LSH Families of Hash Functions

Definition

Combining hash functions

Making steep S-Curves

Mining of Massive Datasets
Leskovec, Rajaraman, and Ullman
Stanford University

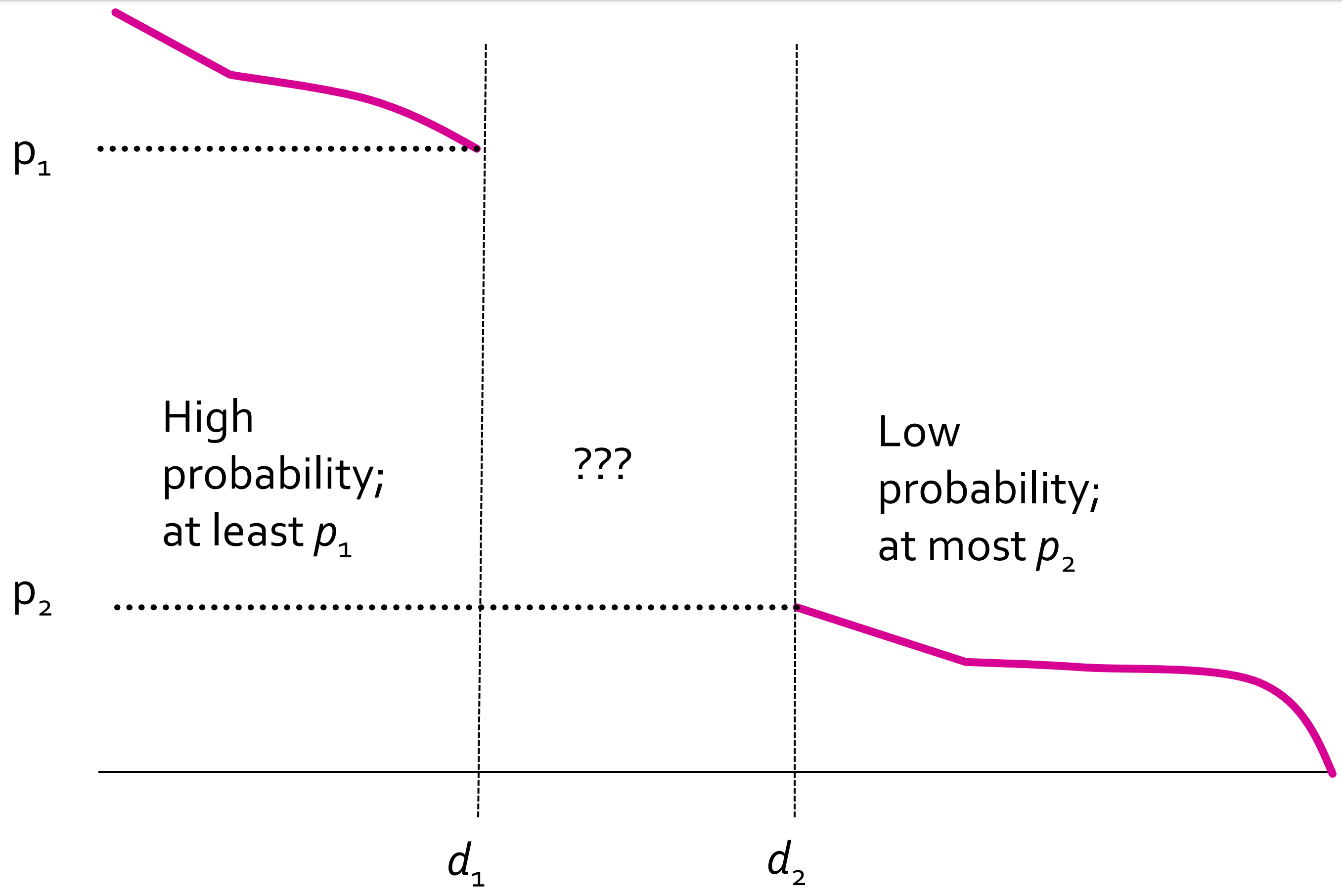
Hash Functions Decide Equality

- There is a subtlety about what a "hash function" really is in the context of LSH families.
- A hash function h really takes two elements x and y, and returns a decision whether x and y are candidates for comparison.
- Example: the family of minhash functions computes minhash values and says "yes" iff they are the same.
- Shorthand: "h(x) = h(y)" means h says "yes" for pair of elements x and y.

