LOCALITY-SENSITIVE HASHING

A frequent issue is, given a set S of items, each one with d features, to find the largest group of similar items (where similarity is a function that, taken the features of two items, returns a value in $[0,1]$).

The LSH alg. generates a fingerprint for every item of the set, that is much shorter than the vector of features, and transforms the problem of the similarity between features in the equality between fingerprints.

This alg. is correct with high probability and guarantees local access to data, which reduces the number of I/O operations.

How it works

Assuming binary features only, for each p, q binary vectors, LSH uses an hash table to execute the similarity check:

if $h(p) == h(q)$, then p and q are similar.

which h do we need?

h chooses a set I with size k of random coordinates

EXAMPLE

If $I = \{1, 4\}$ (here $k = 2$)

first coordinate

fourth coordinate

then $h(01011) = 01$

binary vector

What about false positive?

Given two binary vector p and q

$$P(\text{pick } x \text{ such that } p(x) \neq q(x)) = \frac{D(p,q)}{d}$$

where $D(p,q)$ is the Hamming Distance, which returns the number of different bits between p and q

Therefore:

$$P(\text{pick } x \text{ such that } p(x) == q(x)) = 1 - \frac{D(p,q)}{d}$$

That said:

$$P(h(p) == h(q)) =$$

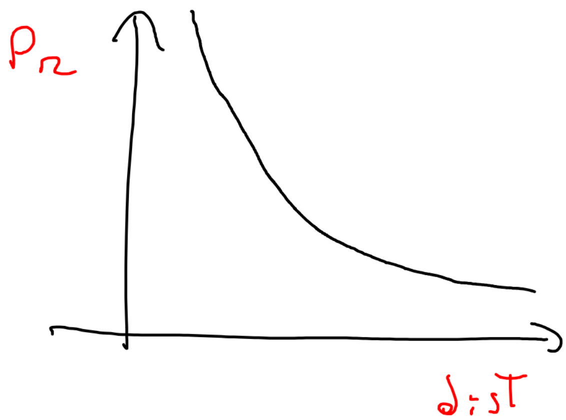
$$= P(\text{pick } x \text{ such that } p(x) == q(x))^K =$$

$$= \left(1 - \frac{D(p,q)}{d}\right)^K$$

It is clear now that the probability of a false positive is bounded by K

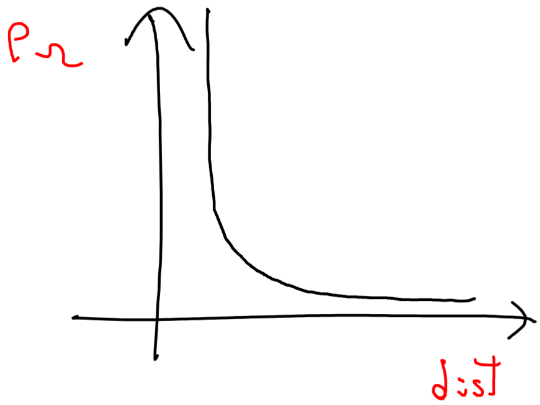
- Small K

more False Positive



- Large K

Less False Positive



<u>But how do we address the false negative then?</u>

Repeat the hashing L times using different set I of random coordinates:

(1) Set up L hashes : $h_1(p), \dots, h_L(p)$

(2) p is similar to q if there is at least an i such that $h_i(p) == h_i(q)$

$$P(p \text{ matches } q) = P(\exists i : h_i(p) == h_i(q)) =$$

$$= 1 - P(\forall i : h_i(p) \neq h_i(q)) =$$

$$= 1 - P(h_i(p) \neq h_i(q))^L =$$

$$= \boxed{1 - \left(1 - \left(1 - \frac{D(p,q)}{d}\right)^K\right)^L}$$

The probability of a false negative is bounded by L = larger L, fewer false negatives

EXAMPLE OF REITERED MATCH

$L = 3$, $K = 2$, $p = 01001$, $q = 01101$

- $I_1 = \{3, 4\}$ $\begin{cases} h_1(p) = 00 \\ h_1(q) = 10 \end{cases}$ ✗
- $I_2 = \{1, 3\}$ $\begin{cases} h_2(p) = 00 \\ h_2(q) = 01 \end{cases}$ ✗
- $I_3 = \{1, 5\}$ $\begin{cases} h_3(p) = 01 \\ h_3(q) = 01 \end{cases}$ ✓

p matches q

IN PRACTICE

p matches q if they fall in the same bucket at least once