Due date: February 04, 2025

Problem 1: Potential problems:

- (Easiest) Prove XXX is LSH. For example what would a good LSH on $\{0,1\}^d$ with respect to the d-dimensional Hamming distance?
- (Moderate) If you don't plan to cover amplication in details, we can make create a question asking them to analyze the effects of having k independent hash functions.
- (Another moderate problem) Random hyperplanes and the cosine distance, see section 3.7.2 of this book. Each function is defined by a randomly chosen vector v_f. Given two vectors x, y, say $f(x) = f(y)$ if and only if the dot products $\langle v_f, x\rangle$ and $\langle v_f, y\rangle$ have the same sign. Find the right parameters to make the collection of such functions LSH.
- (Hard, probably multi-part) Reducing ANN search to LSH (see here). Can only assign this if lecture covers amplification. Or maybe I can make amplification itself a problem and leave the remaining ANN query as a bonus part.

Problem 2: Let S be a set of n points in $\mathbb{R}^2$, let q_i be the charge of point p_i in S, and let ϵ be a parameter. Assume that the spread of S is $n^{O(1)}$. Define

$$F(S) = \sum_i \sum_{j \neq i} \frac{q_i q_j}{\|p_i - p_j\|^2}.$$

Describe an $O(n/\varepsilon^{-2} \log n)$ algorithm to compute an estimate $\tilde{F}(S)$ of $F(S)$ that has a multiplicative error of $\leqslant \epsilon$, i.e., $|\tilde{F}(S) - F(S)| \leq \varepsilon F(S)$.

Problem 3: In this problem, you will implement and analyze three clustering algorithms: (i) Lloyd's algorithm (k-means), (ii) a randomized greedy algorithm with local search, and (iii) a hybrid approach that combines the two for seed selection. Please choose one of the following datasets for your experiments:

- Iris Dataset: $n = 150$, $d = 4$, $k = 3$. If you are using Python, you may load this directly via `sklearn.datasets.load_iris()`.
- Pen-Based Handwritten Digits: $n = 10992$, $d = 16$, $k = 10$.
- If you are using Python: Gaussian Blobs via `sklearn.datasets.make_blobs`. Configure $n = 300$, $d = 2$, and $k = 4$.

Please finish the following tasks:

(i) Implement Lloyd's algorithm. Start by initializing k cluster centers arbitrarily. Track the clustering cost (sum of squared distances to centers) and analyze how it evolves over iterations.

(ii) Implement a randomized greedy algorithm that selects more than k centers before refining the selection with local search. How does increasing the number of initial selections impact performance?

(iii) Compare arbitrary seed selection with randomized greedy initialization. How does the choice of initial centers affect clustering cost and convergence?

(iv) Track and plot the clustering cost over iterations for different values of k. How does performance vary as k increases? *Optionally, if you are familiar with dimensionality reduction techniques such as PCA or t-SNE, consider down-projecting the dataset onto* $\mathbb{R}^2$ *and directly visualize the clustering results. No extra credits will be awarded, but it may make your solution look nicer. :)*

(v) Compare different termination conditions for Lloyd's algorithm, for example stopping after a fixed number of iterations or when the cluster centers stabilize, or your own criteria. Which one(s) work best in practice, and why do you think this is the case?