Warm-up

Problem 1. Give a data structure for the Nearest Neighbour problem over a d-dimensional universe using space $O(nd)$, for which QUERY runs in time $O(nd)$). (Also, show that it can maintain S dynamically, and implement INSERT and REMOVE methods running in time $O(nd)$.)

Problem 2. Give a data structure for the Nearest Neighbour problem over $\{0,1\}^d$ using space $O(2^d)$, for which QUERY runs in time $O(2^d)$ (independent of n). (Also, can maintain S dynamically, and implement INSERT and REMOVE methods running in time $O(1)$.)

Problem 3. Check your understanding: since we want very efficient lookups and are willing to accept a small probability of failure for QUERY, can we use Bloom filters for the "baby version" of LSH instead of hash tables? What fails?

Problem solving

Problem 4. Prove Theorem 38 from the lecture notes, showing how to solve the "general" ANN from the "baby version," at the cost of only a logarithmic factor in the dimension d.

Problem 5. Analyse the LSH family described in the lecture notes for the Euclidean case, where a locally-sensitive hash function $h_g \colon \mathbb{R}^d \to \{-1,1\}$ is obtained by drawing a d-dimensional Gaussian random vector $g \sim \mathcal{N}(0_d, I_d)$ (all coordinates are independent $\mathcal{N}(0,1)$ normal random variables) and setting

$$h_g \colon x \in \mathbb{R}^d \to \operatorname{sign}\Big(\sum_{i=1}^{d} g_i x_i\Big)$$

Show that, for every $r > 0, C > 1$, this defines an (r, C, p, q)-LSH family with p, q such that $\rho \leq 1/C$. *[Note: this is called the SimHash scheme.]*

Problem 6. For the set $[d] = \{1, 2, \ldots, d\}$, let the universe $\mathcal{X}$ be the set of all 2^d subsets of $[d]$, along with the *Jaccard distance*:

$$\operatorname{dist}(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}, \qquad A, B \in \mathcal{X}$$

Consider the following hash family $\mathcal{H}$: for every permutation $\pi \colon [d] \to [d]$, define $h_\pi \colon \mathcal{X} \to [d]$ by setting

$$h_\pi(A) = \min_{a \in A} \pi(a)$$

and $\mathcal{H} = \{h_\pi\}_\pi$.

a) Verify that the Jaccard distance is a metric on $\mathcal{X}$. What is its range?

b) What is the size of $\mathcal{H}$?

c) Show that, for every $r \in (0, 1]$ and $C > 1$, $\mathcal{H}$ is an (r, C, p, q)-LSH family for $p = 1 - r$ and $q = 1 - Cr$. What is its sensitivity parameter ρ?

Advanced

Problem 7. Give a data structure for the Nearest Neighbour problem over the Euclidean space $(\mathbb{R}^d, \ell_2)$ based on kd-trees. Analyse the space complexity of the data structure and its query time.