JD}(S_i, S_j) + \text{JD}(S_j, S_k) \geq \text{JD}(S_i, S_k)$.

Sensitive Function Family

Sensitive Family of Functions

Let d be a distance measure and let $d_1 < d_2$ be two distances. Let $0 \leq p_2 < p_1 \leq 1$. A family of functions $\mathcal{F}$ is said to be (d_1, d_2, p_1, p_2)-sensitive if for every $f \in \mathcal{F}$ the following two conditions hold;

1. If $d(x, y) \leq d_1$ then $Pr[f(x) = f(y)] \geq p_1$.
2. If $d(x, y) \geq d_2$ then $Pr[f(x) = f(y)] \leq p_2$.



Family of MinHash Signatures

Consider the Jaccard distance measure for finding similar sets in a collection of sets $\mathcal{S}$.

Min-Hash Signature Family

Let $0 \leq d_1 < d_2 \leq 1$. The family of minhash-signatures is $(d_1, d_2, p_1 = 1 - d_1, p_2 = 1 - d_2)$-sensitive.

Proof: Suppose that the Jaccard similarity between two sets is at least s. Then their Jaccard distance is at most $d_1 = 1 - s$. The probability that they will be hashed to the same bucket by minhash signatures is $\geq p_1 = s = 1 - d_1$.

Now suppose that the Jaccard similarity is at most s'. Then their Jaccard distance is at least $d_2 = 1 - s'$. The probability that the minhash signatures map them to the same bucket is at most $p_2 = s' = 1 - d_2$.

□

LSH Family for Hamming Distance

Consider two d-dimensional Boolean vectors u and v.
HAM(u, v)= Number of coordinates in which u and v differ
Let $f_i(x) = i$-th coordinate of u.
For a randomly chosen index i, $Pr[f_i(u) = f_i(v)] = 1 - \frac{\text{HAM}(u,v)}{d}$

Example:

$u =$	1	0	0	1	1	0	1	1
$v =$	1	1	0	0	1	1	1	1

$Pr[f_i(u) = f_i(v)] = 1 - \frac{\text{HAM}(u,v)}{d} = 1 - \frac{3}{8} = \frac{5}{8}$

Sensitive-family for Hamming distance

For any $d_1 < d_2$, $\mathcal{F} = \{f_1, f_2, \ldots, f_d\}$ is a $(d_1, d_2, 1 - d_1/d, 1 - d_2/d)$-sensitive family of functions.

Proof: Let $p_1 = 1 - d_1/d$ and $p_2 = 1 - d_2/d$.
A family of functions $\mathcal{F}$ is said to be (d_1, d_2, p_1, p_2)-sensitive if for every $f_i \in \mathcal{F}$ the following two conditions hold:

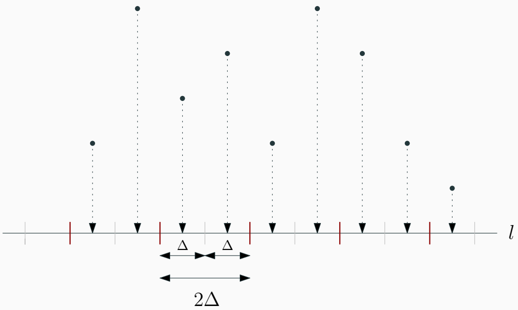
1. If HAM$(u, v) \leq d_1$ then $Pr[f_i(u) = f_i(v)] \geq p_1$
2. If HAM$(u, v) \geq d_2$ then $Pr[f_i(u) = f_i(v)] \leq p_2$

□

LSH Family for Near Neighbors in 2D

P= Set of points in 2D and $\Delta > 0$ a parameter.
Define hash function f_l by a line l with random orientation as follows:

Partition l into intervals of equal size 2Δ.
Orthogonally project all points of P on l.
Let $f_l(x)$ be the interval in which $x \in P$ projects to.



