Complete all problems.

You are not permitted to look at solutions of previous years’ assignments. You can work together in groups, but all solutions must be written up individually. If you get information from sources other than the course notes and slides, please cite the information, even if from Wikipedia or a textbook.

**Problem 1: The Exacting LSH (25pt)**
A similarity function \( \text{sim}(\cdot, \cdot) \) takes any pair of items and outputs a number between 0 and 1. A similarity function is called *exact-LSHable* if there exists a hash family \( H \) such that
\[
\text{sim}(x, y) = \Pr_{h \leftarrow H} [h(x) = h(y)].
\]
Recall the Jaccard similarity \( \text{sim}_{\text{Jac}}(S, T) := \frac{|S \cap T|}{|S \cup T|} \) was exact-LSHable.

(i) If a similarity function \( \text{sim} \) is exact-LSHable, show that the distance function \( d(x, y) = 1 - \text{sim}(x, y) \) must satisfy the triangle inequality: \( d(x, y) \leq d(x, z) + d(z, y) \) for all \( x, y, z \).

(ii) Show that Dice’s similarity function
\[
\text{sim}_{\text{Dice}}(S, T) = \frac{|S \cap T|}{\frac{1}{2}(|S| + |T|)}
\]
is not exact-LSHable.

(iii) Show that the Overlap similarity function
\[
\text{sim}_{\text{over}}(S, T) = \frac{|S \cap T|}{\min(|S|, |T|)}
\]
is not exact-LSHable.

For the latter two parts, give examples of three sets \( A, B, C \) such that the distance function obtained from these similarity functions does not satisfy the triangle inequality. (Hint: you only need to consider sets of size at most 2.)

**Problem 2: Finding Your Neighbors (25pt)**
One application of LSH used a lot in practice is to approximately answer nearest-neighbor queries quickly. Suppose we have a dataset \( \mathcal{A} \) of \( n \) points in a metric space with distance metric \( d \). Given a query point \( z \) with the promise that there is some \( x \in \mathcal{A} \) with \( d(x, z) \leq \lambda \),
we want to return a point $x' \in A$ with $d(x', z) \leq c\lambda$ for some $c > 1$. Such a point is called a $(c, \lambda)$-approximate nearest neighbor (ANN).

Choose a $(\lambda, c\lambda, p_1, p_2)$-LSH family $\mathcal{H}$ for the distance metric $d$, and let $\mathcal{G}$ be the set of $k$-tuples of functions from $\mathcal{H}$, where $k = \lceil \log_{1/p_2}(n) \rceil$. If the functions in $\mathcal{H}$ map into the set $[m]$, then functions in $\mathcal{G}$ map into the set $[M] := [m]^k$. To find $(c, \lambda)$-ANNs, we propose the following algorithm. First, build a data structure:

1. Sample $L = n^\rho$ members $g_1, \ldots, g_L$ of $\mathcal{G}$, where $\rho = \frac{\log(1/p_1)}{\log(1/p_2)}$.
2. For each $x \in A$, compute the hash values $g_1(x), g_2(x), \ldots, g_L(x)$ and store the index of $x$ in those $L$ buckets.

To query a point $z$:

1. Compute the $L$ hash values $g_i(z)$, and figure out which buckets it lands in.
2. Choose at most $3L$ points from those buckets uniformly at random. (If there are fewer than $3L$ matches, just take them all.)
3. For those $3L$ points, compute $d(x, z)$ and return the one whose distance is minimal.

We will first show that this algorithm yields a correct answer with constant probability.

(i) Let $W_j = \{x \in A \mid g_j(x) = g_j(z)\}$, for $1 \leq j \leq L$, be the set of data points which $g_j$ maps to the same bucket as $z$. Define $T = \{x \in A \mid d(x, z) > c\lambda\}$. Prove that:

$$\Pr \left[ \sum_{j=1}^{L} |T \cap W_j| \geq 3L \right] \leq \frac{1}{3}.$$  

*Hint:* try the Markov inequality.

(ii) Let $x^* \in A$ be a point with $d(x^*, z) \leq \lambda$. Prove that

$$\Pr [g_j(x^*) \neq g_j(z) \forall 1 \leq j \leq L] < \frac{1}{e}.$$  

(iii) Conclude that, with a constant probability, the reported point is a $(c, \lambda)$-ANN.

Now that we’ve shown that the algorithm works, let’s try it out. Matlab code along with a dataset of images is available at [http://realworld.herokuapp.com/ann/lsh.zip](http://realworld.herokuapp.com/ann/lsh.zip) (Matlab, if you don’t have it, is available from Computing Services or on cluster computers.)

Each column in this dataset is a $20 \times 20$ image patch, represented as a 400-dimensional vector. We’ll empirically compare the performance of LSH-based ANN search with linear

---

4 Code and dataset adapted from Greg Shakhnarovich at Brown by way of Jure Leskovec at Stanford; this question is also largely adapted from the latter.

[http://www.cmu.edu/computing/software/all/matlab/](http://www.cmu.edu/computing/software/all/matlab/)
search in retrieving image patches based on $L_1$ distance. (This is far too high-dimensional for traditional, exact nearest-neighbor indices like kd-trees to be helpful.) You should use the provided code, which is explained in the README.txt file; in particular, you’ll need the functions lsh and lshlookup.

(iv) For each of the image patches in columns 100, 200, 300, . . . , 1000, find the top 3 nearest neighbors (excluding the query patch, which of course is in the index as well) using both LSH and linear search. What is the average search time for each method?

Note that sometimes, nnlsh may return less than the requested number of neighbors. If that happens, you should rebuild the indices so that it returns the correct number of neighbors.

(v) Call \{z_j\} the set of query points (1 ≤ j ≤ 10; i.e. $z_j$ is the patch in column 100j), \{x_{ij}\}_{i=1}^3 the approximate neighbors found via LSH, and \{x_{ij}^*\}_{i=1}^3 the true nearest neighbors found via linear search. Compute the following error measure:

$$\frac{1}{10} \sum_{j=1}^{10} \frac{\sum_{i=1}^{3} d(x_{ij}, z_j)}{\sum_{i=1}^{3} d(x_{ij}^*, z_j)}.$$

Plot this error value as a function of $L$, for $L = 10, 12, 14, \ldots, 20$, with $k = 24$. Also plot the error value as a function of $k$, for $k = 16, 18, 20, 22, 24$, with $L = 10$. Briefly describe the trend shown in each plot (one or two sentences each).

(vi) Finally, plot the top 10 nearest neighbors found via the two methods (with $L = 10$, $k = 24$ for LSH) for the image patch in column 200. (The readme includes examples of plotting these patches.) How do the results look visually?

For parts (iv)-(vi), hand in a printout of Matlab code to produce the answer to each section, as well as the plots and your descriptions.