LSH Families Defined

- Suppose we have a space *S* of points with a distance measure *d*.
- A family **H** of hash functions is said to be *(d_1,d_2,p_1,p_2)-sensitive* if for any *x* and *y* in *S*:
  1. If $d(x,y) \leq d_1$, then the probability over all *h* in **H**, that $h(x) = h(y)$ is at least p_1.
  2. If $d(x,y) \geq d_2$, then the probability over all *h* in **H**, that $h(x) = h(y)$ is at most p_2.

LS Families: Illustration



Example: LS Family

- Let *S* = sets, *d* = Jaccard distance, **H** is formed from the minhash functions for all permutations.
- Then Prob[h(x)=h(y)] = 1-d(x,y).
  - Restates theorem about Jaccard similarity and minhashing in terms of Jaccard distance.

Example: LS Family – (2)

- Claim: **H** is a ($\boxed{1/3}$, 2/3, $\boxed{2/3}$, 1/3)-sensitive family for *S* and *d*.

If distance $\leq$ 1/3
(so similarity $\geq$ 2/3)

Then probability
that minhash values
agree is $\geq$ 2/3

For Jaccard similarity, minhashing gives us a
$(d_1, d_2, (1-d_1), (1-d_2))$-sensitive family for any $d_1 < d_2$.

Amplifying a LSH-Family

- The “bands” technique we learned for signature matrices carries over to this more general setting.
  - Goal: the “S-curve” effect seen there.
- AND construction like “rows in a band.”
- OR construction like “many bands.”

AND of Hash Functions

- Given family **H**, construct family **H'** whose members each consist of *r* functions from **H**.
- For $h = \{h_1,\ldots,h_r\}$ in **H'**, $h(x)=h(y)$ if and only if $h_i(x)=h_i(y)$ for all i.
- Theorem: If **H** is (d_1,d_2,p_1,p_2)-sensitive, then **H'** is $(d_1,d_2,(p_1)^r,(p_2)^r)$-sensitive.
  - Proof: Use fact that h_i's are independent.

OR of Hash Functions

- Given family **H**, construct family **H'** whose members each consist of *b* functions from **H**.
- For $h = \{h_1,\ldots,h_b\}$ in **H'**, h(x)=h(y) if and only if $h_i(x)=h_i(y)$ for some *i*.
- Theorem: If **H** is (d_1,d_2,p_1,p_2)-sensitive, then **H'** is $(d_1,d_2,1-(1-p_1)^b,1-(1-p_2)^b)$-sensitive.

Effect of AND and OR Constructions

- AND makes all probabilities shrink, but by choosing *r* correctly, we can make the lower probability approach 0 while the higher does not.
- OR makes all probabilities grow, but by choosing *b* correctly, we can make the upper probability approach 1 while the lower does not.

Composing Constructions

- As for the signature matrix, we can use the AND construction followed by the OR construction.
  - Or vice-versa.
  - Or any sequence of AND's and OR's alternating.

AND-OR Composition

- Each of the two probabilities *p* is transformed into $1-(1-p^r)^b$.
  - The "S-curve" studied before.
- Example: Take **H** and construct **H'** by the AND construction with *r* = 4. Then, from **H'**, construct **H''** by the OR construction with *b* = 4.

Table for Function $1-(1-p^4)^4$

p	$1-(1-p^4)^4$
.2	.0064
.3	.0320
.4	.0985
.5	.2275
.6	.4260
.7	.6666
.8	.8785
.9	.9860

Example: Transforms a (.2,.8,.8,.2)-sensitive family into a (.2,.8,.8785,.0064)-sensitive family.

OR-AND Composition

- Each of the two probabilities *p* is transformed into $(1-(1-p)^b)^r$.
  - The same S-curve, mirrored horizontally and vertically.
- Example: Take **H** and construct **H'** by the OR construction with *b* = 4. Then, from **H'**, construct **H''** by the AND construction with *r* = 4.

Table for Function $(1-(1-p)^4)^4$

p	$(1-(1-p)^4)^4$
.1	.0140
.2	.1215
.3	.3334
.4	.5740
.5	.7725
.6	.9015
.7	.9680
.8	.9936

Example: Transforms a (.2,.8,.8,.2)-sensitive family into a (.2,.8,.9936,.1215)-sensitive family.

Cascading Constructions

- Example: Apply the (4,4) OR-AND construction followed by the (4,4) AND-OR construction.
- Transforms a (.2,.8,.8,.2)-sensitive family into a (.2,.8,.9999996,.0008715)-sensitive family.

General Use of S-Curves

- For each S-curve $1-(1-p^r)^b$, there is a *threshold* t, for which $1-(1-t^r)^b = t$.
- Above t, high probabilities are increased; below t, they are decreased.
- You improve the sensitivity as long as the low probability is less than t, and the high probability is greater than t.
  - Iterate as you like.

Visualization of Threshold