LSH Family for Near Neighbors

Sensitive Family via Projection on a Random Line

The family of hash functions with respect to the projection on a random line with intervals of size 2Δ is a $(\Delta, 4\Delta, 1/2, 1/3)$-sensitive family.

Proof: Assume l is horizontal.

We first show that if $d(x, y) \leq \Delta$, then $Pr[f_l(x) = f_l(y)] \geq 1/2$.

Let m be the mid-point of the interval $f_l(x)$.

In $f_l(x)$, with probability $1/2$ the projection of x lies to the left of m and with probability $1/2$, the projection of y lies to the right of projection of x.

$\implies$ projection of y lies in $f_l(x)$ (i.e., $f_l(x) = f_l(y)$) as $d(x, y) \leq \Delta$.

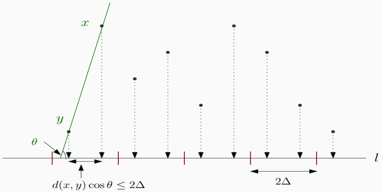
Thus with probability $1/4$, projections of x and y lie in $f_l(x)$ where the projection of x is to the left of m and the projection of y is to the right of the projection of x.

Same reasoning holds when $f_l(x)$ is to the right of m and the projection of y is to the left of the projection of x.

Since the above two cases are mutually exclusive, $Pr[f_l(x) = f_l(y)] \geq 1/2$.

Proof (contd.)

Now consider the case when $d(x, y) > 4\Delta$.



We want to show that $Pr[f_l(x) = f_l(y)] \leq 1/3$.

Let θ be the angle of the line passing through x and y with respect to l.

For the projections of x and y to fall in the same interval, we will need that $d(x, y)\cos\theta \leq 2\Delta$.

For this to happen $\cos\theta \leq 1/2$, or the angle the line xy forms with the horizontal needs to be between 60° and 90°.

This has at most $1/3$-rd chance.

$\square$

AND-OR Family

Let $\mathcal{F}$ be (d_1, d_2, p_1, p_2)-sensitive family.
Construct a new family $\mathcal{G}$ by an *AND-construction* as follows:

AND-Family: Each function $g \in \mathcal{G}$ is formed from a set of r independently chosen functions of $\mathcal{F}$, say $f_1, f_2, \ldots, f_r$ for some fixed value of r.
Now, $g(x) = g(y)$ if and only if for all $i = 1, \ldots, r$, $f_i(x) = f_i(y)$.

AND-Family

$\mathcal{G}$ is an (d_1, d_2, p_1^r, p_2^r)-sensitive AND family.

Proof: This is the probability of all the r independent events to occur simultaneously.

OR-Family: Each member g in $\mathcal{G}$ is constructed by taking b independently chosen members $f_1, f_2, \ldots, f_b$ from $\mathcal{F}$.
Now $g(x) = g(y)$ if and only if $f_i(x) = f_i(y)$ for at least one of the members in $\{f_1, f_2, \ldots, f_b\}$.

OR-Family

$\mathcal{G}$ is an $(d_1, d_2, 1-(1-p_1)^b, 1-(1-p_2)^b)$-sensitive OR family.

Proof: Estimate the probability that none of the b-events occur and then look at the complementary event.

Probabilistic Amplification

	$\mathcal{F}_1$ (AND)	$\mathcal{F}_2$ (OR)	$\mathcal{F}_3$ (AND-OR)	$\mathcal{F}_4$ (OR-AND)
p	p^r	$1-(1-p)^b$	$1-(1-p^r)^b$	$(1-(1-p)^r)^b$
0.2	0.0001	0.6723	0.0079	0.0717
0.4	0.0256	0.9222	0.1216	0.4995
0.6	0.1296	0.9897	0.5004	0.8783
0.7	0.2401	0.9975	0.7446	0.9601
0.8	0.4096	0.9996	0.9282	0.9920
0.9	0.6561	0.9999	0.9951	0.9995

Table 1: Illustration of four families obtained for different values of p. $\mathcal{F}_1$ is the AND family for $r=4$. $\mathcal{F}_2$ is OR family for $b=5$. $\mathcal{F}_3$ is the AND-OR family for $r=4$ and $b=5$. $\mathcal{F}_4$ is the OR-AND family for $r=4$ and $b=5$.

