Weekplan: Approximate Near Neighbor

Philip Bille Inge Li Gørtz Eva Rotenberg

References and Reading

[1] Notes by Aleksandar Nikolov

[2] Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions, Andoni and Indyk, Communications of the ACM, January 2008.

We recommend reading [1] in detail and [2] section 1-3.

Probability theory cheat-sheet

Markov's inequality: For Y being a positive-valued random variable,

$$P[Y \geq t] \leq \frac{\mathbb{E}[Y]}{t} .$$

Exercises

1 Hamming distance

1.1 Solve Exercise 1 and 2 from [1].

1.2 From the proof of Claim 1 on the slides:

(a) Prove that

$$1 - \prod_{\ell=1}^{L} (1 - P[g_\ell(x) = g_\ell(z^*)]) \geq 1 - \prod_{\ell=1}^{L} (1 - p_1^k) .$$

(b) Prove that $Lp_1^k = 2$. *Hint:* Recall that $k = \lg n / \lg(1/p_2)$.

2 Jaccard distance and Sim Hash The *Jaccard similarity* of two sets is defined as $\textsc{Jsim}(A, B) = \frac{|A \cap B|}{|A \cup B|}$. In *Min-Hash* you pick a random permutation π of the elements in the universe and let $h(A) = \min_{a \in A} \pi(a)$.

2.1 Let $S_1 = \{a, e\}, S_2 = \{b\}, S_3 = \{a, c, e\}, S_4 = \{b, d, e\}$. Compute the Jaccard similarity of each pair of sets.

2.2 Let S_1, S_2, S_3, S_4 be as above and let the random permutation be (b, d, e, a, c), i.e., $\pi(a) = 4$, $\pi(b) = 1$, etc. Compute the min-hash value of each of the sets.

2.3 Prove that the probability that the min-hash of two sets is the same is equal to the Jaccard similarity of the two sets, i.e., that $P[h(A) = h(B)] = \frac{|A \cap B|}{|A \cup B|}$.

2.4 Argue that if the Jaccard similarity of two sets are 0 then Min-Hashing always give the correct estimate.

2.5 The Jaccard distance is defined as $d_J(a, b) = 1 - \textsc{Jsim}(A, B)$. Show that the Jaccard distance is a metric. That is, show that:

1. $d_J(A, B) \geq 0$ for all sets A and B,
2. $d_J(A, B) = 0$ if and only if $A = B$,
3. $d_J(A, B) = d_J(B, A)$,
4. $d_J(A, B) \leq d_J(A, C) + d_J(C, B)$ for all sets A, B and C.

Hint: For 4. use $P[h(A) = h(B)] = \frac{|A \cap B|}{|A \cup B|}$.

3 Hamming Distance 2 In this exercise we will analyse the LSH scheme for Hamming Distance. Recall that in a query we stop after we have checked $6L + 1$ strings. Let $F = \{y \in P : d(x, y) > cr\}$ (strings far from x) and let z^* be a fixed string with $d(x, z^*) \leq r$. We say that y collides with x if $g_j(x) = g_j(y)$ for some $i \in \{1, \ldots \ell\}$.

3.1 Explain why it is enough to prove that the following two properties hold:

1. the number of strings in F that collides with x at most $6L$.
2. z^* collides with x.

3.2 Let y be a string in F. Prove that $P[y$ collides with x in $T_j] \leq 1/n$. *Hint:* Recall that $k = \log n / \log(1/p_2)$.

3.3 Let $X_{y,j} = 1$ if y collides with x in T_j and 0 otherwise, and let $X = \sum_{y \in F} \sum_{j=1}^{L} X_{y,j}$. Prove that $E[X] \leq L$.

3.4 Use Markov's inequality to show that $P[X > 6L] < 1/6$.

3.5 Prove that if there exists a string z^* in P with $d(x, z^*) \leq r$ then with probability at least 2/3 we will return some y in P for which $d(x, y) \leq cr$.