Programming and Classification

List 4

Marek Klonowski

XI. Minhashing and LSH I

1. Estimate the probability that a random mapping, i.e., a random function from $\{1, \ldots, m\}$ into $\{1, \ldots, m\}$ is a permutation.
2. Construct the characteristic matrix for the family of sets
$$\{\{a, b, c\}, \{a, b, d\}, \{a, f\}\} .$$
When the characteristic matrix is **not** an efficient set representation?
3. Prove that the probability that the minhash function for a random permutation of rows produces the same value for two sets equals the Jaccard similarity of those sets.
4. Let us assume that a signature of a set consists of 1000 minhashes. For what size of the set the signature is shorter than the regular representation of the set?
5. What is the length of the signature (i.e., the number of minhashes) sufficient to estimate the Jaccard similarity with an error smaller than 1%? (approximation is enough, no exact calcunation is needed)
6. Let $\mathcal{F}$ be a family of a (d_1, d_2, p_1, p_2)-sensitive functions. Construct a family of hash functions that is
   - $(d_1, d_2, (p_1)^2, (p_2)^2)$-sensitive,
   - $(d_1, d_2, 1 - (1 - (p_2)^2)^3), 1 - (1 - (p_2)^2)^3)$-sensitive.
7. Prove that for any $s \in (0, 1)$ one can find r, b, such that
$$1 - (1 - s^b)^r = 1/2.$$
Find an interpretation in terms of families of sensitive hash functions.

XII. Misc

1. Find the point on the line $Ax + By + C = 0$ that is closest to the point (x_0, y_0).
2. We randomly permute a database with n enumerated records. What is the probability that the first record will not change its original position?
3. We radomly generate two vectors $a = (a_1, a_2, \ldots, a_n)$, $b = (b_1, b_2, \ldots, b_n)$. Each number a_i, b_i is -1 or 1 with probability $1/2$ (independently).
   - What is maximal and minimal consie similarity between a and b?
   - Estimate the expected value of the consie similarity between a and b.
   - Estimate the probability that the consie similarity between a and b is bigger then 0.1.
4. Answer the same questions in the case if a_i, b_i are 0 and 1 with probability $1/2$ each.
5. Find a unit vector (in a Euclidean space) orthagonal to
   - vector $[1, 2]$,

- vector $[-1, 0, 2]$,
- a surfice spanned by vectors $[1, 0, 0]$ and $[0, 1, -1]$.

6. Find all eigenvalues and unit eigenvectors of a matrix:

   - $$\begin{bmatrix} 1 & 2 \\ 0 & -1 \end{bmatrix}$$

   - $$\begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$