Probabilistic Amplification Examples

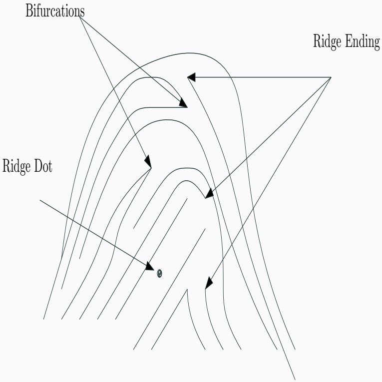
We can apply the AND-OR amplification technique for any sensitive family. For example,

1. $\mathcal{F}$ be a $(d_1, d_2, p_1 = 1 - d_1, p_2 = 1 - d_2)$-sensitive minhash function family for similarity of sets.
2. Hamming distance $(d_1, d_2, 1 - d_1/d, 1 - d_2/d)$-sensitive family for finding similar Boolean strings.
3. Projection on a random line $(\Delta, 4\Delta, 1/2, 1/3)$-sensitive family for finding near points.
4. Metric Property $\rightarrow$ Sensitive Family $\rightarrow$ Probabilistic Amplification

Fingerprints

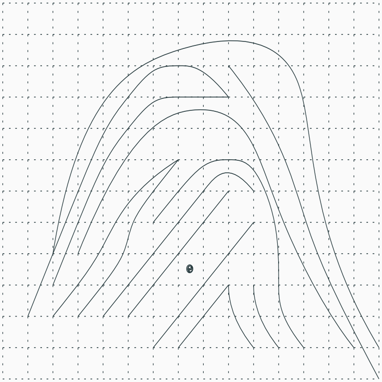
Matching Fingerprints

Fingerprints consists of **minutia points** and patterns that form ridges and bifurcations



Fingerprint with an overlay grid

Fingerprint mapped to a normalized grid cell



Minutia of two fingerprints

Statistical Analysis from fingerprint analyst:

1. Pr(minutia in a random grid cell of a fingerprint) $= 0.2$
2. Pr(given two fingerprints of the same finger and that one fingerprint has a minutia in a grid cell, other fingerprint has the minutia in that cell) $= 0.85$
3. Pick 3 random grid cells and define a (hash) function f that sends two fingerprints to the same bucket if they have minutia in each of those three cells
4. Pr(two arbitrary fingerprints will map to the same bucket by f) $= 0.2^6 = 0.000064$
5. Pr(f maps the fingerprints of the same finger to the same bucket) $= 0.2^3 \times 0.85^3 = 0.0049$

Probabilistic Amplification

Suppose we have 1000 such functions and we take ‘OR’ of these functions

1. Pr(two fingerprints from different fingers map to the same bucket) $= 1 - (1 - 0.000064)^{1000} \approx 0.061$
2. Pr(two fingerprints of the same finger map to the same bucket) $= 1 - (1 - 0.0049)^{1000} \approx 0.992$

Take two groups of 1000 functions each and report a match if it’s a match in both the groups.

1. Pr(two fingerprints from different fingers map to the same bucket) $\approx 0.061^2 = 0.0037$
2. Pr(two fingerprints of the same finger map to the same bucket) $\approx 0.992^2 = 0.984$

References

Conclusions

LSH has abundance of applications
(Image Similarity, Documents Similarity, Nearest Neighbors, Similar Gene-Expressions, . . .)

Main References:

1. Piotr Indyk and Rajeev Motwani, Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality, STOC1998
2. Aristides Gionis, Piotr Indyk and Rajeev Motwani, Similarity Search in High Dimensions via Hashing, VLDB 1999
3. LSH Algorithm and Implementation `http://www.mit.edu/~andoni/LSH/`
4. Chapter 3 in MMDS book (mmds.org)
5. Chapter on LSH in My Notes on Topics in Algorithm Design