Assignment 3

CS-UH-2218: Algorithmic Foundations of Data Science

Assignments are to be submitted in groups of two or three. *Upload the solutions on NYU classes as one PDF file for theoretical assignments and separate source code files for each programming assignment. Submit only one copy per group. Clearly mention the participant names on each file you submit.*

Problem 1 (10 points)**.**
Solve Exercise 3.1.3 of [MMDS]. *Hint: Exercise 3.3.4 !*

Problem 2 (10 points)**.**
Consider a similarity function $s(x, y)$ which returns a similarity measure in the range $[0, 1]$ between two items x and y. We say that s is LSHable if there is a family of hash functions $\mathcal{H}$ so that the following holds: For any set of items x, y, z, w with $s(x, y) \geq s(z, w)$, if h is chosen from $\mathcal{H}$ uniformly at random, then $\Pr[h(x) = h(y)] \geq \Pr[h(z) = h(w)]$. In other words, a higher similarity between items implies a higher probability of being matched under a randomly chosen hash function from $\mathcal{H}$.

1. Prove that if the function s is LSHable then the function $d(x, y) = 1 - s(x, y)$ satisfies the triangle inequality i.e. for any three items x, y and z, $d(x, y) + d(y, z) \geq d(x, z)$.

   Hint. Note that if $h(x) \neq h(z)$ then either $h(x) \neq h(y)$ or $h(y) \neq h(z)$.

2. Consider the following similarity measure for sets called *Sørensen's Dice Similarity Coefficient* : $dsc(A, B) = \frac{2|A \cap B|}{|A|+|B|}$. Use the above to conclude that dsc is not LSHable.

   Hint. Define $d(A, B) = 1 - dsc(A, B)$. Construct sets A, B and C s.t. $d(A, B) + d(B, C) < d(A, C)$. There are very small sets A, B, C satisfying this.

Problem 3 (10 points)**.**
Let $\mathbf{v}$ be any (fixed) vector in $\mathbb{R}^d$. Consider a random gaussian vector $\mathbf{g} = (g_1, \cdots, g_d) \in \mathbb{R}^d$ where each component g_i of $\mathbf{g}$ is chosen independently from the normal distribution $\mathcal{N}(0, 1)$ with mean 0 and variance 1. What is the distribution of the random variable $x = \mathbf{g} \cdot \mathbf{v}$?
Please explain your answer.